

UNIVERSIDADE FEDERAL DE SANTA MARIA
CENTRO DE CIÊNCIAS NATURAIS E EXATAS
PROGRAMA DE PÓS-GRADUAÇÃO EM CIÊNCIAS BIOLÓGICAS:
BIOQUÍMICA TOXICOLÓGICA

Rayana dos Santos Feltrin

**LACUNAS NA EVOLUÇÃO DA VIA DE REPARO POR EXCISÃO DE
NUCLEOTÍDEOS EM EUCARIOTOS**

Santa Maria, RS

2020

Rayana dos Santos Feltrin

**LACUNAS NA EVOLUÇÃO DA VIA DE REPARO POR EXCIÇÃO DE
NUCLEOTÍDEOS EM EUCARIOTOS**

Dissertação apresentada ao Curso de Pós-Graduação em Ciências Biológicas: Bioquímica Toxicológica, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do título de **Mestre em Ciências Biológicas: Bioquímica Toxicológica.**

Orientador: Prof. Dr. André Passaglia Schuch

Coorientadora: Dra. Ana Lúcia Anversa Segatto

Santa Maria, RS
2020

Feltrin, Rayana dos Santos
Lacunas na Evolução da Via de Reparo por Excisão de
Nucleotídeos em Eucariotos / Rayana dos Santos Feltrin.-
2020.

54 p.; 30 cm

Orientador: André Passaglia Schuch
Coorientadora: Ana Lúcia Anversa Segatto
Dissertação (mestrado) - Universidade Federal de Santa
Maria, Centro de Ciências Naturais e Exatas, Programa de
Pós-Graduação em Ciências Biológicas: Bioquímica
Toxicológica, RS, 2020

1. Reparo de DNA 2. Via NER 3. Eucariotos 4.
Estrutura gênica 5. Arquitetura de domínios I. Schuch,
André Passaglia II. Segatto, Ana Lúcia Anversa III.
Título.

Sistema de geração automática de ficha catalográfica da UFSM. Dados fornecidos pelo autor(a). Sob supervisão da Direção da Divisão de Processos Técnicos da Biblioteca Central. Bibliotecária responsável Paula Schoenfeldt Patta CRB 10/1728.

© 2020

Todos os direitos autorais reservados a Rayana dos Santos Feltrin. A reprodução de partes ou do todo deste trabalho só poderá ser feita mediante a citação da fonte.

Endereço: Av. João Machado Soares, 1240, Bloco A1, aptº 201, Bairro Camobi, Santa Maria, RS. CEP: 97110-000.

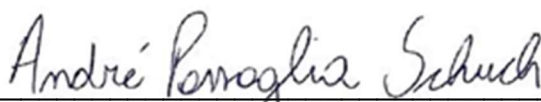
Fone (055) 55 984513141; E-mail: ryanafeltrin@gmail.com

Rayana dos Santos Feltrin

**LACUNAS NA EVOLUÇÃO DA VIA DE REPARO POR EXCIÇÃO DE
NUCLEOTÍDEOS EM EUKARIOTOS**

Dissertação apresentada ao Curso de Pós-Graduação em Ciências Biológicas: Bioquímica Toxicológica, da Universidade Federal de Santa Maria (UFSM, RS), como requisito parcial para a obtenção do título de **Mestre em Ciências Biológicas: Bioquímica Toxicológica**

Aprovada em 21 de outubro de 2020:



André Passaglia Schuch, Dr. (UFSM)
(Presidente/Orientador)



Carlos Frederico Martins Menck, Dr. (USP)



Elgion Lucio da Silva Loreto, Dr. (UFSM)

Santa Maria, RS
2020

DEDICATÓRIA

Dedico este trabalho aos meus pais, aos meus avós – em especial à minha vó Nina (In Memoriam) e ao meu avô Neri – bem como a toda a minha família, que sempre me apoia muito e, sem a qual, eu nada seria! Dedico também aos meus amigos e a toda equipe do Laboratório de Fotobiologia da UFSM: o suporte e convívio de vocês tornou a jornada até aqui muito mais leve, divertida e prazerosa!

AGRADECIMENTOS

No meio acadêmico, sempre ouvimos dizer que, na Ciência, nada se faz sozinho... e, de fato, não se faz! Porém, apesar de vermos grande parte da Ciência ser feita em laboratório, a energia necessária para a reação “fazer ciência” vem de diversas fontes. Assim, gostaria de agradecer a todos que se fizeram e se fazem fonte para o meu “fazer ciência” neste início de minha carreira de cientista.

Primeiramente, agradeço à minha família, que me deu o amor e o calor necessário para aquecer essa reação e o meu coração, principalmente aos meus pais Gladis e Mariano: o amor e gratidão que tenho por vocês tende ao infinito! Também sou muito grata aos meus avós Nina e Neri pelo carinho, muitas vezes dispensado em forma de “precezinhas” para que eu sucedesse em minhas provas e tarefas. Enfim, agradeço à minha família como um todo, que muito me apoiou e entendeu quando eu não podia me fazer muito presente nos encontros! E também ao meu companheiro felino Perseu, que estudou junto comigo para a prova de seleção e sentava no teclado do computador quando era hora de descansar!

Ademais, agradeço imensamente a toda a equipe do Laboratório de Fotobiologia da UFSM, especialmente ao professor André Schuch, que foi e é um “catalisador” no meu “fazer ciência”. Sou muito grata por ter me apresentado o apaixonante mundo do reparo de DNA, me acolhido de volta no laboratório e muitas vezes ter sido um pai pra mim! Sou muito grata por trabalhar contigo e agradeço pela paciência e incentivo incessante quando, por muitas vezes, eu não acreditei em mim. Te admiro muito como pessoa e pesquisador! Também agradeço muito à Ana Lúcia Segatto, que me apresentou e ensinou a amar a Bioinformática e me guiou como uma polimerase nesse processo de aprendizado! Te admiro muito como pessoa, pesquisadora e mulher na ciência! Não posso deixar de agradecer também ao Tiago de Souza, que tem uma participação fundamental neste trabalho e tirou meu medo da “tela preta”, incentivando-me a aprender a programar! Além disso, agradeço aos meus colegas de laboratório: nossa amizade e convivência foi e é um cofator - ou seja, essencial - para a minha reação!

Gratidão também aos meus amigos e psicóloga, que nunca deixaram essa reação se acabar! E à UFSM, minha segunda casa desde o Ensino Médio, bem como ao PPGBTox e à CAPES (PROEX 23038.005848/2018-31) pelo apoio financeiro! Avante, ciência brasileira!!!

Para nós, passou totalmente despercebido o possível papel do reparo de DNA, embora mais tarde eu tenha notado que o DNA é tão precioso que, provavelmente, existam muitos mecanismos de reparo distintos.

(Francis Crick)

RESUMO

LACUNAS NA EVOLUÇÃO DA VIA DE REPARO POR EXCISÃO DE NUCLEOTÍDEOS EM EUCARIOTOS

AUTORA: Rayana dos Santos Feltrin

ORIENTADOR: Prof. Dr. André Passaglia Schuch

COORIENTADORA: Dra. Ana Lúcia Anversa Segatto

A via de reparo por excisão de nucleotídeos (NER) é o mais versátil mecanismo de reparo de DNA, já que está envolvido na remoção de diferentes tipos de lesões que causam grandes distorções na dupla-hélice. Devido à sua grande importância na manutenção da integridade genômica, essa via apareceu cedo na evolução das espécies. Além disso, a maioria dos estudos relacionados ao NER está focada em humanos, camundongos, leveduras e bactérias. Considerando a grande quantidade de dados disponíveis em bancos de dados genômicos, é possível obter sequências de componentes do NER em diferentes organismos. Dessa forma, visamos caracterizar potenciais ortólogos de componentes-chave da via NER em organismos eucarióticos usando diferentes critérios estruturais e de similaridade, através do uso de ferramentas de bioinformática. Essa metodologia nos permitiu caracterizar a estrutura de genes e proteínas de maneira comparativa, bem como esclarecer alguns aspectos evolutivos da via NER. Diante disso, foram obtidos resultados de busca significativos para a maioria das proteínas em grande parte dos organismos analisados, principalmente para aquelas que têm um papel essencial na via. Entretanto, reanalisamos importantes diferenças e encontramos novos aspectos que podem implicar um funcionamento distinto do NER em diferentes organismos. Através da demonstração da heterogeneidade das estruturas gênicas e da variedade na arquitetura das proteínas dessa via, nossos resultados revelam diferenças importantes entre o NER humano e o de eucariotos evolutivamente distantes.

Palavras-chave: Reparo de DNA. Via NER. Eucariotos. Estrutura gênica. Arquitetura de domínios.

ABSTRACT

OPEN GAPS IN THE EVOLUTION OF THE EUKARYOTIC NUCLEOTIDE EXCISION REPAIR

AUTHOR: Rayana dos Santos Feltrin

ADVISOR: Prof. Dr. André Passaglia Schuch

CO-ADVISOR: Dr. Ana Lúcia Anversa Segatto

Nucleotide excision repair (NER) is the most versatile DNA repair pathway as it removes different kinds of bulky lesions. Due to its essential role for genome integrity, it appeared early in the evolution of species. However, most published studies are focused on humans, mice, yeast or bacteria. Considering the large amount of information on genome databases, it is currently possible to retrieve sequences from NER components in many organisms. Therefore, we attempted to characterize the potential orthologs of 10 critical components of the human NER pathway in 12 eukaryotic species by using similarity and structural criteria through the use of bioinformatical tools. This approach has allowed us to characterize gene and protein structures comparatively, taking a glance at some evolutionary aspects of the NER pathway. We obtained significant search results for the majority of the proteins in most of the organisms studied, mainly for factors that play a pivotal role in the pathway. However, we revisited significant differences and found new aspects that may imply a distinct functioning of this pathway in different organisms. Through the demonstration of the heterogeneity of the gene structures and a variety in the protein architecture of the NER components evaluated, our results highlight important differences between human NER and evolutionarily distant eukaryotes.

Keywords: DNA repair. NER pathway. Eukaryotes. Gene structure. Domain architecture.

SUMÁRIO

	APRESENTAÇÃO DA DISSERTAÇÃO	9
1	INTRODUÇÃO	10
2	ARTIGO 1 – OPEN GAPS IN THE EVOLUTION OF THE EUKARYOTIC NUCLEOTIDE EXCISION REPAIR.....	18
	Abstract	19
	1. Introduction	19
	2. Material and methods.....	20
	3. Results	21
	4. Discussion	25
	5. Conclusion	27
	References	27
3	ARTIGO 2 – XERODERMA PIGMENTOSUM D HELICASE IN GALLIFORM BIRDS: WHERE IT HAS FLOWN TO?	30
	Abstract	32
	1. Introduction	33
	2. Material and methods	34
	3. Results and discussion	35
	4. Conclusions	40
	References	41
	Supplementary material.....	45
	5. DISCUSSÃO GERAL.....	48
	6. CONCLUSÃO	50
	REFERÊNCIAS.....	51

APRESENTAÇÃO DA DISSERTAÇÃO

A presente dissertação está dividida em cinco capítulos, sendo o primeiro capítulo a introdução geral. O segundo e o terceiro capítulos encontram-se na forma de dois artigos: um publicado, sendo este uma revisão de literatura sobre o mecanismo do NER, juntamente com a busca de ortólogos em organismos eucarióticos; e o outro, que investiga um resultado intrigante encontrado no primeiro artigo, está em preparação para ser submetido. As seções Introdução, Materiais e Métodos, Resultados, Discussão, Conclusão e Referências de cada um desses capítulos encontram-se nos seus respectivos artigos e representam a íntegra deste estudo. Uma discussão geral entre o segundo e o terceiro capítulos é apresentada no quarto capítulo, e o quinto capítulo contém a conclusão geral da dissertação. As referências bibliográficas apresentadas no final da dissertação referem-se às citações que aparecem nos itens Introdução e Discussão Geral.

1 INTRODUÇÃO

1.1 Agentes causadores de danos no DNA e mecanismos de reparo

A preservação da informação contida nos genomas dos seres vivos é fundamental para a perpetuação da vida. Não obstante, a mutagênese exerce um papel imprescindível para a evolução, já que consiste na fonte geradora de variabilidade genética e seleção natural (CHATTERJEE e WALKER, 2017; TUBBS e NUSSENZWEIG, 2017). Dessa forma, os genomas dos organismos estão continuamente expostos a potenciais agentes causadores de danos que alteram a estrutura do DNA e, conseqüentemente, comprometem sua função (DE LAAT; JASPERS; HOEIJMAKERS, 1999; MACHADO et al., 2014). Danos no DNA podem ser provenientes tanto de fontes exógenas, tais como radiação ultravioleta (UV), radiação ionizante e poluentes químicos, quanto de fontes endógenas relacionadas ao metabolismo celular, como espécies reativas de oxigênio (EROs) e até mesmo erros de replicação (KUMAR et al., 2020) (Figura 1a). Entretanto, ao longo da evolução das espécies, vários mecanismos de reparo de DNA foram sendo desenvolvidos a fim de garantir a manutenção da integridade genômica.

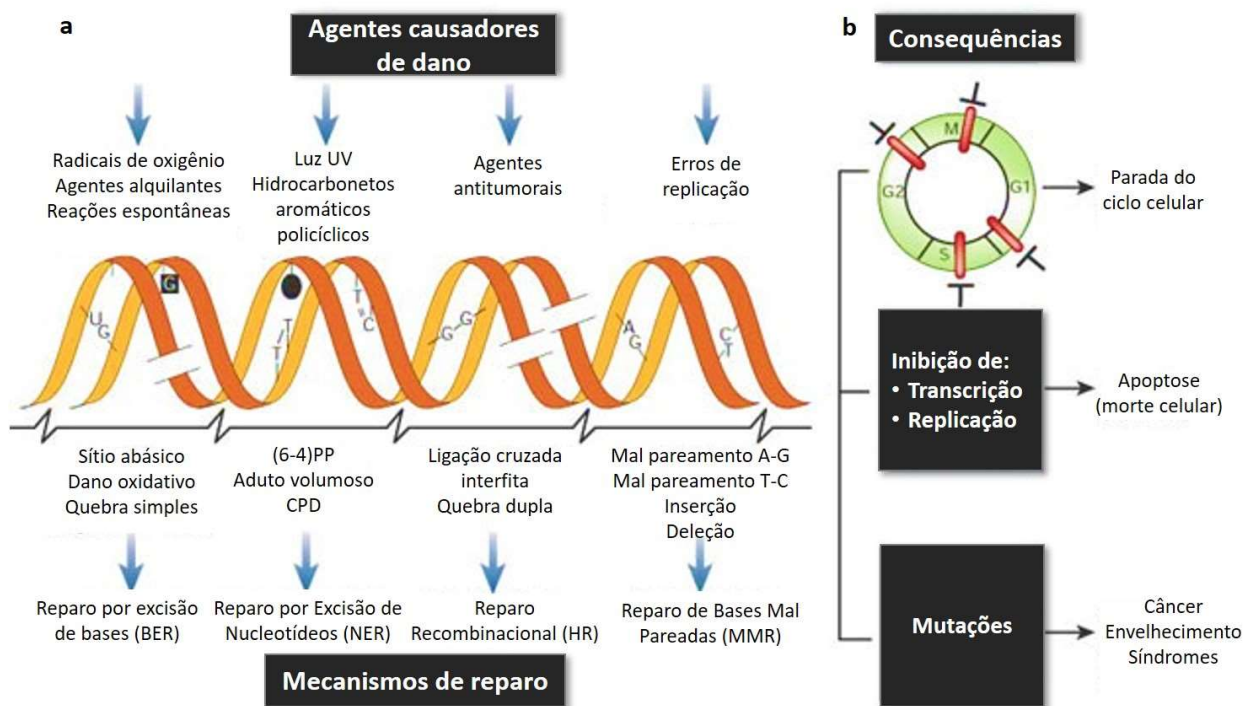


Figura 1: Danos de DNA, mecanismos de reparo e consequências, adaptado de Hoeijmakers (2001). (a) Agentes comuns causadores de danos no DNA, exemplos de danos e os principais mecanismos de reparo de DNA envolvidos na sua correção. (b) Consequências dos danos de DNA na progressão do ciclo celular, no metabolismo da célula, bem como as decorrentes de mutações.

Basicamente, de acordo com o tipo de dano de DNA, há cinco principais vias de reparo: reversão direta da lesão, reparo por excisão de bases (BER), reparo por excisão de nucleotídeos (NER), reparo de bases mal pareadas (MMR), reparo por recombinação homóloga (HR), junção de extremidades não homólogas (NHEJ) e reparo de ligações cruzadas interfita (ICL) (KUMAR et al., 2020). Dentre eles, o NER é o mecanismo mais versátil e flexível, já que é capaz de reparar uma ampla variedade de lesões que distorcem a dupla-hélice, mas são estruturalmente distintas (COSTA et al., 2003). Todavia, há lesões que são tipicamente reparadas pelo NER, tais como os fotoprodutos da radiação UV, que incluem dímeros de pirimidina ciclobutano (CPDs) e 6-4 pirimidina-pirimidona (6-4PPs), adutos formados por agentes mutagênicos ambientais, como benzo(a)pireno, bem como danos induzidos por quimioterápicos (SCHÄRER, 2013). Aliás, é importante ressaltar que o NER é a única via de reparo de DNA capaz de remover fotoprodutos da radiação UV em mamíferos placentários (MENCK, 2002; VERMEULEN e FOUSTERI, 2013).

1.2 Reparo por excisão de nucleotídeos (NER)

O NER ocorre basicamente em quatro passos consecutivos: reconhecimento da lesão, abertura da região de dupla-hélice danificada, clivagem e excisão do dano, ressíntese do DNA excisado e ligação das fitas (MENCK e MUNFORD, 2014). Entretanto, de acordo com a região do genoma em que a lesão se encontra, o NER é dividido em duas subvias distintas: reparo do genoma global (GG-NER) e reparo acoplado à transcrição (TC-NER). Enquanto o GG-NER realiza o reparo ao longo do genoma inteiro, inclusive em regiões não transcritas de genes ativos, o TC-NER se encarrega de corrigir preferencialmente lesões localizadas em fitas transcritas de genes ativos (HANAWALT, 2002). Além disso, essas subvias refletem a diferença temporal em que as lesões são removidas, já que danos situados em genes transcricionalmente ativos são reparados mais rapidamente do que em genes inativos (COSTA et al., 2003; MELLON; SPIVAK; HANAWALT, 1987). Entretanto, em relação ao mecanismo, o GG-NER e o TC-NER diferem apenas quanto ao reconhecimento das lesões (KAMILERI; KARAKASILIOTI; GARINIS, 2012).

O início do NER depende do reconhecimento das lesões de DNA em ambas as subvias (COSTA et al., 2003). Em humanos, o GG-NER começa com o reconhecimento do dano por um complexo (Figura 2a) formado por XPC, hHR23B e

centrina-2 (CETN2) (SPIVAK, 2015). Esse complexo não reconhece as lesões propriamente ditas, mas as distorções que elas provocam na dupla-hélice. Diante disso, ele se liga à fita complementar não lesionada, o que explica a ampla gama de lesões que podem ser reparadas pelo NER (LEE et al., 2014). Uma vez que essa ligação ocorre, a proteína hHR23B se dissocia do complexo e não participa do restante do processo (BERGINK et al., 2012; KAMILERI; KARAKASILITI; GARINIS, 2012). Entretanto, quando uma lesão causa uma distorção menor, como no caso de um CPD, o reconhecimento é feito inicialmente pelo complexo DDB, formado por DDB1 e DDB2 (XPE) (RASTOGI et al., 2010; SPIVAK, 2015). Desse modo, DDB recruta XPC, estimulando o reparo de CPDs (COSTA et al., 2003; FITCH et al., 2003).

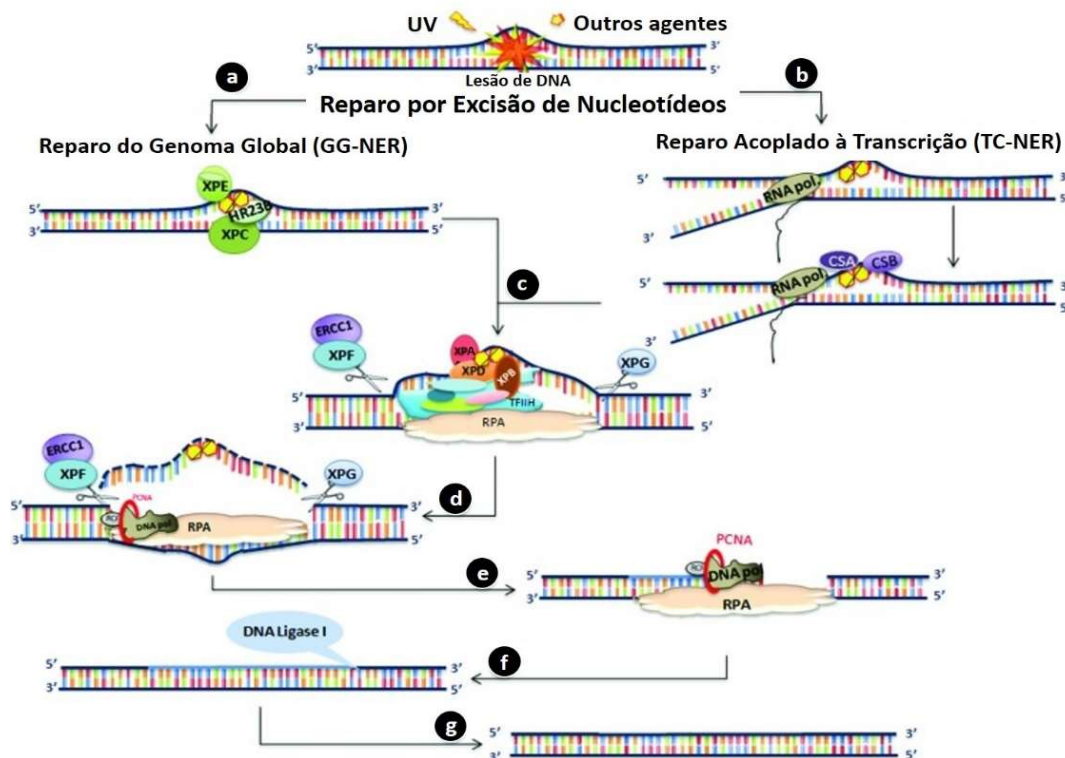


Figura 2: Representação esquemática da via de reparo por excisão de nucleotídeos (NER) em humanos, adaptada de Menck e Munford (2014). A subvia de reparo por genoma global (GGR) inicia com o reconhecimento de dano (a) pelo complexo XPC-HR23B, podendo receber auxílio da proteína XPE (DDB2). Na subvia de reparo acoplada à transcrição (TCR) (b), o reconhecimento tem início com a parada da RNA-polimerase II no sítio da lesão. Isso sinaliza para a proteína CSA se ligar à polimerase, o que recruta CSB a fim de direcionar a ubiquitinação de CSB. Na sequência, a maquinaria de transcrição é removida para permitir a continuação do reparo. Após o reconhecimento das lesões, ambas as subvias convergem para o mesmo mecanismo, que envolve a abertura da molécula de DNA (c) no sítio da lesão pelas proteínas XPB e XPD, que compõem o TFIIH. Para evitar a renaturação do DNA, bem como a degradação da fita simples, ocorre a ligação de RPA. XPA, por sua vez, oferece estabilidade estrutural à maquinaria de reparo, bem como recruta ERCC1-XPF e XPG, de atividade endonucleásica, para a clivagem e excisão do dano (d). Em seguida, o fragmento de DNA excisado é ressintetizado (e) por uma DNA-polimerase com o auxílio de PCNA. As extremidades então são unidas por uma DNA ligase (f), reconstituindo a dupla-hélice (g).

O TC-NER, por sua vez, tem seu início desencadeado pelo bloqueio da RNA-polimerase II (RNA-pol II) no sítio do dano, o que exerce um papel na etapa de reconhecimento (PANI e NUDLER, 2017; SCHUCH et al., 2017). Diante disso, a proteína ERCC6 (CSB) liga-se fortemente à RNA-pol II e envolve o DNA ao redor de si mesma, alterando a conformação da dupla-hélice, bem como a interface entre a RNA-pol II e o DNA (BEERENS et al., 2005; SPIVAK e GANESAN, 2014;). Então, CSB recruta a proteína ERCC8 (CSA) (Figura 2b) e outros fatores do NER para o sítio de bloqueio da RNA-pol II (SPIVAK, 2016). Assim, CSA se junta ao complexo ubiquitina-ligase CRL4CSA e conduz a ubiquitinação de CSB, acarretando sua degradação pelo proteassomo (GROISMAN et al., 2006; JAARSMA et al., 2013). Entretanto, a fim de garantir que CSB dure tempo suficiente para exercer sua função no NER, as proteínas UVSSA e USP7 atuam na sua proteção contra a degradação proteassomal (SCHWERTMAN; VERMEULEN; MARTEIJN, 2013). Segundo um modelo integrado, a subsequente ubiquitinação de CSB – ou até mesmo da subunidade maior da RNA-pol II – por CSA pode funcionar como um sinal para a desmontagem do complexo inicial do TC-NER, a fim de permitir a continuidade do processo de reparo (PASCUCCI et al., 2011). Nesse sentido, o TC-NER desempenha uma importante função ao evitar a obstrução permanente da RNA-pol II: além de prevenir o bloqueio do ciclo celular, também evita a ocorrência de apoptose (PANI e NUDLER, 2017) (Figura 1b).

Após o reconhecimento das lesões por ambas as subvias, GG-NER e TC-NER convergem para um mecanismo em comum, que consiste no recrutamento, por um domínio de XPC, do fator de transcrição IIH (TFIIH) para o local da lesão (Figura 2c). O TFIIH é um complexo formado por 10 proteínas, dentre as quais estão XPB e XPD (SPIVAK, 2015; UCHIDA et al., 2002). A fim de formar o complexo de pré-incisão, as proteínas XPG, RPA e XPA também são recrutadas (COSTA et al., 2003). Nesse processo, o TFIIH realiza a abertura inicial da dupla-hélice, na qual a atividade de ATPase de XPB se combina com a função helicase de XPD. Em seguida, ocorre a ligação de XPG, RPA e XPA, levando à abertura total de aproximadamente 30 nucleotídeos ao redor da lesão (MULLENDERS, 2018). Ao passo que XPG está envolvida na estabilização da bolha de DNA formada com a abertura da dupla-hélice, XPA e RPA garantem que há dano no DNA, além de também auxiliarem a formar um complexo de pré-incisão mais estável (RASTOGI et al., 2010). Ademais, considerando

que RPA é uma proteína de ligação à fita simples, ela também atua na proteção da fita de DNA não danificada contra a ação de endonucleases (MARTEIJN et al., 2014).

Assim, antes que ocorra a incisão do dano, é necessário que o complexo de pré-incisão esteja montado e a lesão esteja posicionada corretamente (COSTA et al., 2003). Para tanto, a proteína RPA que está ligada à fita simples não lesionada ajuda a posicionar as endonucleases ERCC1-XPF e XPG na fita de DNA contendo o dano, na qual elas executam uma incisão dupla próximo às extremidades 5' e 3' da lesão, respectivamente (DE LAAT et al., 1998; MULLENDERS, 2018; SCHÄRER, 2013) (Figura 2d). Na sequência, o oligonucleotídeo contendo a lesão é removido com o TFIIH ligado a ele (KEMP et al., 2012). Então, uma vez que a incisão pelo complexo ERCC1-XPF gera uma extremidade 3'-OH livre, esta pode ser empregada como um *primer* para a ressíntese de DNA na lacuna gerada pela remoção da lesão (FAGBEMI; ORELLI; SCHÄRER, 2011). Desse modo, o novo fragmento de DNA é sintetizado pelas polimerases δ e ϵ , com o auxílio do antígeno nuclear de proliferação celular (PCNA), que regula a processividade da síntese (SCHUCH et al., 2017; SHIVJI; KENNY; WOOD, 1992) (Figura 2e) juntamente com o fator de replicação C (RFC). Por fim, o fragmento de DNA sintetizado é unido à fita original pela ação da DNA ligase I ou II (COSTA et al., 2003; MACHADO et al., 2014) (Figura 2f-g).

1.3 Aspectos funcionais da via NER em humanos e demais organismos

A relevância funcional do NER pode ser notoriamente observada através de síndromes e demais doenças hereditárias autossômicas causadas por mutações em genes dessa via (MENCK e MUNFORD, 2014). Dentre elas, estão o xeroderma pigmentoso, a síndrome de Cockayne, a tricotiodistrofia, a síndrome cérebro-óculo-fácio-esquelética e a síndrome de sensibilidade à UV (MULLENDERS, 2018). O xeroderma pigmentoso decorre majoritariamente de mutações nos genes XP (*XPA-XPG*), aumentando em quase dez mil vezes a incidência de câncer de pele em relação à média geral da população (BRADFORD et al., 2011). Além disso, essa doença genética também pode ser causada por mutações em genes não relacionados ao NER, como o gene da DNA polimerase η , a qual é responsável por prosseguir com a síntese de DNA apesar da presença de danos, evitando assim a morte celular (MORENO, SOUZA et al., 2020). Por outro lado, a síndrome de Cockayne está, na maioria dos casos, associada a mutações nos genes *CSA* (*ERCC8*) e/ou *CSB*

(ERCC6) (HANAWALT, 2000). Seus sintomas incluem fotossensibilidade, retardo mental, envelhecimento precoce e a incapacidade, em nível celular, de retomar a síntese de DNA após a indução de danos (FOUSTERI e MULLENDERS, 2008; VERMEULEN e FOUSTERI, 2013). Adicionalmente, a síndrome de sensibilidade à UV é caracterizada por mutações em CSA, CSB e UVSSA. Indivíduos afetados apresentam sensibilidade ao sol, pigmentação anormal na área exposta da pele, bem como sardas (SPIVAK, 2016; VERMEULEN e FOUSTERI, 2013).

Em um contexto evolutivo, a via NER está presente na história dos organismos desde arqueas (ROUILLON e WHITE, 2011), passando pelas bactérias com um mecanismo mais simplificado do que o observado em eucariotos, com o qual não apresenta homologia. No NER bacteriano, poucas proteínas são necessárias, dentre as quais estão as exonucleases UvrA, UvrB e UvrC e a helicase UvrD, que exercem as mesmas etapas básicas do NER em humanos (PETIT e SANCAR, 1999). Por outro lado, é interessante observar que, em certas espécies de eucariotos, essa via de reparo apresenta algumas diferenças em relação ao modo como ocorre em humanos. *Schizosaccharomyces pombe*, por exemplo, possui dois mecanismos distintos para o reparo de danos de UV: um NER clássico e um processo de reparo de excisão alternativo (UVER) (MCCREADY; OSMAN; YASUI, 2000). Já *Trypanosoma brucei* demonstrou a ausência da função de vários genes, sendo o reparo acoplado à transcrição a principal subvia relacionada ao NER para esses organismos (MACHADO et al., 2014). Pelo contrário, em *Plasmodium falciparum* foram encontrados homólogos para quase todos os componentes do NER, exceto para XPC e p62, que é uma proteína componente do TFIIH (TAJEDIN et al., 2015).

Além disso, eventos de duplicação de genes do NER já foram reportados em alguns organismos. Em plantas como *Arabidopsis thaliana*, foram encontradas duplicações em quatro genes: *RAD23*, *DDB1*, *CSA* e *XPB* (SPAMPINATO, 2017). Para ambas as cópias de CSA, porém, foram encontrados padrões de expressão sobrepostos, o que sugere uma sobreposição na função (ZHANG et al., 2010). Em relação às duas cópias de XPB, há indícios de que estejam passando por uma especialização funcional: enquanto um dos parálogos pode estar envolvido em reparo e proliferação celular, o outro está possivelmente associado ao reparo em células altamente especializadas (MASUDA et al., 2020). Sendo assim, é provável que tanto as cópias de CSA quanto de XPB tenham divergido há pouco tempo em *A. thaliana*.

Em adição a esses eventos, verificou-se que também há uma duplicação de XPB em *T. brucei*. No entanto, provavelmente trata-se de um acontecimento mais remoto, já que as duas proteínas têm funções distintas: enquanto uma das cópias está envolvida no reparo de DNA, a outra possivelmente desempenha uma função na transcrição (MACHADO et al., 2014).

Desse modo, a via NER tem uma dinâmica evolutiva complexa em muitas linhagens, com duplicações e aparentes perdas de funções em algumas proteínas da via para determinados organismos. Essa afirmativa difere da literatura a respeito da difundida informação de que o NER é extremamente conservado na árvore da vida, especialmente nos eucariotos (COSTA et al., 2003; RASTOGI et al., 2010). Assim, é de extrema importância para o entendimento de como ocorreu a evolução da via NER, que entendamos as suas diferenças entre as linhagens evolutivas, uma vez que há poucos estudos focados na história evolutiva de seus genes. Ademais, a maioria dos trabalhos envolvendo o NER está voltada para humanos, camundongos, leveduras e bactérias. O presente trabalho, por sua vez, inclui espécies de interesse médico, econômico e ecológico, o que torna ainda mais importante entender como esse sistema de reparo funciona nesses organismos.

Diante disso e levando em consideração a grande quantidade de dados de sequenciamento disponíveis em bancos de dados, é possível obter sequências de componentes dessa via de reparo de DNA para vários organismos modelo e não modelo, incluindo espécies nas quais o NER ainda apresenta aspectos não muito bem esclarecidos. Assim, através do uso de ferramentas de busca por similaridade, pôde-se obter uma noção estrutural e evolutiva com relação aos componentes da via. Ademais, a informação gerada neste trabalho poderá ser utilizada para futuros estudos do NER em organismos eucarióticos e, inclusive, a metodologia aplicada poderá ser empregada também para a investigação de outras vias de reparo de DNA.

Sendo assim, o objetivo geral deste trabalho consistiu em encontrar e caracterizar os potenciais ortólogos de componentes-chave da via NER humana em demais organismos eucarióticos (i) pesquisando, em bancos de dados, as sequências de aminoácidos de dez componentes-chave da via NER humana; (ii) buscando sequências através de *tblastn*, utilizando como isca proteínas humanas, e como alvos, os genomas de outros 12 organismos eucarióticos, a fim de obter suas sequências;

(iii) verificando o número de íntrons e o tamanho de cada gene nos diferentes organismos; (iv) identificando os domínios proteicos conservados nos diferentes organismos; e (v) realizando análises de reconstrução filogenética para cada proteína.

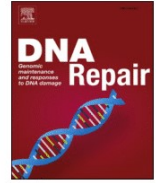
2 ARTIGO 1 – OPEN GAPS IN THE EVOLUTION OF THE EUKARYOTIC NUCLEOTIDE EXCISION REPAIR

Este estudo busca e caracteriza estruturalmente componentes centrais da via NER em organismos eucarióticos, trazendo diferenças importantes entre o NER eucariótico e o NER humano. Nesse sentido, são mostradas algumas lacunas na evolução do NER em eucariotos, junto a possíveis explicações em uma ótica evolutiva e funcional. Além disso, alguns testes de hipóteses são sugeridos como perspectivas para futuros estudos. O presente capítulo encontra-se publicado na revista científica *DNA Repair* (FELTRIN et al., 2020).



Contents lists available at ScienceDirect

DNA Repair

journal homepage: www.elsevier.com/locate/dnarepair

Open gaps in the evolution of the eukaryotic nucleotide excision repair

Rayana dos Santos Feltrin^{a,b}, Ana Lúcia Anversa Segatto^a, Tiago Antonio de Souza^c, André Passaglia Schuch^{a,b,*}^a Department of Biochemistry and Molecular Biology, Federal University of Santa Maria, RS, Brazil^b Postgraduate Program in Biological Sciences: Toxicological Biochemistry^c Institute of Biomedical Sciences, University of São Paulo, SP, Brazil

ARTICLE INFO

Keywords:

DNA repair
NER pathway
Eukaryotes
Gene structure
Domain architecture

ABSTRACT

Nucleotide excision repair (NER) is the most versatile DNA repair pathway as it removes different kinds of bulky lesions. Due to its essential role for genome integrity, it has appeared early in the evolution of species. However, most published studies are focused on humans, mice, yeast or bacteria. Considering the large amount of information on genome databases, it is currently possible to retrieve sequences from NER components in many organisms. Therefore, we have characterized the potential orthologs of 10 critical components of the human NER pathway in 12 eukaryotic species by using similarity and structural criteria through the use of bioinformatics tools. This approach has allowed us to characterize gene and protein structures comparatively, taking a glance at some evolutionary aspects of the NER pathway. We have obtained significant search results for the majority of the proteins in most of the organisms studied, mainly for factors that play a pivotal role in the pathway. However, we have revisited significant differences and found new aspects that may imply a distinct functioning of this pathway in different organisms. Through the demonstration of the heterogeneity of the gene structures and a variety in the protein architecture of the NER components evaluated, our results show important differences between human NER and evolutionarily distant eukaryotes. We highlight the lack of a canonical XPD in chicken, the divergence of XPA in plants and protozoans and the absence of XPE in the invertebrate species analyzed. In spite of this, it is remarkable the presence of this excision repair mechanism in a high number of evolutionary distant organisms, being present since the origin of eukaryotes.

1. Introduction

Genomes are continually being exposed to several genotoxic agents that can alter the DNA structure and compromise its function. These agents can be both endogenous, coming from cellular metabolism, and exogenous, involving the interaction with the environment. The former comprises reactive oxygen and nitrogen species, whereas the latter encompasses ultraviolet (UV) radiation, ionizing radiation and harmful chemicals [1,2]. Moreover, some DNA lesions can also be induced spontaneously, through deamination, depurination or depyrimidination [3]. Nevertheless, a variety of DNA repair mechanisms have developed throughout the evolution of species to ensure the maintenance of genome integrity [4]. Among them, the nucleotide excision repair (NER) pathway is the most versatile DNA repair mechanism as it corrects a wide range of bulky lesions that thermodynamically destabilize the double helix. UV photoproducts, such as cyclobutane pyrimidine dimers

(CPDs) and 6–4 pyrimidine-pyrimidone photoproducts (6–4PPs), adducts induced by environmental mutagens, such as benzo[a]pyrene, as well as damage formed by chemotherapeutic drugs are some lesions commonly repaired by NER [3,5].

1.1. General NER mechanism

NER occurs in four main steps: damage recognition, excision, DNA resynthesis, and strand ligation [5,6]. However, the efficiency of lesion removal is different depending on its location in the genome. Base adducts situated in actively transcribed genes are repaired faster than in inactive genes, which indicates the existence of two distinct processes of damage response [7]. Therefore, NER can be divided into two sub-pathways: global genome NER (GG-NER) and transcription-coupled NER (TC-NER). The first one is associated with a progressive screening of the entire genome for lesions to be repaired, and the second one refers

* Corresponding author at: Av. Roraima, 1000, P.O. Box 5021, room 3010, Camobi, Santa Maria, RS, 97110-970, Brazil.
E-mail address: andre.schuch@ufsm.br (A.P. Schuch).

<https://doi.org/10.1016/j.dnarep.2020.102955>

Received 8 January 2020; Received in revised form 6 August 2020; Accepted 16 August 2020

Available online 23 August 2020

1568-7864/© 2020 Elsevier B.V. All rights reserved.

to a preferential repair of transcription-blocking lesions located in transcribed DNA strands [8]. Concerning the genes involved in NER, there are two main groups: XP genes - whose mutations cause Xeroderma Pigmentosum disease, and CS genes - whose mutations result in Cockayne Syndrome, being both related to photosensitivity.

1.1.1. Global genome NER (GG-NER)

In humans, GG-NER initiates with direct recognition of the lesion by XPC-hHR23B complex, with the aid of CETN2 protein [9]. HHR23B may be involved in XPC stabilization, and CETN2 is known to improve the damage recognition function of XPC [6]. Conversely, adducts that subtly distort the DNA duplex, such as CPDs, are previously recognized by the DDB complex, formed by DDB1 and DDB2 (XPE) proteins. The interaction with XPC-hHR23B complex recruits the transcription initiation factor IIIH (TFIIH), which is composed of 10 proteins, including the 3' and 5' helicases XPB and XPD [8,9]. The engagement of XPD in the damage site allows the assembly of the pre-incision complex, constituted of XPA, RPA, and XPG proteins [3,9]. XPA provides protein complex stability, as well as XPG ensures TFIIH structural support. In turn, RPA protects the single-stranded DNA from endonuclease action. XPA is also responsible for the recruitment of the ERCC1-XPF complex, which makes the incision 5' of the lesion, followed by the action of XPG, a 3' endonuclease [1,9]. Then, TFIIH is removed, and the excised DNA fragment is resynthesized by polymerases δ and ϵ , helped by PCNA. Finally, the ends are joined by ligase I or III [5].

1.1.2. Transcription-coupled NER (TC-NER)

TC-NER assumes great importance in maintaining the flow of genetic information, as it removes DNA lesions that block the translocation of RNA-polymerase II (RNA-pol II) along the template strand [10]. Consequently, it plays a fundamental role in preventing cell cycle arrest and apoptosis by avoiding the permanent obstruction of the polymerase [11]. Furthermore, recent knowledge about the extent of pervasive transcription demonstrates, even more, the importance of RNA-pol II and TC-NER in DNA repair [12]. The TC-NER mechanism starts with the stalling of RNA-pol II, which takes part in lesion recognition [6]. Therefore, the human ERCC6 (CSB) protein, which is a DNA-dependent ATPase that moves together with RNA-pol II, strongly binds to the polymerase, altering the DNA conformation [9]. In turn, ERCC8 (CSA) is a binding protein containing several WD40 repeats from binding domains that is essential to signal transduction (scaffold protein). This protein is a member of the CRL4CSA ubiquitin complex, composed of Cullin4-DDB1-RING enzymes [13].

In order to ensure the resumption of transcription, CSA directs CSB ubiquitination, leading to its proteasomal degradation. This is followed by the removal of the transcription machinery that is blocked at the lesion site, as well as the recruitment and assembly of other factors needed to proceed with damage excision [6,14]. In this process, two essential proteins are involved, UVSSA and USP7, which act in concert to protect CSB from the proteasomal degradation. Paradoxically, this assures that CSB will last time enough to perform its function in TC-NER. After the arrest of RNA-pol II is solved, the TC-NER complex is destabilized, proceeding to CSB degradation [15]. The release of RNA-pol II from the damage strand also involves the action of UVSSA, allowing the recovery of a fast repair activity [16]. A mutation in this gene is associated with the UV sensitivity syndrome, characterized by photosensitivity and some mild skin abnormalities [17].

1.2. Research interests

In an evolutionary context, NER is in the tree of life since the bacterial and archaeal domains [18,19]. Some NER proteins from archaea are probably homologous to eukaryotic NER proteins, even though the NER pathway from bacteria and eukaryotes appear to have separate origins [20]. Probably, these repair mechanisms are as old as life on Earth [21]. However, despite the knowledge about the functions of the

enzymes from this pathway and that it is well conserved in eukaryotes [8], there are few current studies concerning the evolutionary history of this DNA repair pathway. Additionally, most studies are focused on humans, mice, yeast or bacteria. Thus, considering the large number of available sequences from eukaryotic species on genome databases, this study aims to find potential orthologs of 10 key components of the human NER pathway in other twelve eukaryotic organisms by using similarity and structural criteria. Consequently, this work demonstrates the characterization of their protein architectures and gene structures in a comparative approach, taking a glance at some evolutionary aspects of the eukaryotic NER pathway.

2. Material and methods

2.1. Search for DNA and protein sequences

In order to get the query sequences for BLAST searches, the bait "protein symbol + *Homo sapiens*" has been used on GenBank Protein section [22], considering the RefSeq database [23]. In this step, 10 key proteins of the NER pathway have been searched: CSA, CSB, hHR23B (homologue to RAD23B from yeast), XPA, XPB, XPC, XPD, XPE (DDB2), XPF, and XPG. The resulting sequences corresponding to the longest isoform from each protein, whose coding sequences have had the highest exon number, have been employed as queries for individual *tblastn* searches [24] in the genomes of *Mus musculus*, *Danio rerio*, *Xenopus laevis*, *Alligator mississippiensis*, *Gallus gallus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, *Trypanosoma cruzi*, and *Giardia lamblia*. Genome assembly codes are available in Table S1.

2.2. Defined criteria for the homology search

Firstly we have defined four filtering criteria to select BLAST homologue sequence candidates: *e-value*¹, percent identity, protein sequence length, and conserved protein domains in common. The first criterion filters as statistically significant only hits with *e-value* lower than 1×10^{-6} [25]. Then, sequences have been selected based on the similarity of annotated features from the resulting hit to human query sequences, aligning them on *Molecular Evolutionary Genetic Analysis 7 (MEGA7)* software [26] using MUSCLE [27]. We have chosen the sequences with the highest similarity and the closest length to the query. If the resulting proteins have been similar, even in the amino acid number, we have compared the transcript sequences using the same criteria. Thus, the transcript and protein sequences corresponding to the selected genomic sequences have been retrieved.

The second homology criterion chooses the hits with higher percent identities. If the chosen hit had been the second one, we have used the third criterion to compare protein sequence lengths from both hits with the length of the query. Then, we have opted for the hit whose protein length was more similar to its query. Lengths have been obtained from the respective protein accession numbers on GenBank. Also, when there was no annotated feature for a genomic sequence, we have used it as a query to a *blastx* search [24] directed to the organism of interest. All BLAST results and retrieved sequences can be accessed in Table S2.

The fourth homology criterion is based on protein conserved domains. We have used the normal mode of *Simple Modular Architecture Research Tool (SMART)* [28] to identify protein conserved domains, considering PFAM database [29]. The analyzes have been made using selected sequences of each target protein.

When no hit had been obtained in a given search or no homologue candidate had been found following our criteria, we have performed a refined search to confirm if this NER component is in fact not found in a

¹ The *e-value* is a statistic measure that indicates the expectation of obtaining the same search result by chance on a database with the same size.

certain species. To do that, we have used the closest species as query to carry out both *tblastn* and *blastx* searches. If some retrieved sequence had been in agreement with our search criteria, then we have included it in our analyzes. The results of this refined search, as well as the corresponding sequences, are available in Table S3.

2.3. Estimates of evolutionary distance and phylogenetic analyzes

To estimate the evolutionary distance among the retrieved sequences, first we have aligned the corresponding sequences from each protein on MEGA7 software [26] using MUSCLE [27]. Conserved regions from the resulting alignments have been obtained with Gblocks 0.91b [30], allowing half of the gaps for all proteins except CSA, for which all gaps have been allowed, and CSB, with no gaps (default parameter). Thereafter, we have evaluated the results on ProtTest 3.4.2 [31] to set the best evolutionary models. The evolutionary distance analysis has been run on MEGA7 using the substitution model JTT + G (number of gamma categories = 4), 1000 bootstrap replicates, and pairwise deletion. Here we have considered only the evolutionary distances to *H. sapiens*. To test the homology hypothesis among the sequences retrieved for each protein, we have conducted the phylogenetic analyzes with the same parameters from the evolutionary distance analysis by using the Neighbor-Joining method.

2.4. Phylogenetic reconstruction of the eukaryotic species tree

To compare the phylogenetic trees of NER proteins with a species tree, we have used the mitochondrial molecular marker 18S rRNA to infer the phylogeny of the eukaryotic species analyzed in this work. To

do that, we have employed sequences retrieved by name from GenBank, whose accession numbers are available in Table S4. These sequences have been aligned on MEGA7 software [26] using MUSCLE [27], and the resulting alignment has been submitted to Gblocks 0.91b (default parameters) to get the conserved regions [30]. We have predicted the best evolutionary model with jModelTest v2.1.10 [32], which have resulted in TrN + G as the best-fit model according to the Bayesian Information Criterion (BIC). This has also been the best model among the ones that are available on MEGA7 for distance estimation when the evaluation has been based on the Akaike Information Criterion (AIC). Thus the phylogenetic reconstruction has been performed as described in item 2.3, but rather using the TrN + G model.

2.5. Analysis of gene structure and length

Additional analyzes have been performed to investigate the structure of NER genes based on intronic regions. The intron numbers have been inferred by submitting the genomic and transcript sequences of each gene to *Gene Structure Display Server 2.0 (GSDS 2.0)* [33]. All gene lengths have been obtained from GenBank using their respective GeneID.

3. Results

3.1. BLAST results and evolutionary distances

Although the majority of the resulting sequences from the first search, which have used human sequences as queries, are statistically significant (Table 1), some sequences have presented non-significant *e*-

Table 1

e-value threshold and evolutionary distance from human sequences regarding the best homologue candidates to NER proteins from both the first search and the refined search.

Species	CSA	CSB	HHR23B	XPA	XPB	XPC	XPD	XPE	XPF	XPG
<i>Mmu</i>	0.046 (1E-27)	0.011 (1E-105)	0.005 (3E-67)	0.102 (2E-43)	0.019 (3E-98)	0.108 (2E-112)	0.019 (5E-140)	0.16 (9E-64)	0.066 (0.0)	0.08 (7E-60)
<i>Gga</i>	0.069 (1E-26)	0.039 (2E-36)	0.01 (2E-19)	0.244 (9E-24)	0.045 (6E-58)	0.174 (3E-44)	2.402 (2E-03)	0.53 (1E-18)	0.101 (2E-26)	0.175 (9E-35)
<i>Ami</i>	0.046 (2E-25)	0.062 (7E-38)	0.005 (2E-18)	0.205 (1E-23)	0.037 (2E-58)	0.158 (6E-42)	0.103 (1E-44)	0.428 (6E-19)	0.096 (1E-128)	0.148 (5E-34)
<i>Xla</i>	0.106 (0.0)	0.091 (3E-30)	0.037 (3E-18)	0.253 (4E-21)	0.046 (8E-59)	0.212 (7E-29)	0.18 (5E-45)	0.529 (2E-45)	0.155 (3E-124)	0.166 (3E-32)
<i>Dre</i>	0.167 (6E-26)	0.075 (1E-93)	0.193 (2E-20)	0.378 (2E-19)	0.059 (2E-58)	0.36 (9E-44)	0.146 (9E-89)	0.788 (8E-42)	0.284 (3E-97)	0.32 (4E-29)
<i>Cel</i>	2.577 (4E-07)	94.105 (1E-78)	0.882 (4E-22)	0.984 (1E-28)	0.226 (8E-110)	1.102 (1E-15)	0.521 (1E-37)	NS (1E-06)	1.504 (1E-39)	1.204 (2E-14)
<i>Dme</i>	2.26 (3E-12)	1.383 (2E-74)	0.674 (4E-10)	0.699 (5E-54)	0.201 (0.0)	0.782 (8E-81)	0.385 (5E-123)	NS (2E-06)	0.782 (4E-95)	0.668 (4E-67)
<i>Ath</i>	0.753 (3E-23)	0.486 (3E-165)	0.807 (1E-08)	–	0.468 (3E-38)	1.266 (2E-07)	0.653 (4E-82)	1.743 (7E-10)	0.907 (5E-65)	0.882 (1E-21)
<i>Sce</i>	1.24 (1E-13)	0.475 (0.0)	1.227 (4E-22)	1.862 (4E-11)	0.432 (0.0)	1.894 (8E-26)	0.676 (0.0)	NS (8E-06)	1.18 (8E-64)	1.093 (6E-48)
<i>Spo</i>	0.771 (2E-47)	0.414 (0.0)	0.718 (2E-14)	1.437 (1E-15)	0.474 (0.0)	1.603 (2E-36)	0.63 (0.0)	NS (1E-06)	1.021 (2E-66)	0.887 (5E-59)
<i>Tcr</i>	1.54 (6E-10)	0.721 (5E-128)	1.369 (2E-07)	NS (5.9E-01)	1.124 (6E-107)	1.816 (2E-25)	1.003 (0.0)	NS (2.1E-02)	1.816 (8E-30)	1.66 (3E-18)
<i>Gla</i>	3.117 (2E-07)	1.25 (2E-90)	–	–	1.261 (5E-82)	–	1.708 (5E-81)	NS (1E-03)	2.595 (4.6E-01)	1.831 (7E-09)



The evolutionary distance values are displayed only for sequences whose hits have had significant *e*-values (< 1E-06). The *e*-values corresponding to each sequence are between parentheses, in blue. Numbers in bold refer to sequences resulting from the refined searches. CSB *Cel* has a very discrepant value, then it has remained out of the color range and is displayed in red. NS indicates hits with nonsignificant *e*-values, whereas the hyphen represents searches which have resulted in no hit. *Hsa*: *Homo sapiens*, *Mmu*: *Mus musculus*, *Gga*: *Gallus gallus*, *Ami*: *Alligator mississippiensis*, *Xla*: *Xenopus laevis*, *Dre*: *Danio rerio*, *Cel*: *Caenorhabditis elegans*, *Dme*: *Drosophila melanogaster*, *Ath*: *Arabidopsis thaliana*, *Sce*: *Saccharomyces cerevisiae*, *Spo*: *Schizosaccharomyces pombe*, *Tcr*: *Trypanosoma cruzi*, *Gla*: *Giardia lamblia*.

values, such as XPA of *T. cruzi*, XPD of *G. gallus*, XPE of invertebrates, and XPF of *G. lamblia*. Also, we have found no hits for XPA in *A. thaliana*, neither for hHR23B, XPA, and XPC in *G. lamblia*. In almost all cases, higher percent of protein identities ($75\% \pm 17\%$) can be observed among vertebrates, whereas invertebrates and plants present lower identities to human queries ($40\% \pm 11\%$) (Table S2). Conversely, we have observed that the helicases XPB and XPD maintain relatively high identities in all analyzed species. For the helicases, we have observed identities from 96% (*M. musculus* and *A. mississippiensis*) to 35% (*T. cruzi*) in XPB, and 81% (*A. mississippiensis*) to 26% (*G. lamblia*) in XPD. For the endonucleases, we have reported identities from 84% (*M. musculus*) to 27% (*T. cruzi*) in XPF, and 93% to 30% in XPG (Table S2).

Refined searches have been performed using queries from closer species to verify if the searches that resulted in no hit or no homologue candidate do refer to the absence of the NER component in a certain species (Table S3). Through these additional searches, it has been possible to find better homologue candidates for CSA of *T. cruzi*, as well as CSB of *D. rerio* and *C. elegans* (Table 1). We also have found a putative sequence with significant *e-value* for XPD of *G. gallus*, which is annotated as an ATP-dependent DNA helicase DDX11. A homologue candidate for XPF of *G. lamblia* has been obtained as well. To find it, we have submitted the genomic sequence without annotation retrieved in the *tblastn* search to ORFfinder [34]. Thus, the longest resulting ORF has been used as query to a *blastp* search in *G. lamblia* and, as a result, we have obtained XPF. Since we have not found a corresponding annotated CDS, ORFfinder has been also used to carry out this prediction.

Regarding the evolutionary distances to *H. sapiens* (Table 1), as expected, we have reported that the smallest values are generally related to vertebrate species, except for the putative XPD sequence of *G. gallus*, which have presented a more discrepant distance (2.402). In contrast, the most divergent sequences have resulted from invertebrates, mainly unicellular species. Very larger distances have been obtained for the sequences from protozoan in all the analyzed proteins, except for XPE, in which no homologous has been found. In the yeast species, this also has occurred with CSA (1.24), hHR23B (1.227), XPA (1.862), XPC (1.894),

and XPF (1.18) of *S. cerevisiae*, as well as with XPA (1.437) and XPC (1.603) of *S. pombe*. However, we also have found greater distances for CSA (2.577), CSB (94.105), XPF (1.504), and XPG (1.204) of *C. elegans*, CSA (2.26) and CSB (1.383) of *D. melanogaster*, as well as XPC (1.266) and XPE (1.743) of *A. thaliana*.

3.2. Phylogenies and conserved protein domains

The phylogenies of NER proteins, in general, have resulted in a well-supported clade formed by vertebrate proteins, whereas the positions of the other species could not be determined.

The CSA protein sequences present WD40 domains that are similar to all the analyzed eukaryotic species (Fig. 1A). We have observed five WD40 domain units with very similar size and positions among vertebrate species. Also, *A. thaliana* presents a domain pattern that resembles vertebrates, with a mild variation in the last two domains. Conversely, some species show more striking differences, such as *D. melanogaster*, which has six WD40 domains and a different N-terminal domain named CAF 1C_H4-bd. The CSA putative homologue of *C. elegans* has seven WD40 domains and a Ubox domain in the N-terminus. In addition, we have detected more domains in *T. cruzi*, which presents seven WD40 units. A different number of domains has also been observed in the two kinds of yeast: *S. cerevisiae* has four WD40 domains (three of their units display similar positions to vertebrates), whereas *S. pombe* presents six WD40 domains, being five of them with similar positions to vertebrates. *G. lamblia* also has six N-terminal WD40 domains, and other two C-terminal domains, named Coatomer_WDAD and COPI_C. Regarding the phylogenetic trees, in comparison with the species phylogeny (Figure S1), the CSA tree clusters *A. thaliana* and both yeasts species are clustered closer to vertebrates.

A pair of domains, DEXDc and HELICc, has been found in all CSB protein sequences (Fig. 1B). Positions and size of these domains are quite similar among amniotes. However, we have observed distinct domains in fruit fly, such as one DBINO domain in the N-terminus, and two SANT domains near the C-terminus. *G. lamblia* presents the same CSB typical

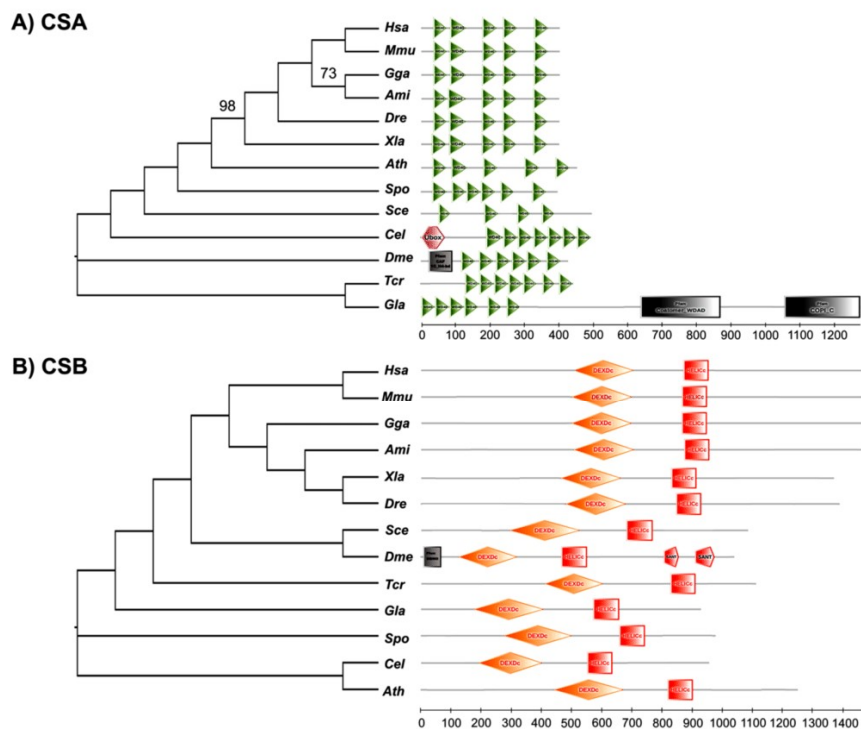


Fig. 1. Phylogenetic trees and protein conserved domains of TC-NER damage recognition proteins in the analyzed eukaryotic species. **A)** CSA. **B)** CSB. Domains have been obtained from the SMART server, and trees have been reconstructed using the Neighbor-Joining method on MEGA7. Only bootstrap supports above 70 are displayed. Hsa: *Homo sapiens*, Mmu: *Mus musculus*, Gga: *Gallus gallus*, Ami: *Alligator mississippiensis*, Xtr: *Xenopus tropicalis*, Xla: *Xenopus laevis*, Dre: *Danio rerio*, Cel: *Caenorhabditis elegans*, Dme: *Drosophila melanogaster*, Ath: *Arabidopsis thaliana*, Sce: *Saccharomyces cerevisiae*, Spo: *Schizosaccharomyces pombe*, Tcr: *Trypanosoma cruzi*, Gla: *Giardia lamblia*.

domains, with an approximated position to those of *C. elegans*. Regarding the CSB phylogeny, we have obtained very low bootstrap values supporting the branches. In comparison with the species tree (Figure S1), we have observed that in CSB tree, the vertebrate clade is clustered with a branch that groups *S. cerevisiae* and *D. melanogaster*. Additionally, *T. cruzi* and *G. lamblia* are closer to the vertebrate-containing clade, and *S. pombe* remains in a separate branch, whereas *C. elegans* and *A. thaliana* form a new clade.

In XPC protein sequences, four domains have been found (Rad4, BHD_1, BHD_2, and BHD_3), whose number, size, and positions are similar among amniotes (Fig. 2A). However, the plant species also displays an additional Transglut_core domain. Furthermore, we have observed a considerable divergence in the domain positions of *D. melanogaster*, followed by *C. elegans*. In addition, both yeast species have similar domain positions. Concerning the phylogenetic analyzes of the species (Figure S1), there are two new clades: one formed by *A. thaliana* and *T. cruzi*, and the other one, composed only by *S. cerevisiae* and *S. pombe*.

The hHR23B protein sequences have three specific domains: UBQ, UBA (two units), and STI1 (Fig. 2B). Their positions, number, and size are very similar among vertebrates, except for *D. rerio*, whose two UBA domains and also STI1 have shuffled positions. Also, hHR23B of *A. mississippiensis* has a slightly smaller UBQ domain. Furthermore, it is noteworthy that hHR23B of *D. melanogaster* displays a divergent protein structure due to the absence of the UBQ domain in the N-terminus, as well as the other domains present different positions. The plant and the unicellular species share approximately similar structures. Conversely, *T. cruzi* has an XPC_binding domain in place of STI1. Regarding the species phylogeny (Figure S1), in hHR23B tree, *D. melanogaster* and *A. thaliana* group with vertebrates. Moreover, there are two new clades, one formed by *C. elegans* and *T. cruzi*, and the other one, by *S. cerevisiae* and *S. pombe*.

The XPE protein sequences are characterized by WD40 domains,

from four to five units (Fig. 2C). We have reported that the two analyzed mammal species are composed by five WD40 domains at similar positions. From alligator to chicken, these vertebrate species present four domain units that share relatively approximated positions. The exception is the fourth domain of *D. rerio*, which has a larger shift in relation to the position of the third one. Similarly, *A. thaliana* has five more spaced WD40 domains, besides a ZnF_C2HC domain in the N-terminus. In comparison with the species tree (Figure S1), the XPE phylogeny has also a vertebrate clade, besides a branch formed only by *A. thaliana*.

The XPB protein sequences are characterized by the Helicase_C_3, DEXDc, and HELICc domains (Fig. 3A). In general, the protein architectures are highly similar among all species, notably among multicellular species. In unicellular species, we have observed a larger shift in the domain positions of the trypanosome, which differs from similar positions observed in both yeast species. Moreover, *G. lamblia* also shows a shift in the position of the Helicase_C_3 domain, but not for DEXDc and HELICc, which share similar positions to yeast. Compared to the species phylogeny (Figure S1), the XPB tree topology is quite similar.

The XPD protein sequences present two domains, DEXDc and HELICc, whose number, size, and positions are extremely similar among the species in which this protein has been retrieved (Fig. 3B). Only *G. lamblia* displays a domain structure slightly shifted to the right, besides a larger DEXDc domain. Regarding the species tree (Figure S1), in the XPD phylogeny *D. melanogaster* and *C. elegans* are in separate but close clades. Also, the branch containing *G. gallus* remains in a position that is closer to protozoans. In the sequence alignment, this protein is discordant (data not shown). In turn, the clade formed by plant and yeasts in the species tree is splitted in a branch formed by both yeast species, whereas *A. thaliana* remains in a separate but adjacent branch.

We have found two domains in the XPA protein sequences, XPA_N and XPA_C, related to the N- and C-terminal regions, respectively (Fig. 3C). Their positions, number, and size are also quite similar among vertebrates, except for *A. mississippiensis*, whose XPA_N position varies

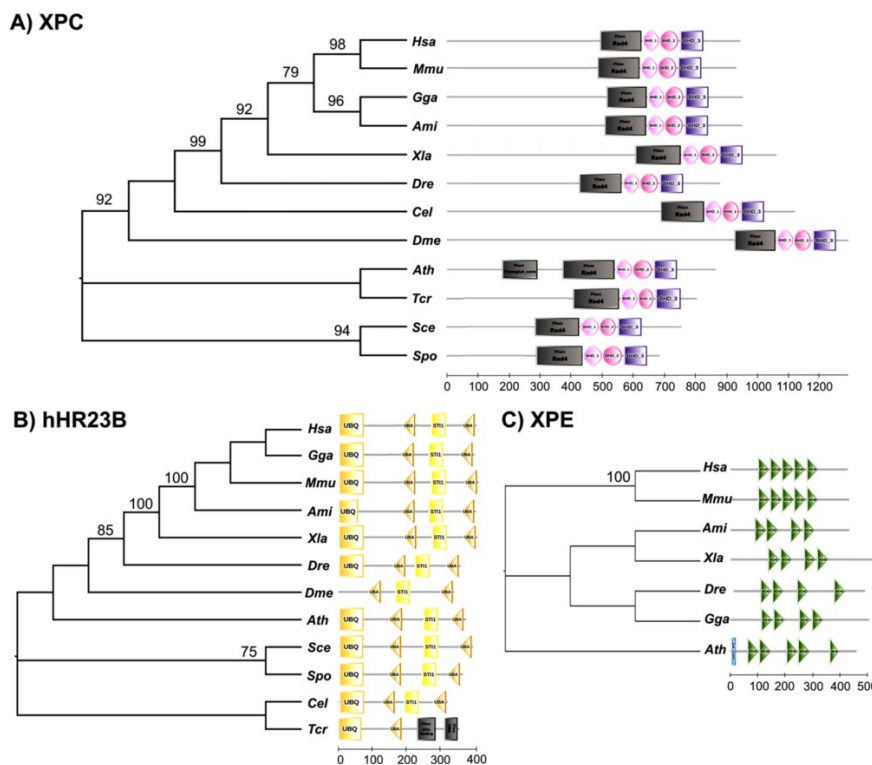


Fig. 2. Phylogenetic trees and protein conserved domains of GG-NER damage recognition proteins in the analyzed eukaryotic species. A) XPC. B) hHR23B. C) XPE. Domains have been obtained from the SMART server, and trees have been reconstructed using the Neighbor-Joining method on MEGA7. Only bootstrap supports above 70 are displayed. Hsa: *Homo sapiens*, Mmu: *Mus musculus*, Gga: *Gallus gallus*, Ami: *Alligator mississippiensis*, Xtr: *Xenopus tropicalis*, Xla: *Xenopus laevis*, Dre: *Danio rerio*, Cel: *Caenorhabditis elegans*, Dme: *Drosophila melanogaster*, Ath: *Arabidopsis thaliana*, Sce: *Saccharomyces cerevisiae*, Spo: *Schizosaccharomyces pombe*, Tcr: *Trypanosoma cruzi*, Gla: *Giardia lamblia*.

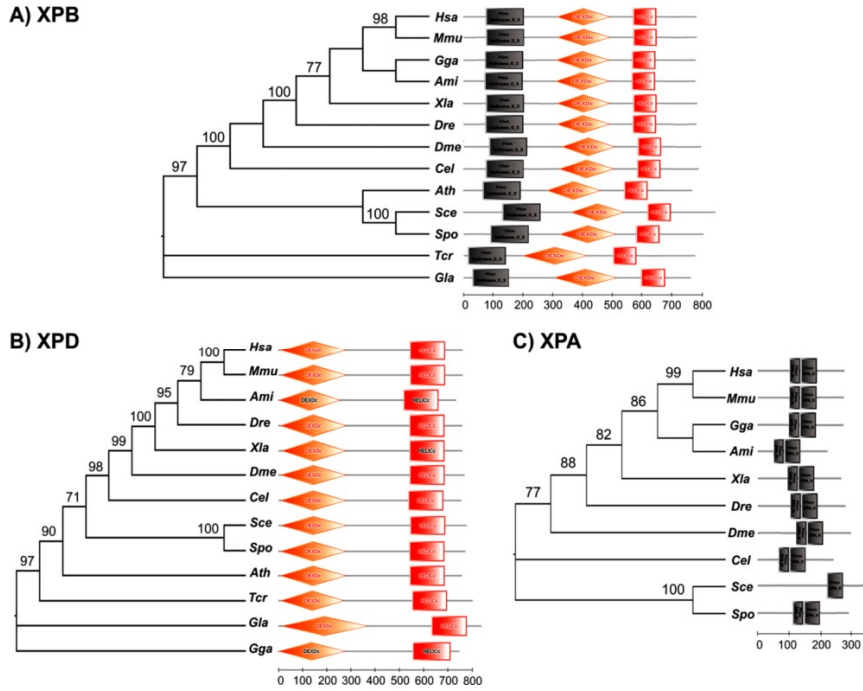


Fig. 3. Phylogenetic trees and protein conserved domains of TFIIH and opening of damaged region related proteins. **A)** XPB. **B)** XPD. **C)** XPA. Domains have been obtained from the SMART server, and trees have been reconstructed using the Neighbor-Joining method on MEGA7. Only bootstrap supports above 70 are displayed. *Hsa*: Homo sapiens, *Mmu*: Mus musculus, *Gga*: Gallus gallus, *Ami*: Alligator mississippiensis, *Xtr*: Xenopus tropicalis, *Xla*: Xenopus laevis, *Dre*: Danio rerio, *Cel*: Caenorhabditis elegans, *Dme*: Drosophila melanogaster, *Ath*: Arabidopsis thaliana, *Sce*: Saccharomyces cerevisiae, *Spo*: Schizosaccharomyces pombe, *Tcr*: Trypanosoma cruzi, *Gla*: Giardia lamblia.

towards the N-terminus. In contrast, *S. cerevisiae* presents a distinctive structure, since the XPA_N domain has been not detected and XPA_C has a different position compared to the other analyzed species. In relation to the species phylogeny (Figure S1), the XPA tree has separate branches for *D. melanogaster* and *C. elegans*, besides a clade that clusters both yeast

species.

Regarding XPF protein sequences, only an ERCC4 domain was observed in all analyzed species (Fig. 4A). Similar size, number, and positions are observed among the multicellular species, whereas unicellular organisms have more divergent positions. Concerning the

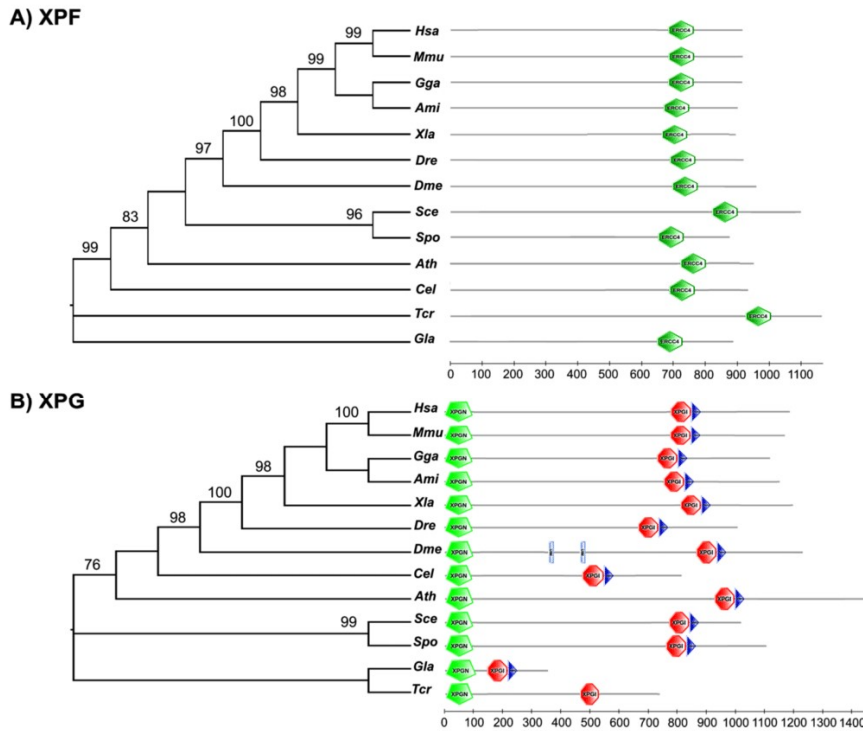


Fig. 4. Phylogenetic trees and protein conserved domains of the endonucleases. **A)** XPF. **B)** XPG. Domains have been obtained from the SMART server, and trees have been reconstructed using the Neighbor-Joining method on MEGA7. Only bootstrap supports above 70 are displayed. *Hsa*: Homo sapiens, *Mmu*: Mus musculus, *Gga*: Gallus gallus, *Ami*: Alligator mississippiensis, *Xtr*: Xenopus tropicalis, *Xla*: Xenopus laevis, *Dre*: Danio rerio, *Cel*: Caenorhabditis elegans, *Dme*: Drosophila melanogaster, *Ath*: Arabidopsis thaliana, *Sce*: Saccharomyces cerevisiae, *Spo*: Schizosaccharomyces pombe, *Tcr*: Trypanosoma cruzi, *Gla*: Giardia lamblia.

species tree (Figure S1), in the XPF phylogeny there was a split of the clade composed by *C. elegans* and *D. melanogaster*, and the former is closer to *A. thaliana*. In turn, this plant species is not grouped with yeasts anymore.

In XPG protein sequences, three domains were found: XPGN right in the N-terminus, XPGI, and HhH2 near the middle of the protein towards the C-terminus (Fig. 4B). The number, size, and positions of XPGN domains are identical among all species analyzed. Moreover, we have verified that XPGI and HhH2 have similar structural aspects between *H. sapiens* and *M. musculus*, as well as in *G. gallus* and *A. mississippiensis*. Different from the other species, we have detected two UIM domains in the protein structure of *D. melanogaster*, located between XPGN and the two other domains. Furthermore, we have observed a large variation in the domain positions in *C. elegans*, *G. lamblia*, and *T. cruzi*. This may be attributed to the smaller size of these proteins. Interestingly, the HhH2 domain has not been detected in the trypanosomatid. Concerning the yeast species, both share similar positions between themselves. The XPG tree, compared to the species phylogeny, has two new clades: one formed only by the yeast species, and the other one, by the protozoan species. Also, *A. thaliana* remains in a separate branch.

3.3. Intron number and gene length

Generally, we have found that vertebrates present a higher and similar intron number, although CSA of *X. laevis* has only one intron (Figure S2). Conversely, invertebrates present a lower intron number, whereas unicellular species generally have few or even no intron. In almost all cases, *A. thaliana* exhibits an intermediate intron density between vertebrates and invertebrates. Regarding gene length, we have observed that the longest genes are those from mammalian, followed by the other analyzed vertebrates. An exception has been verified in *XPE* of *G. gallus* (Figure S2, H), which presents the longest length for this gene. Furthermore, we have found that *A. thaliana* commonly has an intermediate gene size, and invertebrates have smaller genes, chiefly unicellular organisms.

4. Discussion

From archaea and bacteria to eukaryotes, NER has the same overall scheme of DNA lesion removal [20]. In eukaryotes, NER proteins are strongly conserved [1], as we have showed in our results regarding the significant findings of potential homologues for the majority of proteins searched in the eukaryotic organisms. Notably, the XPB helicase and the endonucleases XPF and XPG are found in all species evaluated in this study. Actually, some of the NER proteins are widespread, and they probably represent ancient proteins that could have taken part in a primitive NER mechanism [5]. Although NER proteins are well conserved, there are essential differences among distant phylogenetic groups, which may imply a distinct functioning of this pathway in different organisms [1,5,35]. Besides substantial evolutionary distances, non-significant *e-values* might suggest the nonoccurrence of the respective proteins or large sequence divergences. However, it is important to mention that this could be also associated with annotation errors that could hinder the homology search [35] or even false negatives [25].

We have found some constraints on sequences that have significant *e-values*, but are probably not analogous to the human proteins due to aspects related to conserved domains, length, as well as to the evolutionary analyzes. CSA is a protein organized in five WD40 repeats [5], essential structures that work as platforms in protein-protein and DNA-protein interactions [36]. We have reported that this architecture has been evolutionarily maintained in vertebrates, but substantial differences have been encountered in some invertebrates (Fig. 1A). For instance, *D. melanogaster* presents six WD40 repeats and a chromatin assembly factor-related domain, which suggests a distinct function from human CSA. Furthermore, its sequence has a certain evolutionary distance (2.26) from *H. sapiens*. Indeed, the fruit fly genome does not

encode a potential counterpart for CSA [37], and curiously this gene seems to be missing even from Holometabola, that is, insects that undergo complete metamorphosis [38,39].

Caenorhabditis elegans also comprises a different structure of seven WD40 domains shifted to the right, besides a U-box domain (Fig. 1A), probably associated with a ubiquitin-ligase function [40]. Also, the protein sequence retrieved for CSA in *C. elegans* has a distance of 2.577 from human CSA and a quite bigger length than the observed average. These characteristics strongly suggest that, possibly, it is not analogous to CSA. However, an analogous protein has been found for CSA in *C. elegans* through a sequence profile made from a multiple sequence alignment of established CSA sequences, a more effective methodology to find distant homologues [41,42].

Additionally, we have observed that the protein sequence obtained for CSA of *S. cerevisiae* has resulted in only four WD40 domains, and three of them have similar positions to vertebrates (Fig. 1A). Moreover, this sequence has a distance of 1.24 from human CSA and a much bigger length than the average verified for this protein, even though it has clustered in the same clade containing the vertebrate species. Actually, a study has reported that a homologue of CSA in budding yeast has been identified with five WD40 repeats. It has been demonstrated that three of these WD40 repeats are structurally related to repeats in the human CSA [43]. However, the fifth repeat have had the lowest percent identity with its human correspondent [43], which might explain why this have not been detected with our approach. Notwithstanding, although a CSA homologue has been detected in *S. cerevisiae*, it is not in the TC-NER of this organism [20,44].

In the corresponding hit for CSA of *S. pombe*, we have detected six WD40 domains, and five share similar positions with vertebrates (Fig. 1A). In fact, it has been demonstrated that there is a CSA homologue in *S. pombe*, termed Ckn1, which is involved in TC-NER, but does not seem to play a role in the ubiquitination of the CSB counterpart [45]. Regarding TC-NER in budding yeast, it is known that its CSB homologue, Rad26, is entirely or partially dispensable in the presence of a subunit of the RNA-pol II [46,47]. However, TC-NER in *S. cerevisiae* is primarily executed by the Rad26-dependent mechanism, with the exception of a small region immediately downstream of the transcription start site [47, 48]. Therefore, we can infer that TC-NER exists since the origin of eukaryotes. However, the genes involved in this process may have lost and gained functions during the evolution, which can be supported by the variability of gene structure and domains (Figs. 1 and S2). This implies that, in some organisms, homologous proteins may not participate in TC-NER and then analogous proteins might perform these functions. Nonetheless, functional studies, as well as studies of the molecular evolution of each gene family, are necessary to corroborate this hypothesis.

Furthermore, in the hit obtained for CSA of *T. cruzi*, we have verified a different domain frequency corresponding to seven WD40 repeats (Fig. 1A), besides a distance of 1.54 from *H. sapiens*. Similarly, no homologue sequence of CSA has been detected in *T. brucei*, although a CSB ortholog has been identified, as we also have reported for *T. cruzi*. These observations imply a different functioning from human TC-NER in these trypanosomatids, which is mentioned as the main subpathway of NER in *T. brucei*, probably due to its multigenic transcription [1].

We also have found an entirely different structure for CSA of *G. lamblia*. Besides a longer length, this protein has two different domains in the C-terminus, a coatomer WD-associated region, and a COPLC domain (Fig. 1A). The first one seems to be related to protein transport between the endoplasmic reticulum and the Golgi apparatus [49], whereas the second one may be involved in the intra-Golgi transport [50]. Besides a different domain composition with a very distinct function, the sequence of this protein corresponds to a divergent branch from the central clustering in the CSA tree. Therefore, it is likely that this is not analogous to human CSA.

The CSB protein belongs to the DNA-dependent ATPase SWI/SNF family, which is known to play a role in chromatin remodeling [5]. In

the screening of the CSB protein structure, we have found two main helicase-related conserved domains, DEXDc and HELICc (Fig. 1B). Indeed, the enzymes within this family harbor a central ATPase domain composed of seven conserved helicase motifs. These are probably contained in DEXDc and HELICc, since the most kept positions along the analyzed species share approximately the same 510–960 region described for the ATPase domain [51]. However, we have found remarkable differences in some structures of putative CSB proteins retrieved in our BLAST searches.

A distinct protein structure has been found for the retrieved sequence of a putative CSB of *D. melanogaster*, whose length is smaller in comparison to the CSB candidates encountered for the other species. Besides having a considerable shift to the left in the positions of DEXDc and HELICc domains, we have verified a DBINO domain in the N-terminus and two SANT domains in the C-terminus of this protein (Fig. 1B). DBINO is a DNA-binding domain of INO80, which is a subfamily of the SNF2 protein family. A striking feature of the INO80 superfamily is the presence of a DBINO domain near the N-terminus, approximately 100 residues upstream to the SNF2 helicase domain [52], which has been reported in our study for *D. melanogaster*. Additionally, SANT domains have a function in chromatin remodeling, assuming an important role in SWI/SNF complexes [53]. Although these domains are involved in chromatin remodeling, which is also a function of CSB, this sequence integrates a divergent clade from the main clustering of the CSB tree, then probably it does not have an analogous role to human CSB. It has been already reported that CSB is missing from fruit flies and even from the order Diptera [38]. Although the *D. melanogaster* genome have also failed to reveal a CSA homologue, a recent methodology using an *in vivo* excision assay and excision repair-sequencing (XR-Seq) surprisingly have showed that fruit flies do perform TC-NER at comparable levels to human cells [54].

The low branch supports obtained in CSB tree indicates the sequences have large evolutionary distances, which has led to the loss of the phylogenetic signal. However, despite the great evolutionary sequence (94.105) observed in CSB of *C. elegans*, a CSB counterpart has indeed been identified in this species, containing 957 amino acids with 37% identity to its human homologue. Also, it is composed of an SNF2-like ATPase domain, which has the same protein domain pattern as the structure we have found in vertebrates [55,56]. This suggests CSB in *C. elegans* performs a similar function to the vertebrate proteins, which in fact has been observed experimentally. Suppression of CSB in *C. elegans* by RNA interference has caused hypersensitivity to UV radiation, leading to cell proliferation arrest, apoptosis and embryonic mortality. This implies that CSB in *C. elegans* does perform a similar role to its known function in TCR [55].

The hHR23B protein is characterized by a ubiquitin-like domain in the N-terminus, two ubiquitin-related domains, and a STI1 domain (Fig. 2B). STI1 is a heat-shock chaperone binding domain involved in XPC-binding [5,57]. We have reported a conserved structure for hHR23B homologues in the eukaryotic species analyzed, but, surprisingly, we have not detected the ubiquitin-like domain in *D. melanogaster*. In budding yeast, the deletion of this domain has been found to be associated with UV sensitivity, revealing its importance in DNA repair [58]. As the ubiquitin-like domain mediates the interaction between the proteasome pathway and DNA repair, it has been presumed that the UV sensitivity caused by its deletion might be due to the inability of RAD23B to interact with the proteasome [59]. Whether this apparent lack of the ubiquitin-like domain in fruit fly has some functional implication or it is an artifact of the conserved domain search is a matter of further research. Regarding *T. cruzi*, the protein retrieved presents an XPC-binding domain, which is homologous to the STI1 domain and also classified as heat-shock chaperone binding [60].

It is also interesting to point out that we have retrieved no hit for hHR23B and XPC in *G. lamblia*, as well as no significant *e-values* for XPE (Table 1). In fact, not all classical NER proteins could be identified in *G. lamblia* genome, but it does not mean that NER is not functional in this

organism. Hypothetically, this may suggest a different NER mechanism or an offset by another repair pathway, since there is a homologous recombination (HR) repair system acting on DNA damage removal in trophozoites [61]. However, *Giardia* cysts have demonstrated a limited ability to repair DNA damage when exposed to UV, which has been attributed to the inactivity of DNA replication in this dormant life stage [62].

The less frequent NER protein in the eukaryotic organisms here evaluated is XPE. This protein is probably lacking in all the analyzed invertebrates, for which we have not obtained significant *e-values* (Table 1, Fig. 2C). The non-detection of a DDB2 subunit in *C. elegans* may imply the absence of its function, which probably is being compensated by the action of DDB1, the other DDB subunit [63]. In *T. brucei*, an apparent homologue of DDB2 has also been not identified, and even the DDB1 subunit has presented low similarity to their human and plant counterparts [1]. Moreover, *S. cerevisiae* has no homologue of any DDB subunit, but this function is likely to be supplied by another binding complex, Rad7-Rad16 [5], which does not have a human homologue detected [64]. Furthermore, a DDB2 homologue has been also not found in *S. pombe* [65]. The proteins Rhp7 and Rhp16, homologues of Rad7 and Rad16 from budding yeast, have demonstrated to be essential for GG-NER and share at least some properties of DDB. Therefore, one may suggest that Rhp7 and Rhp16 are analogues of DDB2 in fission yeast, although their sequences do not have similarity to the human DDB2 [66].

Moreover, we have reported for the first time that a canonical XPD homologue is surprisingly not found in chicken according to our previously described criteria. However, since the helicase activity of XPD is fundamental for NER [67] and also takes part in the initiation of RNA polymerase II transcription [5], it is very likely that some other protein compensates this function in chicken, which needs further studies to be better elucidated. Regarding *tblastn* results, the corresponding hit of the primary search with the human XPD query does not have a significant *e-value*, although it displays a feature named BRIP1, which is also known as Fanconi Anemia group J protein (FANCI). It is a DNA helicase paralogue of XPD, then they indeed share a helicase-related domain [67]. Interestingly, FANCI plays a role in homologous recombination repair and translesion synthesis [68].

Concerning the refined searches for XPD in *G. gallus*, the resulting protein contains a significant *e-value* and indicates the same function of XPD, since it has the same domains at very similar positions (Fig. 3B). Despite its divergence in the XPD phylogeny, this sequence, which is annotated as an ATP-dependent DNA helicase DDX11, is a member of the XPD family [69] and also displays a 5'-3' DNA helicase activity [70]. Additionally, DDX11 acts as a backup for the Fanconi Anemia pathway regarding the repair of intrastrand crosslinks in DT40 chicken cells and also facilitates repair by homologous recombination [69]. Thus, considering the absence of a canonical XPD, we suppose that DDX11 may have an analogous function to XPD in chicken, but functional studies are strongly required to test this hypothesis. Moreover, the presence of annotation errors should also be investigated. Nonetheless, XPD homologues are found in many species of birds other than chicken, comprising at least ten different taxonomic orders [71].

XPA is a UV-damage DNA-binding protein that acts as a scaffold in both GG-NER and TC-NER subpathways [5,72]. XPA is composed of a N-terminus domain containing a zinc-finger motif, a globular central domain, and a C-terminus domain enclosing a shallow cleft [72]. Although we have found both N- and C-terminus domains in the majority of the analyzed eukaryotic species, we could not identify a N-terminal domain in *S. cerevisiae* (Fig. 3A). Despite this yeast homologue has a zinc-finger motif in the N-terminus, this region have diverged considerably. In humans, the N-terminal domain seems to be fundamental for the interaction with RPA [73]. However, experimental observations have suggested that N-terminus is not essential for XPA function in *S. cerevisiae* [74].

It is outstanding that no hit was obtained for XPA in *A. thaliana*

(Table 1), and probably there is no potential homologous of XPA in plants [5,75], which suggests some difference in the primary DNA repair mechanism between plants and animals [75]. Therefore, we have tried to investigate the lack of XPA in the base of plant phylogeny by performing a search in Characeae, since these green algae are close to the ancestors of land plants [76]. To do that, we have used XPA of *S. cerevisiae* to search *Chara braunii* genome, but no hit has been obtained (data not shown).

Curiously, it has been reported that XPA is an intrinsically disordered protein, meaning it does not fold spontaneously into an organized tertiary structure [77,78]. Also, protein disordered regions have a smaller sequence conservation across species [77]. Probably, this explains why an XPA ortholog has not been detected, although an XPA sequence not retrieved by standard search methods may likely participate in plant NER or another protein sharing similar properties to XPA substitutes its function [79]. This may clarify why *A. thaliana* has an ortholog of XAB-1, an XPA-binding protein from humans, even though an XPA itself was not detectable [5].

We also have reported non-significant *e-values* for XPA in *G. lamblia* and *T. cruzi* (Table 1). Actually, XPA has not been detected in the genome of any trypanosomatid [80]. In *T. brucei*, the apparent absence of an XPA homologue has been supposed to be related to the lack of TFIIH function in NER, since XPA seems to play a role in stabilizing the association between TFIIH and the NER machinery [1]. However, the nondetection of XPA in these protozoans can also be explained by the low sequence conservation due to XPA intrinsic disorder [77]. This suggests that XPA have had a greater divergence in plants and protozoans.

Different from *G. lamblia*, *Plasmodium falciparum* parasites have almost all NER components, except p62, a TFIIH component, and XPC [81]. This clearly indicates that NER is present in the tree of life since the origins of eukaryotes. However, in spite of a study had reported the occurrence of XPA in *P. falciparum* [81], when employing our methodology for a more refined search by using XPA of *S. cerevisiae* as query, we could not retrieve any significant result. Also it has not been possible to detect an XPE ortholog when using the corresponding sequence of *A. thaliana* as query (data not shown). Thereby, the question regarding XPA can probably be explained by its intrinsically disordered nature [77]. Concerning XPE, this result supports the hypothesis of its probable nonoccurrence in invertebrates, suggesting it may have appeared in the evolution of species after the diversification of protozoans.

When comparing the putative species tree (Figure S1) with the NER phylogenies, the same pattern could be observed for vertebrates, which form a well-supported clade for most genes, while no pattern has been observed for invertebrates. Regarding gene structures (Figure S2), we have observed clear differences between vertebrates and invertebrates. In general, there is relative maintenance of the intron architecture in vertebrates, which supposes the intervention of a certain selective pressure. On the contrary, invertebrates have a more relaxed pattern, considering a larger variation in their gene structures. Therefore, the results of gene structure and phylogenies reinforce the intricate evolutionary story of NER pathway in eukaryotes, with divergence of functions and the possible presence of analogy.

5. Conclusion

Through the demonstration of the heterogeneity of the gene structures and a particular variety in the protein architecture of the NER components here evaluated, besides the apparent absence of certain proteins in some organisms, our results indicate important differences regarding the eukaryotic NER to the human NER. In this sense, we highlight the lack of a canonical XPD in chicken, a greater divergence of XPA in plants and protozoans and the absence of XPE in the invertebrate species analyzed. Despite this, it is remarkable the presence of this excision repair mechanism in a high number of evolutionarily distant organisms, being present since the origin of eukaryotes. However, the

proteins involved in this pathway are a mix of homologous and analogous proteins, making difficult the understanding of this repair mechanism in different branches of eukaryotes. This opens new perspectives of studies focused on the gaps found in the evolution of eukaryotic NER. Therefore, the many differences here presented do not allow a direct connection among several NER proteins with their human counterparts.

Declaration of Competing Interest

The authors declare that there are no conflicts of interest.

Acknowledgements

We thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior/ Programa de Excelência Acadêmica – Brasil (CAPES/PROEX – 23038.005848/2018-31; 88887.212689/2018-00), and Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq - 407103/2018-0; 307063/2018-6) for the financial support. We also thank Dr. Mauro de Freitas Ortiz for reviewing some BLAST analyzes.

Appendix A. Supplementary data

Supplementary material related to this article can be found, in the online version, at doi:<https://doi.org/10.1016/j.dnarep.2020.102955>.

References

- [1] C.R. Machado, J.P. Vieira-da-Rocha, I.C. Mendes, M.A. Rajão, L. Marcello, M. Bitar, M.G. Drummond, P. Grynberg, D.A.A. Oliveira, C. Marques, B. Van Houten, R. McCulloch, Nucleotide excision repair in *Trypanosoma brucei*: specialization of transcription-coupled repair due to multigenic transcription, *Mol. Microbiol.* 92 (4) (2014) 756–776, <https://doi.org/10.1111/mmi.12589>.
- [2] W.L. de Laat, N.G.J. Jaspers, J.H.J. Hoeijmakers, Molecular mechanism of nucleotide excision repair, *Genes Dev.* 13 (7) (1999) 768–785, <https://doi.org/10.1101/gad.13.7.768>.
- [3] O.D. Schärer, Nucleotide excision repair in eukaryotes, *Cold Spring Harb. Perspect. Biol.* 5 (10) (2013), <https://doi.org/10.1101/cshperspect.a012609>.
- [4] J.A. Marteijn, H. Lans, W. Vermeulen, J.H.J. Hoeijmakers, Understanding nucleotide excision repair and its roles in cancer and ageing, *Nat. Rev. Mol. Cell Biol.* 15 (2014) 465–481, <https://doi.org/10.1038/nrm3822>.
- [5] R.M.A. Costa, V. Chiganças, R.S. Galhardo, H. Carvalho, C.F.M. Menck, The eukaryotic nucleotide excision repair pathway, *Biochimie.* 85 (11) (2003) 1083–1099, <https://doi.org/10.1016/j.biochi.2003.10.017>.
- [6] A.P. Schuch, N.C. Moreno, N.J. Schuch, C.F.M. Menck, C.C.M. Garcia, Sunlight damage to cellular DNA: focus on oxidatively generated lesions, *Free Radic. Biol. Med.* 107 (2017) 110–124, <https://doi.org/10.1016/j.freeradbiomed.2017.01.029>.
- [7] I. Mellon, G. Spivak, P.C. Hanawalt, Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene, *Cell.* 51 (2) (1987) 241–249, [https://doi.org/10.1016/0092-8674\(87\)90151-6](https://doi.org/10.1016/0092-8674(87)90151-6).
- [8] R.P. Rastogi, Richa, A. Kumar, M.B. Tyagi, R.P. Sinha, Molecular mechanisms of ultraviolet radiation-induced DNA damage and repair, *J. Nucleic Acids* 2010 (2010) 1–32, <https://doi.org/10.4061/2010/592980>.
- [9] G. Spivak, Nucleotide excision repair in humans, *DNA Repair (Amst).* 36 (2015) 13–18, <https://doi.org/10.1016/j.dnarep.2015.09.003>.
- [10] P.C. Hanawalt, G. Spivak, Transcription-coupled DNA repair: two decades of progress and surprises, *Nat. Rev. Mol. Cell Biol.* 9 (2008) 958–970, <https://doi.org/10.1038/nrm2549>.
- [11] B. Pani, E. Nudler, Mechanistic insights into transcription coupled DNA repair, *DNA Repair (Amst).* 56 (2017) 42–50, <https://doi.org/10.1016/j.dnarep.2017.06.006>.
- [12] L.E. Giono, N.N. Moreno, A.E.C. Botto, G. Dujardin, M.J. Muñoz, A.R. Kornblihtt, The RNA response to DNA damage, *J. Mol. Biol.* 428 (12) (2016) 2636–2651, <https://doi.org/10.1016/j.jmb.2016.03.004>.
- [13] D. Jaarsma, I. van der Pluijm, G.T.J. van der Horst, J.H.J. Hoeijmakers, Cockayne syndrome pathogenesis: lessons from mouse models, *Mech. Ageing Dev.* 134 (5–6) (2013) 180–195, <https://doi.org/10.1016/j.mad.2013.04.003>.
- [14] G. Spivak, A.K. Ganesan, The complex choreography of transcription-coupled repair, *DNA Repair (Amst).* 19 (2014) 64–70, <https://doi.org/10.1016/j.dnarep.2014.03.025>.
- [15] P. Schwertman, W. Vermeulen, J.A. Marteijn, UVSSA and USP7, a new couple in transcription-coupled DNA repair, *Chromosoma.* 122 (4) (2013) 275–284, <https://doi.org/10.1007/s00412-013-0420-2>.
- [16] J.E. Cleaver, Photosensitivity syndrome brings to light a new transcription-coupled DNA repair cofactor system, *Nat. Genet.* 44 (2012) 477–478, <https://doi.org/10.1038/ng.2255>.

- [17] A. Sarasin, UVSSA and USP7: new players regulating transcription-coupled nucleotide excision repair in human cells, *Genome Med.* 4 (5) (2012) 44, <https://doi.org/10.1186/gm343>.
- [18] C. Petit, A. Sancar, Nucleotide excision repair: from *E. coli* to man, *Biochimie* 81 (1–2) (1999) 15–25, [https://doi.org/10.1016/S0300-9084\(99\)80034-0](https://doi.org/10.1016/S0300-9084(99)80034-0).
- [19] C. Rouillon, M.F. White, The evolution and mechanisms of nucleotide excision repair proteins, *Res. Microbiol.* 162 (1) (2011) 19–26, <https://doi.org/10.1016/j.resmic.2010.09.003>.
- [20] J.A. Eisen, P.C. Hanawalt, A phylogenomic study of DNA repair genes, proteins, and processes, *Mutat. Res. - DNA Repair* 435 (3) (1999) 171–213, [https://doi.org/10.1016/S0921-8777\(99\)00050-6](https://doi.org/10.1016/S0921-8777(99)00050-6).
- [21] C.F.M. Menck, V. Munford, DNA repair diseases: what do they tell us about cancer and aging? *Genet. Mol. Biol.* 37 (1) (2014) 220–233, <https://doi.org/10.1590/S1415-47572014000200008>.
- [22] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D.J. Lipman, J. Ostell, E. W. Sayers, GenBank, *Nucleic Acids Res.* 41 (Database issue) (2013) D36–D42, <https://doi.org/10.1093/nar/gks1195>.
- [23] K.D. Pruitt, T. Tatusova, D.R. Maglott, NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic Acids Res.* 35 (Database issue) (2007) D61–D65, <https://doi.org/10.1093/nar/gkl842>.
- [24] S.F. Altschul, T.L. Madden, A.A. Schäffer, J. Zhang, Z. Zhang, W. Miller, D. J. Lipman, Gapped BLAST and PSI-BLAST: a new generation of protein database search programs, *Nucleic Acids Res.* 25 (17) (1997) 3389–3402, <https://doi.org/10.1093/nar/25.17.3389>.
- [25] W.R. Pearson, An introduction to sequence similarity (“homology”) searching, *Curr. Protoc. Bioinforma.* (2013), <https://doi.org/10.1002/0471250953.bi0301s42>.
- [26] S. Kumar, G. Stecher, K. Tamura, MEGA7: molecular evolutionary genetics analysis version 7.0 for bigger datasets, *Mol. Biol. Evol.* 33 (7) (2016) 1870–1874, <https://doi.org/10.1093/molbev/msw054>.
- [27] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput, *Nucleic Acids Res.* 32 (5) (2004) 1792–1797, <https://doi.org/10.1093/nar/gkh340>.
- [28] J. Schultz, F. Milpetz, P. Bork, C.P. Ponting, SMART, a simple modular architecture research tool: identification of signaling domains, *Proc. Natl. Acad. Sci. U. S. A.* 95 (11) (1998) 5857–5864, <https://doi.org/10.1073/pnas.95.11.5857>.
- [29] R.D. Finn, A. Bateman, J. Clements, P. Coggill, R.Y. Eberhardt, S.R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, E.L.L. Sonnhammer, J. Tate, M. Punta, Pfam: the protein families database, *Nucleic Acids Res.* 42 (Database issue) (2014) D222–D230, <https://doi.org/10.1093/nar/gkt1223>.
- [30] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis, *Mol. Biol. Evol.* 17 (4) (2000) 540–552, <https://doi.org/10.1093/oxfordjournals.molbev.a026334>.
- [31] F. Abascal, R. Zardoya, D. Posada, jModelTest: selection of best-fit models of protein evolution, *Bioinformatics* 21 (9) (2005) 2104–2105, <https://doi.org/10.1093/bioinformatics/bti263>.
- [32] D. Posada, jModelTest: phylogenetic model averaging, *Mol. Biol. Evol.* 25 (7) (2008) 1253–1256, <https://doi.org/10.1093/molbev/msn083>.
- [33] B. Hu, J. Jin, A.Y. Guo, H. Zhang, J. Luo, G. Gao, GSDS 2.0: an upgraded gene feature visualization server, *Bioinformatics* 31 (8) (2015) 1296–1297, <https://doi.org/10.1093/bioinformatics/btu817>.
- [34] D.L. Wheeler, D.M. Church, S. Federhen, A.E. Lash, T.L. Madden, J.U. Pontius, G. D. Schuler, L.M. Schriml, E. Sequeira, T.A. Tatusova, L. Wagner, Database resources of the National Center for Biotechnology Information, *Nucleic Acids Res.* 44 (1) (2003) 28–33, <https://doi.org/10.1093/nar/gkg033>.
- [35] D.G. Passos-Silva, M.A. Rajão, P.H.N. Aguiar, J.P. Vieira-da-Rocha, C.R. Machado, C. Furtado, Overview of DNA repair in *Trypanosoma cruzi*, *Trypanosoma brucei* and *Leishmania major*, *J. Nucleic Acids* 2010 (2010), <https://doi.org/10.4061/2010/840768>.
- [36] C. Zhang, F. Zhang, The multifunctions of WD40 proteins in genome integrity and cell cycle progression, *J. Genomics* 3 (2015) 40–50, <https://doi.org/10.7150/jgen.11015>.
- [37] J.J. Sekelsky, M.H. Brodsky, K.C. Burtis, DNA repair in *Drosophila*: insights from the *Drosophila* genome sequence, *J. Cell Biol.* 150 (2) (2000) F31–F36, <https://doi.org/10.1083/jcb.150.2.F31>.
- [38] J. Sekelsky, DNA repair in *Drosophila*: mutagens, models, and missing genes, *Genetics* 205 (2) (2017) 471–490, <https://doi.org/10.1534/genetics.116.186759>.
- [39] J. Rolff, P.R. Johnston, S. Reynolds, Complete metamorphosis of insects, *Philos. Trans. R. Soc. B Biol. Sci.* 374 (2019), 20190063, <https://doi.org/10.1098/rstb.2019.0063>.
- [40] S. Hatakeyama, K.I. Nakayama, U-box proteins as a new family of ubiquitin ligases, *Biochem. Biophys. Res. Commun.* 302 (4) (2003) 635–645, [https://doi.org/10.1016/S0006-291X\(03\)00245-6](https://doi.org/10.1016/S0006-291X(03)00245-6).
- [41] V. Babu, K. Hofmann, B. Schumacher, A.C. *elegans* homolog of the Cockayne syndrome complementation group A gene, *DNA Repair (Amst)* 24 (2014) 57–62, <https://doi.org/10.1016/j.dnarep.2014.09.011>.
- [42] M. Gribskov, A.D. McLachlan, D. Eisenberg, Profile analysis: detection of distantly related proteins, *Proc. Natl. Acad. Sci. U.S.A.* 84 (13) (2006) 4355–4358, <https://doi.org/10.1073/pnas.84.13.4355>.
- [43] P.K. Bhatia, R.A. Verhage, E.C. Friedberg, Molecular cloning and characterization of *Saccharomyces cerevisiae* RAD28, the yeast homolog of the human Cockayne syndrome A (CSA) gene, *J. Bacteriol.* 178 (20) (1996) 5977–5988, <https://doi.org/10.1128/jb.178.20.5977-5988.1996>.
- [44] A.J. van Gool, R. Verhage, S.M.A. Swagemakers, P. van de Putte, J. Brouwer, C. Troelstra, D. Bootsma, J.H.J. Hoeijmakers, RAD26, the functional *S. cerevisiae* homolog of the Cockayne syndrome B gene *ERCC6*, *EMBO J.* 13 (22) (1994) 5361–5369, <https://doi.org/10.1002/j.1460-2075.1994.tb06871.x>.
- [45] Y. Fukumoto, N. Dohmae, F. Hanaoka, *Schizosaccharomyces pombe* Ddb1 recruits substrate-specific adaptor proteins through a novel protein motif, the DDB-box, *Mol. Cell. Biol.* 28 (22) (2008) 6746–6756, <https://doi.org/10.1128/mcb.00757-08>.
- [46] S. Li, M.J. Smerdon, Rpb4 and Rpb9 mediate subpathways of transcription-coupled DNA repair in *Saccharomyces cerevisiae*, *EMBO J.* 21 (21) (2002) 5921–5929, <https://doi.org/10.1093/emboj/cdf589>.
- [47] S. Li, Transcription coupled nucleotide excision repair in the yeast *Saccharomyces cerevisiae*: the ambiguous role of Rad26, *DNA Repair (Amst)* 36 (2015) 43–48, <https://doi.org/10.1016/j.dnarep.2015.09.006>.
- [48] M. Tijsterman, R.A. Verhage, P. van de Putte, J.G. Tasseron-De Jong, J. Brouwer, Transitions in the coupling of transcription and nucleotide excision repair within RNA polymerase II-transcribed genes of *Saccharomyces cerevisiae*, *Proc. Natl. Acad. Sci. U. S. A.* 94 (15) (1997) 8027–8032, <https://doi.org/10.1073/pnas.94.15.8027>.
- [49] V.T.K. Chow, H.H. Quek, HEP-COP, a novel human gene whose product is highly homologous to the α -subunit of the yeast coatomer protein complex, *Gene* 169 (2) (1996) 223–227, [https://doi.org/10.1016/0378-1119\(95\)00738-5](https://doi.org/10.1016/0378-1119(95)00738-5).
- [50] J. B ethune, F. Wieland, J. Moellegen, COPI-mediated transport, *J. Membr. Biol.* 211 (2) (2006) 65–79, <https://doi.org/10.1007/s00232-006-0859-7>.
- [51] R.J. Lake, H.-Y. Fan, Structure, function and regulation of CSB: a multi-talented gymnast, *Mech. Ageing Dev.* 134 (0) (2013) 202–211, <https://doi.org/10.1016/j.mad.2013.02.004>.
- [52] R. Bakshi, T. Prakash, D. Dash, V. Brahmachari, In silico characterization of the INO80 subfamily of SWI2/SNF2 chromatin remodeling proteins, *Biochem. Biophys. Res. Comm.* 320 (1) (2004) 197–204, <https://doi.org/10.1016/j.bbrc.2004.05.147>.
- [53] L.A. Boyer, R.R. Latek, C.L. Peterson, The SANT domain: a unique histone-tail-binding module? *Nat. Rev. Mol. Cell Biol.* 5 (2004) 158–163, doi:10.1038/nrm1314.
- [54] N. Deger, Y. Yang, L.A. Lindsey-Boltz, A. Sancar, C.P. Selby, *Drosophila*, which lacks canonical transcription-coupled repair proteins, performs transcription-coupled repair, *J. Biol. Chem.* 294 (48) (2019) 18092–18098, <https://doi.org/10.1074/jbc.AC119.011448>.
- [55] M.H. Lee, B. Ahn, I.S. Choi, H.S. Koo, The gene expression and deficiency phenotypes of Cockayne syndrome B protein in *Caenorhabditis elegans*, *FEBS Lett.* 522 (1–3) (2002) 47–51, [https://doi.org/10.1016/S0014-5793\(02\)02880-6](https://doi.org/10.1016/S0014-5793(02)02880-6).
- [56] H. Lans, J.A. Marteijn, B. Schumacher, J.H.J. Hoeijmakers, G. Jansen, W. Vermeulen, Involvement of global genome repair, transcription coupled repair, and chromatin remodeling in UV DNA damage response changes during development, *PLoS Genet.* 6 (5) (2010), <https://doi.org/10.1371/journal.pgen.1000941>.
- [57] M. Yokoi, F. Hanaoka, Two mammalian homologs of yeast Rad23, HR23A and HR23B, as multifunctional proteins, *Gene* 597 (2017) 1–9, <https://doi.org/10.1016/j.gene.2016.10.027>.
- [58] J.F. Watkins, P. Sung, L. Prakash, S. Prakash, The *Saccharomyces cerevisiae* DNA repair gene *RAD23* encodes a nuclear protein containing a ubiquitin-like domain required for biological function, *Mol. Cell. Biol.* 13 (12) (1993) 7757–7765, <https://doi.org/10.1128/mcb.13.12.7757>.
- [59] C. Schaubert, L. Chen, P. Tongaonkar, I. Vega, D. Lambertson, W. Potts, K. Madura, Rad23 links DNA repair to the ubiquitin/proteasome pathway, *Nature* 391 (1998) 715–718, <https://doi.org/10.1038/35661>.
- [60] B. Kim, K.S. Ryu, H.J. Kim, S.J. Cho, B.S. Choi, Solution structure and backbone dynamics of the XPC-binding domain of the human DNA repair protein hHR23B, *FEBS J.* 272 (2005) 2467–2476, <https://doi.org/10.1111/j.1742-4658.2005.04667.x>.
- [61] A. Sandoval-Cabrera, A.L. Zarzosa- lvarez, R.M. Mart nez-Miguel, R.M. Berm dez-Cruz, MR (Mre11-Rad50) complex in *Giardia duodenalis*: in vitro characterization and its response upon DNA damage, *Biochimie* 111 (2015) 45–57, <https://doi.org/10.1016/j.biochi.2015.01.008>.
- [62] E. Einarsson, S.G. Sv rd, K. Troell, UV irradiation responses in *Giardia intestinalis*, *Exp. Parasitol.* 154 (2015) 25–32, <https://doi.org/10.1016/j.exppara.2015.03.024>.
- [63] H. Lans, W. Vermeulen, Nucleotide excision repair in *Caenorhabditis elegans*, *Mol. Biol. Int.* 2011 (2011), <https://doi.org/10.4061/2011/542795>.
- [64] S. Prakash, L. Prakash, Nucleotide excision repair in yeast, *Mutat. Res.* 451 (1–2) (2000) 13–24, [https://doi.org/10.1016/S0027-5107\(00\)00037-3](https://doi.org/10.1016/S0027-5107(00)00037-3).
- [65] F. Zolezzi, S. Linn, Studies of the murine *DDBI* and *DDB2* genes, *Gene* 245 (1) (2000) 151–159, [https://doi.org/10.1016/S0378-1119\(00\)00022-6](https://doi.org/10.1016/S0378-1119(00)00022-6).
- [66] M. Lombaerts, P.H. Peltola, R. Visse, H. den Dulk, J.A. Brandsma, J. Brouwer, Characterization of the *rhp7+* and *rhp16+* genes in *Schizosaccharomyces pombe*, *Nucleic Acids Res.* 27 (17) (1999) 3410–3416, <https://doi.org/10.1093/nar/27.17.3410>.
- [67] M.F. White, Structure, function and evolution of the XPD family of iron-sulfur-containing 5 → 3 DNA helicases, *Biochem. Soc. Trans.* 37 (2009) 547–551, <https://doi.org/10.1042/BST0370547>.
- [68] R. Ceccaldi, P. Sarangi, A.D. D’Andrea, The Fanconi anemia pathway: new players and new functions, *Nat. Rev. Mol. Cell Biol.* 17 (2016) 337–349, <https://doi.org/10.1038/nrm.2016.48>.
- [69] T. Abe, M. Ooka, R. Kawasumi, K. Miyata, M. Takata, K. Hirota, D. Branzi, Warsaw breakage syndrome DDX11 helicase acts jointly with RAD17 in the repair of bulky lesions and replication through abasic sites, *Proc. Natl. Acad. Sci. U. S. A.* 115 (33) (2018) 8412–8417, <https://doi.org/10.1073/pnas.1803110115>.

- [70] Y. Hirota, J.M. Lahti, Characterization of the enzymatic activity of hChR1, a novel human DNA helicase, *Nucleic Acids Res.* 28 (4) (2000) 917–924, <https://doi.org/10.1093/nar/28.4.917>.
- [71] ERCC2 orthologs - NCBI, (n.d.). <https://www.ncbi.nlm.nih.gov/gene/2068/ortholog/?scope=8782#genes-tab> (accessed March 25, 2020).
- [72] N. Sugitani, R.M. Sivley, K.E. Perry, J.A. Capra, W.J. Chazin, XPA: a key scaffold for human nucleotide excision repair, *DNA Repair (Amst)*. 44 (2016) 123–135, <https://doi.org/10.1016/j.dnarep.2016.05.018>.
- [73] B.C. Feltes, D. Bonatto, Overview of xeroderma pigmentosum proteins architecture, mutations and post-translational modifications, *Mutat. Res. - Rev. Mutat. Res.* 763 (2015) 306–320, <https://doi.org/10.1016/j.mrrev.2014.12.002>.
- [74] M. Bankmann, L. Prakash, S. Prakash, Yeast *RAD14* and human xeroderma pigmentosum group A DNA-repair genes encode homologous proteins, *Nature* 355 (1992) 555–558, <https://doi.org/10.1038/355555a0>.
- [75] S. Kimura, K. Sakaguchi, DNA repair in plants, *Chem. Rev.* 106 (2) (2006) 753–766, <https://doi.org/10.1021/cr040482n>.
- [76] M.J. Beilby, *Chara braunii* genome: a new resource for plant electrophysiology, *Biophys. Rev.* 11 (2019) 235–239, <https://doi.org/10.1007/s12551-019-00512-7>.
- [77] L.M. Iakoucheva, A.L. Kimzey, C.D. Masselon, J.E. Bruce, E.C. Garner, C.J. Brown, A.K. Dunker, R.D. Smith, E.J. Ackerman, Identification of intrinsic order and disorder in the DNA repair protein XPA, *Protein Sci.* 10 (3) (2001) 560–571, <https://doi.org/10.1110/ps.29401>.
- [78] H.J. Dyson, P.E. Wright, Intrinsically unstructured proteins and their functions, *Nat. Rev. Mol. Cell Biol.* 6 (2005) 197–208, <https://doi.org/10.1038/nrm1589>.
- [79] F. Canturk, M. Karaman, C.P. Selby, M.G. Kemp, G. Kulaksiz-Erkmen, J. Hu, W. Li, L.A. Lindsey-Boltz, A. Sancar, Nucleotide excision repair by dual incisions in plants, *Proc. Natl. Acad. Sci. U. S. A.* 113 (17) (2016) 4706–4710, <https://doi.org/10.1073/pnas.1604097113>.
- [80] N.M. El-Sayed, P.J. Myler, D.C. Bartholomeu, D. Nilsson, G. Aggarwal, S. J. Westenberger, A. Tran, E. Ghedin, E.A. Worthey, A.L. Delcher, E. Caler, G. C. Cerqueira, C. Branche, B. Haas, A. Anupama, E. Arner, A. Lena, P. Burton, E. Cadag, D.A. Campbell, P. Attipoe, E. Bontempi, M. Carrington, J. Crabtree, H. Darban, J. Franco, P. De Jong, A.C. Frasch, K. Gull, D. Horn, L. Hou, Y. Huang, E. Kindlund, M. Klingbeil, S. Kluge, H. Koo, D. Lacerda, M.J. Levin, H. Lorenzi, T. Louie, C.R. Machado, R. Mcculloch, A. Mckenna, Y. Mizuno, J.C. Mottram, S. Nelson, S. Ochaya, K. Osoegawa, G. Pai, M. Parsons, M. Pentony, U. Pettersson, M. Pop, J.L. Ramirez, J. Rinta, L. Robertson, S.L. Salzberg, D.O. Sanchez, A. Seyler, R. Sharma, J. Shetty, A.J. Simpson, E. Sisk, M.T. Tammi, R. Tarleton, S. Teixeira, S. Van Aken, C. Vogt, The genome sequence of *Trypanosoma cruzi*, etiologic agent of Chagas disease, *Science* 309 (5733) (2005) 409–415, <https://doi.org/10.1126/science.1112631>.
- [81] L. Tajedin, M. Anwar, D. Gupta, R. Tuteja, Comparative insight into nucleotide excision repair components of *Plasmodium falciparum*, *DNA Repair (Amst)*. 28 (2015) 60–72, <https://doi.org/10.1016/j.dnarep.2015.02.009>.

3 ARTIGO 2 – XERODERMA PIGMENTOSUM D HELICASE IN GALLIFORM BIRDS: WHERE IT HAS FLOWN TO?

Este estudo continua a investigação de um resultado do artigo apresentado anteriormente, referente à ausência de uma proteína XPD canônica na espécie *Gallus gallus*. Utilizando a mesma metodologia, a sequência ortóloga de XPD do crocodiliano *Alligator mississippiensis* foi utilizada para explorar a ocorrência dessa proteína em Galliformes, Anseriformes, Struthioniformes e Tinamiformes com um enfoque estrutural. Nesse sentido, os resultados obtidos conferem uma visão mais ampla de que XPD possa estar ausente em todo o grupo-irmão Galliformes-Anseriformes. Além disso, informações recentes da literatura trazem novas contribuições para uma hipótese proposta no primeiro artigo no que se refere à compensação da função de XPD em *G. gallus*. A partir da página seguinte, é apresentado um manuscrito em fase de preparação para submissão na revista *Biochimica et Biophysica Acta (BBA) – General Subjects*.

Xeroderma Pigmentosum D helicase in Galliform birds: where it has flown to?

Rayana dos Santos Feltrin^{a,b}, Ana Lúcia Anversa Segatto^c, Tiago Antonio de Souza^d, Edward Louis Braun^e, and André Passaglia Schuch^{a,b*}

^a Department of Biochemistry and Molecular Biology, Federal University of Santa Maria, RS, Brazil

^b Postgraduate Program in Biological Sciences: Toxicological Biochemistry

^c Federal Institute of Rio Grande do Sul, Caxias do Sul, RS, Brazil

^d TauGC Bioinformatics, São Paulo, SP, Brazil

^e Department of Biology, University of Florida, Gainesville, FL, United States of America

* Correspondence to: Av. Roraima, 1000, P.O. Box 5021, room 3010, Camobi, Santa Maria, RS, 97110-970, Brazil. Fone: 55.55.3220-8136.

E-mail: schuchap@gmail.com

Abstract

Background: Xeroderma Pigmentosum D (XPD) is an important helicase in the constitution of the basal transcription factor IIIH (TFIIH), playing pivotal roles in nucleotide excision repair (NER) and transcription. Motivated by previous works that have shown the absence of a canonical XPD helicase in some birds, we have tried to better investigate this intriguing finding in an evolutionary fashion.

Methods: We have performed a refined search of ERCC2 (XPD) in genomes and proteomes of birds from the orders Tinamiformes, Struthioniformes, Galliformes and Anseriformes by using similarity and structural criteria. In addition, to analyze the conservation of gene positions, we have explored the occurrence of genomic neighbors of alligator *ERCC2* in the chicken genome.

Results: We have revealed that a canonical XPD is not found in the sister group Galliformes-Anseriformes in spite of being present in the base of the bird phylogeny. However, the genomic context of chicken *ERCC2* is not clear enough to suggest a synteny.

Conclusions: The obtained results suggest that there might have been a loss of the XPD sequence in Galloanseres, in addition to a potential function convergence between XPD of Alligator and DDX11 and FANCI of Galloanseres.

Keywords: TFIIH, XPD, DNA repair, helicase, Galliform birds, conserved protein domains.

1. Introduction

Xeroderma Pigmentosum D (XPD) is an important helicase that acts in concert with Xeroderma Pigmentosum B (XPB) in nucleotide excision repair (NER), the most flexible and versatile DNA repair pathway, as it is capable of removing a myriad of structurally unrelated bulky DNA lesions [1,2]. XPB and XPD are critical subunits of the basal transcription factor IIIH (TFIIH), the former composing its seven-subunit core together with p8 (TTD), p34, p44, p52, and p62. TFIIH is also constituted by a three-subunit kinase complex including CDK7, cyclin H, and MAT1, which is attached to the core by XPD [1,3]. Through the 3'-5' XPB and 5'-3' XPD activities, TFIIH is responsible for opening the double-helix around the site containing the transcription promoter or the DNA damage, thereby allowing the access of RNA-polymerase II (RNA-pol II) or NER proteins, respectively [4,5]. However, it is suggested that in NER, only the ATPase activity of XPB is required, combined with the helicase function of XPD [6].

The XPD helicase is probably present in all domains of life, as most archaea have clear homologues, and also *Escherichia coli*, whose DinG protein contains an iron-sulfur-cluster (FeS), a characteristic feature of this helicase family [7]. Additionally, in spite of having been demonstrated that XPD enzymatic activity is dispensable for transcription initiation, a recently published work showed that the interaction between core TFIIH and the kinase complex (CAK) by XPD plays a fundamental role in the phosphorylation of the C-terminal domain of RNA-pol II and in RNA synthesis as well [8,9]. In fact, the deletion of the *ERCC2* gene, which encodes for XPD, results in lethality in both mice embryos and budding yeast (*RAD3*), probably due to its essential function in transcription [10,11]. Furthermore, mutations in human XPD can cause distinct inherited diseases such as Xeroderma Pigmentosum (XP), Cockayne Syndrome (CS), XP combined with CS, and trichothiodystrophy (TTD) [5]. All of these genetic conditions share

a skin photosensitivity phenotype as a result of a defective NER mechanism in response to UV-induced damage [12].

By contrast, previous works have surprisingly shown that a canonical XPD could not be identified in some bird species with completely sequenced genomes, such as *Gallus gallus* (chicken). Notwithstanding, XPD counterparts have actually been encountered in at least 10 different taxonomic orders, including *Tinamus guttatus*, (Tinamiformes), which is in the base of the bird phylogeny [13–15]. Therefore, this study aims to better investigate this intriguing absence of XPD sequence in an evolutionary fashion by searching for it in genomes and proteomes of Tinamiformes, Struthioniformes (both in the base of the bird tree), Galliformes, namely the *G. gallus* order, and Anseriformes, its sister group [16]. In addition, we have explored the occurrence of genomic neighbors of alligator *ERCC2* in the chicken genome. Then, we highlight the probable absence of XPD in and the Galliformes-Anseriformes group.

2. Material and methods

2.1 Search for XPD in birds

In order to better investigate the occurrence of XPD in birds other than *G. gallus*, we have used a previously described homology searching method, which takes into account *e-value* (lower than 1×10^{-6}), percent identity, protein sequence length, and conserved protein domains [13], to perform a refined search. To do so, we have used the nucleotide (XM_014604761.2) and protein (XP_014460247.1) sequences of XPD of *Alligator mississippiensis*, formerly obtained [13], as queries for *blastx* and *tblastn* searches. The genomes and proteomes of Galliformes, Anseriformes, Tinamiformes and Struthioniformes have been searched on NCBI Genome List by taxonomic order and used as subjects for the search. The information on the resulting sequences is in Tab S1.

2.2 Analysis of genomic context based on *Alligator ERCC2* region

Additionally, to investigate whether genes from the same genomic context of *ERCC2* in *A. mississippiensis* occur in *G. gallus* genome or not, firstly we have examined the chromosomal location of the former on GenBank (Gene ID: 102572247), and also of humans (Gene ID: 2068) for a matter of comparison [13,17]. Then, *tblastn* searches have been carried out in the chicken genome, considering the amino acid sequences of the two immediate genomic neighbors of *A. mississippiensis* XPD, as queries for individual searches. The same selection criteria have been used.

3. Results and discussion

3.1 XPD is present in the base of the bird phylogeny

In a general way, the bird phylogeny is divided in Palaeognathae - which includes the flying tinamous (e.g. *Tinamus guttatus* and *Nothoprocta perdicaria*) and the flightless ratites (such as *Struthio camelus*) – and Neognathae, that is, all of the other birds. In turn, the latter is splitted into Galloanserae, which comprises the sister groups Galliformes and Anseriformes, as well as Neoaves, the one containing the majority of avian diversity [15,16]. Regarding our results of XPD occurrence in birds other than *G. gallus*, we have found putative sequences with good statistics (Tab S1) and similar protein structures (Fig 1A, B) in only three genomes (*N. perdicaria* and *T. guttatus* in Tinamiformes; *S. camelus* in Struthioniformes) out of 28 searched. Moreover, the *NCBI Orthologs* page for ERCC2 (XPD) indicates that it is also present in the palaeognaths *Dromaius novaehollandiae* (emu) and also in two species of the genus *Apteryx* (commonly known as kiwi), the closest relatives to the recently extinct Madagascan elephant birds [15,18,19]. Altogether, this implies that ERCC2 is present in the base of the bird phylogeny.

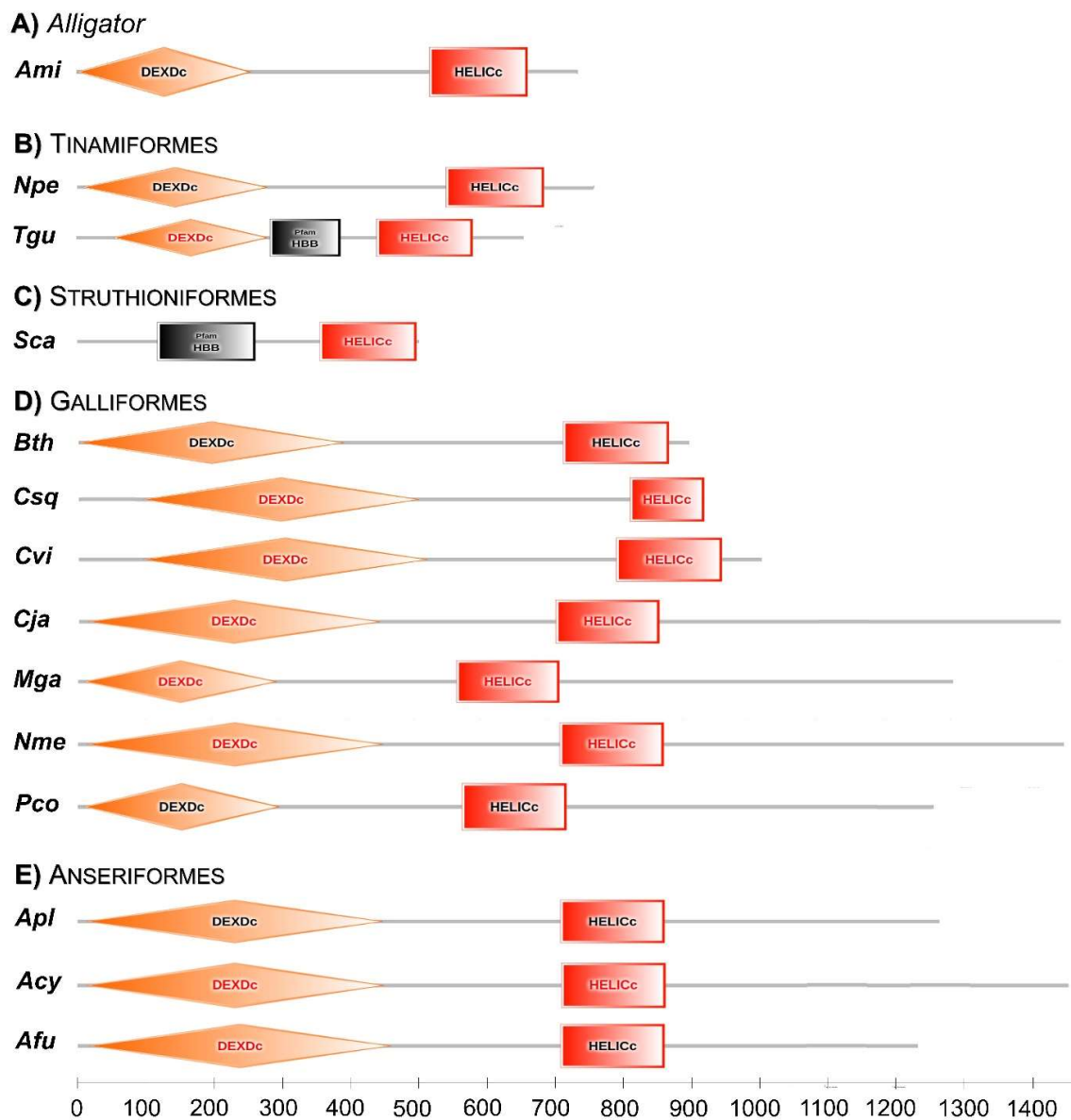


Figure 1: Protein domains of XPD in *Alligator* and the resulting proteins of Tinamiformes, Struthioniformes, Galliformes, and Anseriformes. **A)** *Alligator*. *Ami*: *Alligator mississippiensis*. **B)** Tinamiformes. *Npe*: *Nothoprocta perdicaria*, *Tgu*: *Tinamus guttatus*. **C)** Struthioniformes. *Sca*: *Struthio camelus*. **D)** Galliformes. *Bth*: *Bambusicola thoracicus*, *Csq*: *Callipepla squamata*, *Cvi*: *Colinus virginianus*, *Cja*: *Coturnix japonica*, *Mga*: *Meleagris gallopavo*, *Nme*: *Numida meleagris*, *Pco*: *Phasianus colchicus*. **E)** Anseriformes. *Apl*: *Anas platyrhynchos*, *Acy*: *Anser cygnoides*, *Afu*: *Aythya fuligula*. Structures have been displayed only in species for which there are available protein sequences. Domains have been obtained from the SMART server.

With regard to the protein conserved domains, *N. perdicaria* have similar domains to XPD in *A. mississippiensis*, DEXDc and HELICc [13], whereas *S. camelus* has a smaller protein sequence containing only HELICc and an additional domain HBB, but it is a partial sequence, which would be problematic to claim a domain loss [20] since it is not complete. Nevertheless, based on the results of XPD occurrences in *N. perdicaria*, *T. guttatus* and *S. camelus*, we have used their sequences as queries of *blastx* searches against the chicken genome. However, we also have not obtained a canonical XPD, but the DDX11 helicase has been retrieved in two of the searches with a significant *e-value* (data not shown).

Accordingly, a previous work of our group has recently suggested that this protein may have an analogous function to XPD as they are composed by the same domains [13]. In fact, XPD does perform a paramount role in NER pathway and transcription, and essential genes have a strong tendency to be retained [21]. Despite this, findings of a homology-based model of DDX11 in a fungus species predicted that some of its residues located in the equivalent region of XPD may employ a similar role [9]. Thus, our preceding hypothesis needs to be tested by means of the evolution of redundancy, in which a gene may be possibly considered nonessential by virtue of paralogs or even nonrelated genes which might provide the same function [21].

3.2 A canonical XPD is not found in Galloanserae

The *blastx* searches in 14 Galliform genomes have retrieved proteins with significant *e-value* and protein domains in common with XPD sharing similar positions, such as the sequences found in *Bambusicola thoracicus*, *Callipepla squamata*, and *Colinus virginianus* (Fig 1C). Moreover, proteins annotated as Fanconi anemia group J (FANCI), also containing both XPD domains, have been obtained in *Coturnix japonica* and *Numida meleagris*, even though these proteins are much larger. Indeed, the Fanconi Anemia pathway is known to be crucial for intrastrand crosslink repair in vertebrate cells [22]. In turn, sequences somewhat

shorter than the FANCI proteins, with XPD domains as well, were found in *Meleagris gallopavo* and *Phasianus colchicus*, and are annotated as regulators of telomere elongation helicase 1 (RTEL1), acting in the maintenance of telomere integrity [23].

Concerning *tblastn* searches in Galliformes, we have obtained significant results considering *e-value* and percent identity (71-90%) in nine species, but with very low query cover (5-14%). A similar result pattern has also been reported in Anseriformes (e.g. ducks) [16] of which nine genomes have been searched, seven of them having considerable results in terms of *e-value*. Of these, we have observed higher percent identities (73-97%) in four species, but actually only three of them have annotated features: *Anas platyrhynchos*, *Anser cygnoides*, and *Aythya fuligula*. In spite of this, they also present low coverage (4-16%). Their corresponding proteins (Fig 1D) are annotated as FANCI as well, having both XPD characteristic domains with conserved positions among Anseriformes, which are quite similar to the Galliform species *B. thoracicus*, *C. japonica*, and *N. meleagris* and then, they possibly have the same function as FANCI. Taking this together leads to the fact that XPD is not found in Galliformes (landfowl) and Anseriformes (waterfowl). Thereafter, perhaps it has been lost in the common ancestor of Galloanseres, though this hypothesis needs robust phylogenetic analyzes to be supported [24].

3.3 The genomic context of chicken *ERCC2* is not clear enough to suggest a synteny

We have found the same two genes surrounding *ERCC2* in both human and alligator genomes: a kinesin light chain gene (*KLC3*) upstream and a protein phosphatase gene (*PPP1R13L*) downstream (Fig 2A, B), that is, they are syntenic. Concerning the use of alligator sequences as queries (XP_019351126.1 for *KLC3* and XP_019351129.1 for *PPP1R13L*), the *tblastn* searches have resulted in two putative homologues for these genes in *G. gallus*, both located in the same chromosome 5 region (Fig 2C). However, there are other two genes between them, *XRCC3* and *ZFYVE21*, besides a different genic composition within the referred region. Therefore, albeit synteny helps to infer homology, our results here demonstrated do not allow

us to ascertain that there is in fact a syntenic relationship between alligator and chicken chromosomes [25].

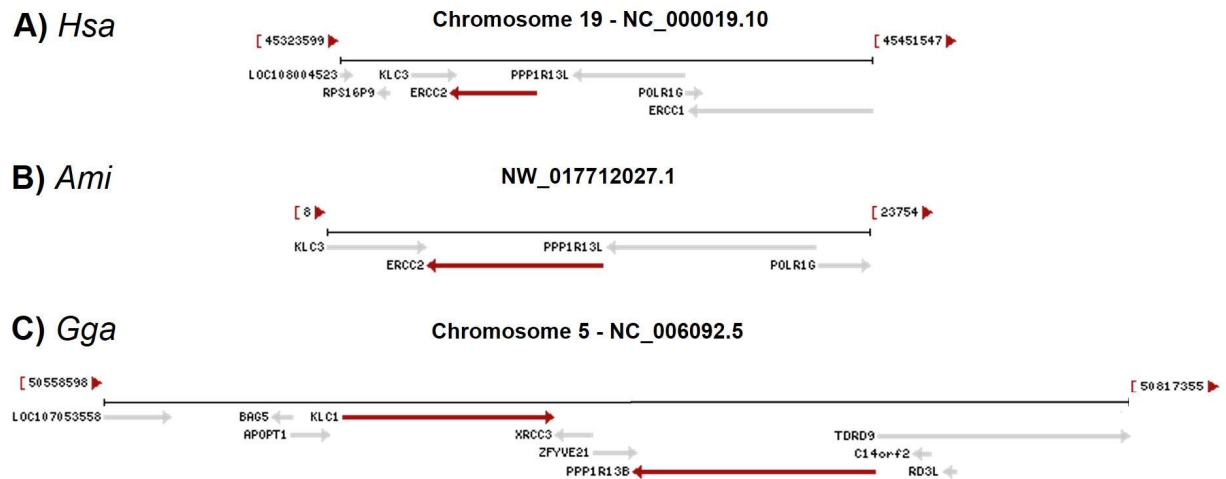


Figure 2: Genomic context of *ERCC2*. **A)** In *H. sapiens* (*Hsa*), *ERCC2*, represented in red, is located in the chromosome 19 (genomic sequence NC_000019.10). **B)** In *Alligator mississippiensis* (*Ami*), *ERCC2*, also in red, is located within the genomic sequence NW_017712027.1 (unplaced scaffold). **C)** In *G. gallus*, the putative homologues for *KLC3* and *PPP1R13L* (in red) are both placed in chromosome 5 (genomic sequence NC_006092.5). The genomic context visualization has been obtained on GenBank.

More specifically about the BLAST results, in the search for *KLC3* and *PPP1R13L* in *G. gallus*, the resulting sequences we have retrieved from the filtering criteria have fitted our *e-value* threshold (Tab S2). The kinesin sequence has 95% identity to the query, both having one TPR_10 domain and four TPR domains somewhat in the middle of the protein towards the C-terminus (Fig 3). However, the one of *G. gallus* has an additional TPR domain near the C-terminal region, and the corresponding sequence is annotated as *KLC1*. Kinesins are a superfamily of proteins involved in the molecular motion along microtubules, having four isoforms (KLC1-4) identified in vertebrates. One of the characteristic features of KLC is the

tetratricopeptide domain (TPR), whose function is associated to the binding of different cargos [26]. Henceforth, despite KLC of chicken has an extra TPR domain, this may still indicate an ortholog relationship.

Regarding the phosphatase sequence, we have observed a percent identity of 53%, presenting two ANK domains and one SH3 in the C-terminal region (Fig 3). Conversely, the domain positions between alligator and chicken have a notable difference of around 250 aa, and the chicken sequence is annotated as *PPP1R13B*. Regardless, that is not a huge difference in the sense that both species have the same conserved domains, and the function of a protein is, in general, determined by its domain architecture [27].

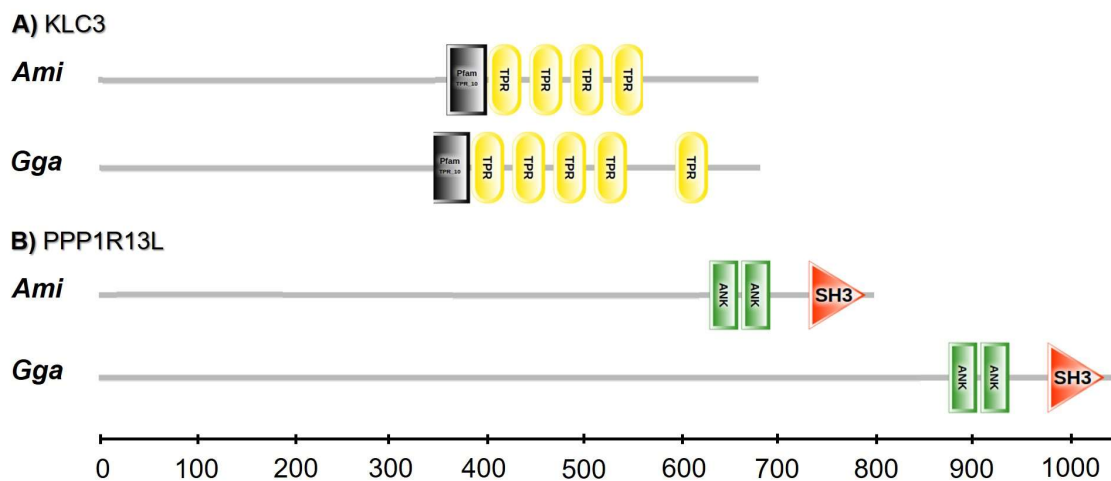


Figure 3: Protein domains of the putative KLC3 and PPP1R13L proteins. **A)** KLC3. **B)** PPP1R13L. Domains have been obtained from the SMART server. *Ami:* *Alligator mississippiensis*, *Gga:* *Gallus gallus*.

4. Conclusions

Our major finding is the absence of XPD in Galloanseres spite of being present in the base of the bird phylogeny. Additionally, our evidences coming from homology searches presented here extended the results of our previous work and do show a potential function convergence between XPD of *Alligator* and DDX11 and FANCI of Galloanseres. Also, the

analyzes of the genomic context based on the *XPD* counterpart of *Alligator* suggest us an unclear syntenic relationship. Thereby, our work gives a glimpse on which wings chicken can be using to deal with DNA damage.

Conflict of interest

The authors declare that there are no conflicts of interest.

Acknowledgements

We thank Coordenação de Aperfeiçoamento de Pessoal de Nível Superior/ Programa de Excelência Acadêmica – Brasil (CAPES/PROEX – 23038.005848/2018-31; 88887.212689/2018-00), and Conselho Nacional de Desenvolvimento Científico e Tecnológico – Brasil (CNPq - 407103/2018-0; 307063/2018-6) for the financial support.

References

- [1] A.A. Galande, N. Perween, M. Saijo, S.S. Ghaskadbi, S. Ghaskadbi, Analysis of the conserved NER helicases (XPB and XPD) and UV-induced DNA damage in Hydra, *Biochim. Biophys. Acta - Gen. Subj.* 1862 (2018) 2031–2042. doi:10.1016/j.bbagen.2018.06.017.
- [2] R.M.A. Costa, V. Chiganças, R.D.S. Galhardo, H. Carvalho, C.F.M. Menck, The eukaryotic nucleotide excision repair pathway, *Biochimie.* 85 (2003) 1083–1099. doi:10.1016/j.biochi.2003.10.017.
- [3] L.H.F. Mullenders, Solar UV damage to cellular DNA: From mechanisms to biological effects, *Photochem. Photobiol. Sci.* 17 (2018) 1842–1852. doi:10.1039/c8pp00182k.
- [4] A.P. Schuch, N.C. Moreno, N.J. Schuch, C.F.M. Menck, C.C.M. Garcia, Sunlight damage to cellular DNA: Focus on oxidatively generated lesions, *Free Radic. Biol. Med.* 107 (2017) 110–124. doi:10.1016/j.freeradbiomed.2017.01.029.
- [5] L.K. Lerner, N.C. Moreno, C.R.R. Rocha, V. Munford, V. Santos, D.T. Soltys, C.C.M. Garcia, A. Sarasin, C.F.M. Menck, XPD/ERCC2 mutations interfere in cellular responses to oxidative stress, *Mutagenesis.* 34 (2019) 341–354.

- doi:10.1093/mutage/gez020.
- [6] F. Coin, V. Oksenysh, J.M. Egly, Distinct Roles for the XPB/p52 and XPD/p44 Subcomplexes of TFIIH in Damaged DNA Opening during Nucleotide Excision Repair, *Mol. Cell.* (2007). doi:10.1016/j.molcel.2007.03.009.
- [7] M.F. White, Biochemical Society Annual Symposium No . 76 Structure , function and evolution of the XPD family of iron – sulfur-containing 5 → 3 DNA helicases, (2009) 547–551. doi:10.1042/BST0370547.
- [8] J. Kuper, C. Braun, A. Elias, G. Michels, F. Sauer, D.R. Schmitt, A. Poterszman, J.M. Egly, C. Kisker, In TFIIH, XPD Helicase Is Exclusively Devoted to DNA Repair, *PLoS Biol.* (2014). doi:10.1371/journal.pbio.1001954.
- [9] S. Peisert, F. Sauer, D.B. Grabarczyk, C. Braun, G. Sander, A. Poterszman, J.M. Egly, J. Kuper, C. Kisker, In TFIIH the Arch domain of XPD is mechanistically essential for transcription and DNA repair, *Nat. Commun.* 11 (2020) 1667. doi:10.1038/s41467-020-15241-9.
- [10] D.R. Higgins, S. Prakash, P. Reynolds, Isolation and characterization of the RAD3 gene of *Saccharomyces cerevisiae* and inviability of rad3 deletion mutants, *Proc. Natl. Acad. Sci. U. S. A.* 80 (1983) 5680–5684. doi:10.1073/pnas.80.18.5680.
- [11] J. De Boer, I. Donker, J. De Wit, J.H.J. Hoeijmakers, G. Weeda, Disruption of the mouse xeroderma pigmentosum group D DNA repair/basal transcription gene results in preimplantation lethality, *Cancer Res.* 58 (1998) 89–94.
- [12] C.F.M. Menck, V. Munford, DNA repair diseases: What do they tell us about cancer and aging?, *Genet. Mol. Biol.* 37 (2014) 220–233. doi:10.1590/S1415-47572014000200008.
- [13] R. S. Feltrin, A.L.A. Segatto, T.A. de Souza, A.P. Schuch, Open gaps in the evolution of the eukaryotic nucleotide excision repair, *DNA Repair (Amst).* 95 (2020). doi:10.1016/j.dnarep.2020.102955.
- [14] K. Voskarides, H. Dweep, C. Chrysostomou, Erratum: Correction to: Evidence that DNA repair genes, a family of tumor suppressor genes, are associated with evolution rate and size of genomes (*Human genomics* (2019) 13 1 (26)), *Hum. Genomics.* 13 (2019) 29. doi:10.1186/s40246-019-0214-6.
- [15] R.O. Prum, J.S. Berv, A. Dornburg, D.J. Field, J.P. Townsend, E.M. Lemmon, A.R. Lemmon, A comprehensive phylogeny of birds (*Aves*) using targeted next-generation DNA sequencing, *Nature.* 526 (2015) 569–573. doi:10.1038/nature15697.
- [16] S.J. Hackett, R.T. Kimball, S. Reddy, R.C.K. Bowie, E.L. Braun, M.J. Braun, J.L.

- Chojnowski, W.A. Cox, K.L. Han, J. Harshman, C.J. Huddleston, B.D. Marks, K.J. Miglia, W.S. Moore, F.H. Sheldon, D.W. Steadman, C.C. Witt, T. Yuri, A phylogenomic study of birds reveals their evolutionary history, *Science* (80-.). 320 (2008) 1763–1768. doi:10.1126/science.1157704.
- [17] D.A. Benson, M. Cavanaugh, K. Clark, I. Karsch-mizrachi, D.J. Lipman, J. Ostell, E.W. Sayers, *GenBank*, 41 (2013) 36–42. doi:10.1093/nar/gks1195.
- [18] ERCC2 orthologs - NCBI, (n.d.).
<https://www.ncbi.nlm.nih.gov/gene/2068/ortholog/?scope=8782#genes-tab> (accessed March 25, 2020).
- [19] K.J. Mitchell, B. Llamas, J. Soubrier, N.J. Rawlence, T.H. Worthy, J. Wood, M.S.Y. Lee, A. Cooper, Ancient DNA reveals elephant birds and kiwi are sister taxa and clarifies ratite bird evolution, *Science* (80-.). 344 (2014) 898–900. doi:10.1126/science.1251981.
- [20] E.L. Braun, Innovation from reduction: gene loss, domain loss and sequence divergence in genome evolution., *Appl. Bioinformatics*. 2 (2003) 13–34.
- [21] D.M. Krylov, Y.I. Wolf, I.B. Rogozin, E. V. Koonin, Gene loss, protein sequence divergence, gene dispensability, expression level, and interactivity are correlated in eukaryotic evolution, *Genome Res*. 13 (2003) 2229–2235. doi:10.1101/gr.1589103.
- [22] T. Abe, M. Ooka, R. Kawasumi, K. Miyata, M. Takata, K. Hirota, D. Branzei, Warsaw breakage syndrome DDX11 helicase acts jointly with RAD17 in the repair of bulky lesions and replication through abasic sites, *Proc. Natl. Acad. Sci. U. S. A.* 115 (2018) 8412–8417. doi:10.1073/pnas.1803110115.
- [23] J.B. Vannier, G. Sarek, S.J. Boulton, RTEL1: Functions of a disease-associated helicase, *Trends Cell Biol.* 24 (2014) 416–425. doi:10.1016/j.tcb.2014.01.004.
- [24] E.D. Jarvis, S. Mirarab, A.J. Aberer, B. Li, P. Houde, C. Li, S.Y.W. Ho, B.C. Faircloth, B. Nabholz, J.T. Howard, A. Suh, C.C. Weber, R.R. Da Fonseca, J. Li, F. Zhang, H. Li, L. Zhou, N. Narula, L. Liu, G. Ganapathy, B. Boussau, M.S. Bayzid, V. Zavidovych, S. Subramanian, T. Gabaldón, S. Capella-Gutiérrez, J. Huerta-Cepas, B. Rekepalli, K. Munch, M. Schierup, B. Lindow, W.C. Warren, D. Ray, R.E. Green, M.W. Bruford, X. Zhan, A. Dixon, S. Li, N. Li, Y. Huang, E.P. Derryberry, M.F. Bertelsen, F.H. Sheldon, R.T. Brumfield, C. V. Mello, P. V. Lovell, M. Wirthlin, M.P.C. Schneider, F. Prosdocimi, J.A. Samaniego, A.M.V. Velazquez, A. Alfaro-Núñez, P.F. Campos, B. Petersen, T. Sicheritz-Ponten, A. Pas, T. Bailey, P. Scofield, M. Bunce, D.M. Lambert, Q. Zhou, P. Perelman, A.C. Driskell, B. Shapiro, Z. Xiong, Y. Zeng, S. Liu, Z. Li, B.

- Liu, K. Wu, J. Xiao, X. Yinqi, Q. Zheng, Y. Zhang, H. Yang, J. Wang, L. Smeds, F.E. Rheindt, M. Braun, J. Fjeldsa, L. Orlando, F.K. Barker, K.A. Jönsson, W. Johnson, K.P. Koepfli, S. O'Brien, D. Haussler, O.A. Ryder, C. Rahbek, E. Willerslev, G.R. Graves, T.C. Glenn, J. McCormack, D. Burt, H. Ellegren, P. Alström, S. V. Edwards, A. Stamatakis, D.P. Mindell, J. Cracraft, E.L. Braun, T. Warnow, W. Jun, M.T.P. Gilbert, G. Zhang, Whole-genome analyses resolve early branches in the tree of life of modern birds, *Science* (80-.). (2014). doi:10.1126/science.1253451.
- [25] R. Ali, S. Muhammad, M. Khan, L. Arvestad, Quantitative synteny scoring improves homology inference and partitioning of gene families., *BMC Bioinformatics*. 14 Suppl 1 (2013). doi:10.1186/1471-2105-14-S15-S12.
- [26] Q. Nguyen, M. Chenon, F. Vilela, C. Velours, M. Aumont-Nicaise, J. Andreani, P.F. Varela, P. Llinas, J. Ménétrey, Structural plasticity of the N-terminal capping helix of the TPR domain of kinesin light chain, *PLoS One*. 12 (2017) 1–20. doi:10.1371/journal.pone.0186354.
- [27] L. Yu, D.K. Tanwar, E.D.S. Penha, Y.I. Wolf, E. V. Koonin, M.K. Basu, Grammar of protein domain architectures, *Proc. Natl. Acad. Sci. U. S. A.* 116 (2019) 3636–3645. doi:10.1073/pnas.1814684116.

Supplementary material

Table S1: *blastx* and *tblastn* search results against XPD in other birds, using XPD of *A. mississippiensis* as query.

Order	Specie	Search	Bit score	e-value	Identity	Query cover	Accession	Reference genome
Tinamiformes	<i>Crypturellus cinnamomeus</i>	tblastn	101	3.00E-20	69%	19%	PTEZ01002580.1	cryCin1
	<i>Eudromia elegans</i>	blastx*	-	-	-	-	-	eudEle1
		tblastn	199	1.00E-70	60%	95%	PTEX01000752.1	
	<i>Nothoprocta perdicaria</i>	blastx	1417	0.0	91%	93%	XP_025905547.1	-
	<i>Tinamus guttatus</i>	blastx	475	2.00E-158	59%	78%	XP_010218114.1	-
Struthioniformes	<i>Struthio camelus</i>	blastx	762	0.0	80%	70%	XP_009666287.1	-
Galliformes	<i>Bambusicola thoracicus</i>	blastx	128	2.00E-30	27%	42%	POI30821.1	Bthov1.0
		tblastn**	45.8	1.40E-02	49%	12%	PPHD01010557.1	
	<i>Callipepla squamata</i>	blastx	110	7.00E-25	29%	36%	OXB59796.1	ASM221830v1
		tblastn	99.4	2.00E-20	90%	6%	MCFN01023643.1	
	<i>Centrocercus minimus</i>	blastx*	-	-	-	-	-	Cmin_1.0
		tblastn**	43.5	1.50E-02	36%	16%	SPOS01000007.1	
	<i>Chrysolophus pictus</i>	blastx*	-	-	-	-	-	Chrysolophus_pictus_GenomeV1.0
		tblastn	70.5	1.00E-10	73%	13%	KZ860042.1	
	<i>Colinus virginianus</i>	blastx	140	4.00E-34	29%	43%	OXB80435.1	Cv_LA_1.0
		tblastn**	50.1	1.00E-04	37%	13%	VONY01000613.1	
	<i>Coturnix japonica</i>	blastx	169	7.00E-43	25%	78%	XP_015736092.1	Coturnix japonica 2.1
		tblastn	77	9.00E-20	85%	13%	NW_015440188.1	
	<i>Lagopus muta</i>	blastx*	-	-	-	-	-	Lagopus muta japonica_ver1.0
		tblastn	107	4.00E-23	74%	8%	BJCA01029474.1	
	<i>Lyrurus tetrix</i>	blastx**	20.4	8.5	34%	4%	AFH75313.1	tetTet1
	tblastn**	47.4	5.00E-05	34%	9%	JDSL01332429.1		
<i>Meleagris gallopavo</i>	blastx	170	2.00E-43	22%	94%	XP_019477738.1	Turkey_5.1	
	tblastn	104	3.00E-22	83%	7%	NW_011216346.1		

	<i>Numida meleagris</i>	blastx	169	1.00E-42	23%	78%	XP_021271791.1	NumMel1.0	
		tblastn**	48.9	3.00E-04	91%	3%	NW_018363296.1		
	<i>Pavo cristatus</i>	blastx*	-	-	-	-	-	AIIM_Pcri_1.0	
		tblastn	57.4	1.00E-08	71%	5%	QZWQ01166974.1		
	<i>Phasianus colchicus</i>	blastx	194	7.00E-51	23%	88%	XP_031445156.1	ASM414374v1	
		tblastn	79	4.00E-16	69%	6%	NW_022221337.1		
	<i>Syrmaticus mikado</i>	blastx*	-	-	-	-	-	NTU_Smik_1.2	
		tblastn	70.1	6.00E-11	73%	5%	QGNR01004031.1		
	<i>Tympanuchus cupido</i>	blastx*	-	-	-	-	-	T_cupido_pinnatus_GPC_3440_v1	
		tblastn	101	2.00E-20	77%	14%	MOXI01000360.1		
	Anseriformes	<i>Anas platyrhynchos</i>	blastx	166	1.00E-41	24%	78%	XP_027327945.1	IASCAAS_PekingDuck_PBH1.5
			tblastn**	45.1	6.00E-03	55%	11%	NC_040065.1	
<i>Anas zonorhyncha</i>		blastx*	-	-	-	-	-	GCA_002224875.1	
		tblastn	142	4.00E-33	49%	62%	NOIK01000931.1		
<i>Anser brachyrhynchus</i>		blastx*	-	-	-	-	-	GCA_002592135.1	
		tblastn	109	4.00E-48	73%	16%	NXHY01002264.1		
<i>Anser cygnoides</i>		blastx	172	8.00E-44	24%	78%	XP_013053520.1	AnsCyg_PRJNA183603_v1.0	
		tblastn	64.3	9.00E-09	97%	4%	NW_013186212.1		
<i>Anser indicus</i>		blastx*	-	-	-	-	-	GCA_006229135.1	
		tblastn**	45.1	6.00E-03	55%	11%	VDDG01000143.1		
<i>Aythya fuligula</i>		blastx	166	5.00E-42	24%	78%	XP_032056908.1	bAytFul2.pri	
		tblastn	176	1.00E-43	89%	13%	NC_045564.1		
<i>Branta canadensis</i>		blastx*	-	-	-	-	-	GCA_006130075.1	
		tblastn	58.5	2.00E-07	92%	3%	ML628616.1		
<i>Cairina moschata</i>		blastx*	-	-	-	-	-	CaiMos1.0	
		tblastn*	-	-	-	-	-		
<i>Cygnus olor</i>		blastx*	-	-	-	-	-	bCygOlo1.pri	
		tblastn*	-	-	-	-	-		

*The corresponding searches resulted in no hit; **The corresponding searches have nonsignificant *e-values*.

Table S2: *tblastn* search results against genomic neighbors of *XPD*

Gene	BLAST results			Accession numbers		
	Bit score	e-value	Identity	Gene ID	Genomic	Protein
KLC3	127	2.00E-28	95%	423484	NC_006092.5	XP_025006858.1
PPP1R13L	91.3	5.00E-17	53%	423487	NC_006092.5	XP_015143326.1

4 DISCUSSÃO GERAL

A presente dissertação contém dois importantes pontos empregados para um melhor entendimento das lacunas da evolução da via NER em eucariotos: 1) uma análise computacional de proteínas-chave dessa via de reparo de DNA tanto em organismos-modelo quanto não-modelo, acompanhada de uma revisão de literatura sobre o seu mecanismo; e 2) um estudo buscando investigar a possível perda de uma importante enzima dessa via de reparo em Galloanseres, tanto na base da filogenia das aves quanto em alguns organismos do clado em questão. Além disso, é inquestionável que os achados deste trabalho têm o potencial de geração de pesquisas futuras baseados na metodologia utilizada, bem como nas hipóteses lançadas no decorrer deste trabalho, seja para esta como para demais vias de reparo de DNA, e até mesmo para a inferência de ortologia em outros processos moleculares.

O primeiro ponto deste trabalho traz contribuições importantes no sentido de identificar as lacunas de informação identificadas na via NER em organismos eucarióticos. Sendo assim, apesar de essa via estar presente desde o domínio Archaea, enfatizamos acontecimentos importantes que ocorreram durante a história evolutiva dessa via em eucariotos. Dentre eles, destaca-se a ausência da proteína XPA em plantas e protozoários, especialmente em *Giardia lamblia*, que carece de estudos nesse mecanismo, e em *Plasmodium falciparum*, para o qual ainda não havia sido reportada a aparente ausência de XPA (KIMURA e SAKAGUCHI, 2006; MACHADO et al., 2014; FELTRIN et al., 2020). Também enfatizamos que, segundo nossos critérios estruturais e de similaridade, a proteína XPE não foi obtida nos invertebrados analisados, podendo indicar uma divergência. Entretanto, o resultado mais intrigante foi não termos encontrado, utilizando nossa metodologia, uma proteína XPD canônica em *G. gallus*.

Sendo assim, o segundo ponto deste estudo teve como foco justamente investigar essa aparente ausência de XPD em galinhas. Além de estar presente até mesmo na espécie de cnidário *Hydra* (GALANDE et al., 2018), XPD tem um papel muito importante tanto no reparo quanto na transcrição (COMPE e EGLY, 2016), sendo estes ambos processos fundamentais para a manutenção da fidelidade e do fluxo da informação genética. Dessa forma, por meio da busca de XPD em quatro ordens taxonômicas, recuperamos sequências ortólogas de XPD na base da filogenia

das aves. Porém, seguindo os mesmos critérios de busca empregados no primeiro artigo, também não foi evidenciada a presença de uma sequência de XPD canônica no grupo-irmão que inclui Galliformes e Anseriformes. Adicionalmente, a análise do contexto genômico de *ERCC2* em *A. mississippiensis* não demonstrou claramente que haja uma relação sintênica entre essa região na espécie de crocodiliano e em *G. gallus*.

Como perspectivas, pretende-se: (i) ampliar o número de espécies analisadas na busca por homologia e também por sintenia, principalmente em Struthioniformes e, além disso, em Neoaves; (ii) utilizar uma metodologia mais robusta para investigar sintenia; (iii) realizar buscas por outros componentes do TFIIH que interagem com XPD durante o NER; e (iv) realizar análises filogenéticas para clarear as relações evolutivas entre as sequências. Com isso, espera-se que a possível ausência de XPD em Galloanseres possa ser melhor compreendida.

5 CONCLUSÃO

O reparo por excisão de nucleotídeos (NER), possivelmente em razão de ser o mecanismo de reparo de DNA mais versátil, faz-se presente em grande parte dos ramos da árvore da vida. Ademais, existe um consenso na literatura de que o NER tenha sido muito conservado ao longo da evolução das espécies, especialmente em eucariotos. Entretanto, neste trabalho demonstramos diferenças importantes entre o NER de humanos e de eucariotos evolutivamente distantes. Apesar de a maioria dos componentes dessa via estarem presentes nas espécies eucarióticas avaliadas, observa-se uma heterogeneidade nas estruturas gênicas, bem como uma variação na arquitetura das proteínas entre esses organismos. Tais achados nos permitem supor que ocorra um diferente funcionamento da via NER nesses organismos, abrindo perspectivas para o desenvolvimento de estudos funcionais. Além disso, apesar de termos encontrado resultados essencialmente preditivos em relação à possível ausência de XPD em Galloanseres, demos um passo considerável a fim de traçar novos horizontes de busca.

REFERÊNCIAS

- BEERENS, N. et al. The CSB Protein Actively Wraps DNA. **The Journal of Biological Chemistry**, v. 280, n. 6, p. 4722–4729, 2005.
- BERGINK, S. et al. Recognition of DNA damage by XPC coincides with disruption of the XPC–RAD23 complex. **Journal of Cell Biology**, v. 196, n. 6, p. 681–688, 2012.
- BRADFORD, P.T. et al. Cancer and neurological degeneration in Xeroderma Pigmentosum: long term follow-up characterizes the role of DNA repair. **Journal of Medical Genetics**, v. 48, i. 3, p. 168-176, 2011.
- CHATTERJEE, N.; WALKER, G. C. Mechanisms of DNA damage, repair, and mutagenesis. **Environmental and Molecular Mutagenesis**, v. 58, n. 5, p. 235–263, 2017.
- COMPE, E.; EGLY, J. Nucleotide Excision Repair and Transcriptional Regulation: TFIIH and Beyond, v. 85, p.265-290, 2016.
- COSTA, R. M. A. et al. The eukaryotic nucleotide excision repair pathway. **Biochimie**, v. 85, n. 11, p. 1083–1099, 2003.
- DE LAAT, W. L. et al. DNA-binding polarity of human replication protein A positions nucleases in nucleotide excision repair. **Genes & Development**, v. 12, n. 16, p. 2598–2609, 1998.
- DE LAAT, W. L.; JASPERS, N. G. J.; HOEIJMAKERS, J. H. J. Molecular mechanism of nucleotide excision repair. **Genes & Development**, v. 13, n. 7, p. 768–785, 1999.
- FAGBEMI, A. F.; ORELLI, B.; SCHÄRER, O. D. Regulation of endonuclease activity in human nucleotide excision repair. **DNA Repair**, v. 10, i. 7, p. 722-729, 2011.
- FELTRIN, R.S. et al. Open gaps in the evolution of the eukaryotic nucleotide excision repair. **DNA Repair**, v. 95, n. 102955, 2020.
- FITCH, M. E. et al. In Vivo Recruitment of XPC to UV-induced Cyclobutane Pyrimidine Dimers by the DDB2 Gene Product. **The Journal of Biological Chemistry**, v. 278, n. 47, p. 46906–46910, 2003.
- FOUSTERI, M.; MULLENDERS, L. H. F. Transcription-coupled nucleotide excision repair in mammalian cells: molecular mechanisms and biological effects. **Cell Research**, v. 18, n. 1, p. 73–84, 2008.
- KIMURA, S.; SAKAGUCHI, K. DNA repair in plants. **Chemical Reviews**, v. 106, i. 2, p. 753–766, 2006.
- GALANDE, A. A. et al. Analysis of the conserved NER helicases (XPB and XPD) and UV-induced DNA damage in *Hydra*. **BBA – General Subjects**, v. 1862, i. 9, p. 2031–2042, 2018.
- GROISMAN, R. et al. CSA-dependent degradation of CSB by the ubiquitin-proteasome pathway establishes a link between complementation factors of the Cockayne syndrome. **Genes & Development**, v. 20, n. 11, p. 1429–1434, 2006.
- HANAWALT, P. C. The bases for Cockayne syndrome. **Nature**, v. 405, n. 6785, p. 415–415, 2000.

HANAWALT, P. C. Subpathways of nucleotide excision repair and their regulation. **Oncogene**, v. 21, n. 58, p. 8949–8956, 2002.

HOEIJMAKERS, J. H. J., Genome maintenance mechanisms for preventing cancer. **Nature**, v. 411, n. 6835, p. 366–374, 2001.

JAARSMA, D. et al. Cockayne syndrome pathogenesis: Lessons from mouse models, **Mechanisms of Ageing and Development**, v. 134, n. 5–6, p. 180–195, 2013.

KAMILERI, I.; KARAKASILIOTI, I.; GARINIS, G. A. Nucleotide excision repair: new tricks with old bricks. **Trends in Genetics**, v. 28, n. 11, p. 566–573, 2012.

KEMP, M. G. et al. Mechanism of Release and Fate of Excised Oligonucleotides during Nucleotide Excision Repair. **The Journal of Biological Chemistry**, v. 287, n. 27, p. 22889–22899, 2012.

KUMAR, N. et al. Cooperation and interplay between base and nucleotide excision repair pathways: From DNA lesions to proteins. **Genetics and Molecular Biology**, v. 43, n. 1 suppl 1, p. e20190104, 2020.

LEE, Y. et al. The relationships between XPC binding to conformationally diverse DNA adducts and their excision by the human NER system: Is there a correlation? **DNA Repair**, v. 19, p. 55–63, 2014.

MACHADO, C. R. et al. Nucleotide excision repair in *Trypanosoma brucei*: specialization of transcription-coupled repair due to multigenic transcription: Specialization of *T. brucei* nucleotide excision repair. **Molecular Microbiology**, v. 92, n. 4, p. 756–776, 2014.

MARTEIJN, J. A. et al. Understanding nucleotide excision repair and its roles in cancer and ageing. **Nature Reviews Molecular Cell Biology**, v. 15, n. 7, p. 465–481, 2014.

MASUDA, H. A. et al. Evidence for sub-functionalization of tandemly duplicated XPB nucleotide excision repair genes in *Arabidopsis thaliana*. **Gene**, v. 754, n. 144818, 2020.

MCCREADY, S. J.; OSMAN, F.; YASUI, A. Repair of UV damage in the fission yeast *Schizosaccharomyces pombe*. **Mutation Research/Fundamental and Molecular Mechanisms of Mutagenesis**, v. 451, n. 1–2, p. 197–210, 2000.

MELLON, I.; SPIVAK, G.; HANAWALT, P. C. Selective removal of transcription-blocking DNA damage from the transcribed strand of the mammalian DHFR gene. **Cell**, v. 51, n. 2, p. 241–249, 1987.

MENCK, C. F. M. Shining a light on photolyases. **Nature Genetics**, v. 32, 338–339, 2002.

MENCK, C. F. M.; MUNFORD, V. DNA repair diseases: what do they tell us about cancer and aging? **Genetics and Molecular Biology**, v. 37, n. 1 suppl 1, p. 220–233, 2014.

MORENO, N. C.; SOUZA, T. A. et al. Whole-exome sequencing reveals the impact of UVA light mutagenesis in xeroderma pigmentosum variant human cells. **Nucleic Acids Research**, v. 48, i. 4, p. 1941–1953, 2019.

MULLENDERS, L. H. F. Solar UV damage to cellular DNA: from mechanisms to biological effects. **Photochemical & Photobiological Sciences**, v. 17, n. 12, p. 1842–1852, 2018.

PANI, B.; NUDLER, E. Mechanistic insights into transcription coupled DNA repair. **DNA Repair**, v. 56, p. 42–50, 2017.

PASCUCCI, B. et al. Role of nucleotide excision repair proteins in oxidative DNA damage repair: an updating. **Biochemistry (Moscow)**, v. 76, n. 1, p. 4–15, 2011.

PETIT, C.; SANCAR, A. Nucleotide excision repair: From E. coli to man. **Biochimie**, v. 81, n. 1–2, p. 15–25, 1999.

RASTOGI, R. P. et al. Molecular Mechanisms of Ultraviolet Radiation-Induced DNA Damage and Repair. **Journal of Nucleic Acids**, v. 2010, p. 1–32, 2010.

ROUILLON, C.; WHITE, M. F. The evolution and mechanisms of nucleotide excision repair proteins. **Research in Microbiology**, v. 162, n. 1, p. 19–26, 2011.

SCHÄRER, O. D. Nucleotide Excision Repair in Eukaryotes. **Cold Spring Harbor Perspectives in Biology**, v. 5, n. 10, p. a012609–a012609, 2013.

SCHUCH, A. P. et al. Sunlight damage to cellular DNA: Focus on oxidatively generated lesions. **Free Radical Biology and Medicine**, v. 107, p. 110–124, 2017.

SCHWERTMAN, P.; VERMEULEN, W.; MARTEIJN, J. A. UVSSA and USP7, a new couple in transcription-coupled DNA repair. **Chromosoma**, v. 122, n. 4, p. 275–284, 2013.

SHIVJI, M. K. K.; KENNY, M. K.; WOOD, R. D. Proliferating cell nuclear antigen is required for DNA excision repair. **Cell**, v. 69, n. 2, p. 367–374, 1992.

SPAMPINATO, C. P. Protecting DNA from errors and damage: an overview of DNA repair mechanisms in plants compared to mammals. **Cellular and Molecular Life Sciences**, v. 74, n. 9, p. 1693–1709, 2017.

SPIVAK, G. Nucleotide excision repair in humans. **DNA Repair**, v. 36, p. 13–18, 2015.

SPIVAK, G. Transcription-coupled repair: an update. **Archives of Toxicology**, v. 90, n. 11, p. 2583–2594, 2016.

SPIVAK, G.; GANESAN, A. K. The complex choreography of transcription-coupled repair. **DNA Repair**, v. 19, p. 64–70, 2014.

TAJEDIN, L. et al. Comparative insight into nucleotide excision repair components of *Plasmodium falciparum*. **DNA Repair**, v. 28, p. 60–72, 2015.

TUBBS, A.; NUSSENZWEIG, A. Endogenous DNA Damage as a Source of Genomic Instability in Cancer. **Cell**, v. 168, n. 4, p. 644–656, 2017.

UCHIDA, A. et al. The carboxy-terminal domain of the XPC protein plays a crucial role in nucleotide excision repair through interactions with transcription factor IIH. **DNA Repair**, v. 1, n. 6, p. 449–461, 2002.

VERMEULEN, W.; FOSTERI, M. Mammalian Transcription-Coupled Excision Repair. **Cold Spring Harbor Perspectives in Biology**, v. 5, n. 8, p. a012625–a012625, 2013.

ZHANG, C. et al. Arabidopsis Cockayne Syndrome A-Like Proteins 1A and 1B Form a Complex with CULLIN4 and Damage DNA Binding Protein 1A and Regulate the Response to UV Irradiation. **The Plant Cell**, v. 22, n. 7, p. 2353–2369, 2010.