DOUTORAMENTO

CIÊNCIAS BIOMÉDICAS

Molecular evolution of DNA topoisomerases in animals

Filipa Moreira

D

2022

Filipa Moreira.Molecular evolution of DNA topoisomerases in animals

D.ICBAS 2022

Molecular evolution of DNA topoisomerases in animals

Filipa Isabel Menezes Moreira

INSTITUTO DE CIÊNCIAS BIOMÉDICAS ABEL SALAZAR

FILIPA ISABEL MENEZES MOREIRA

MOLECULAR EVOLUTION OF DNA TOPOISOMERASES IN ANIMALS

Tese de Candidatura ao grau de Doutor em Ciências Biomédicas submetida ao Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto

Orientador – Doutor Filipe Pereira
Categoria – Investigador
Afiliação – Centro de Ecologia Funcional, Universidade de Coimbra e IDENTIFICA Genetic Testing

Co-orientador – Professor Doutor Arnaldo Videira
Categoria – Professor Catedrático
Afiliação – Instituto de Ciências Biomédicas Abel Salazar da Universidade do Porto

FILIPA ISABEL MENEZES MOREIRA

MOLECULAR EVOLUTION OF DNA TOPOISOMERASES IN ANIMALS

Dissertation in fulfilment of the requirements for the PhD degree in Biomedical Sciences, submitted to Instituto de Ciências Biomédicas Abel Salazar of University of Porto

Supervisor – Filipe Pereira, PhD
Category – Researcher
Affiliation – Centre for Functional Ecology, University of Coimbra and IDENTIFICA Genetic Testing

Co-Supervisor – Professor Arnaldo Videira
Category – Full Professor
Affiliation – Instituto de Ciências Biomédicas Abel Salazar of the University of Porto

The author was in receipt of a PhD Studentship from FCT (POPH - QREN/FSE funding, (SFRH/BD/131584/2017).

“Science is the acceptance of what works and the rejection of what does not.
That needs more courage than we might think”

JACOB BRONOWSKI

Declaração de Honra

Declaro que a presente tese é da minha autoria e não foi utilizada previamente noutro curso ou unidade curricular, desta ou de outra instituição. As referências a outros autores (afirmações, ideias, pensamentos) respeitam escrupulosamente as regras da atribuição e encontram-se devidamente indicadas no texto e nas referências bibliográficas, de acordo com as normas de referenciação. Tenho consciência de que a prática de plágio e auto-plágio constitui um ilícito académico.

Assinado por: **FILIPA ISABEL MENEZES MOREIRA**
Num. de Identificação: 13546636
Data: 2022.10.09 17:03:57+01'00'

CARTÃO DE CIDADÃO

Scientific Output

According to Decreto-Lei no. 230/2009, de 14 de Setembro, I declare to have participated in the design and accomplishment of the experimental work, as well as in the interpretation of the results and in the writing of the following works published and in the phase of publication that integrate this thesis.

Articles published in journals indexed in ISI Web of Knowledge:

1- **Filipa Moreira**, Miguel Arenas, Arnaldo Videira, Filipe Pereira. Molecular Evolution of DNA Topoisomerase III Beta (TOP3B) in Metazoa. Journal of Molecular Evolution, 2021, 89.6: 384-395. DOI: https://doi.org/10.1007/s00239-021-10011-7

2- **Filipa Moreira**, Miguel Arenas, Arnaldo Videira, Filipe Pereira. Evolutionary History of TOPIIA Topoisomerases in Animals. Journal of Molecular Evolution, 2022, 90.2: 149-165. DOI: https://doi.org/10.1007/s00239-022-10048-2

3- **Filipa Moreira**, Miguel Arenas, Arnaldo Videira, Filipe Pereira. Evolution of *TOP1* and *TOP1MT* topoisomerases in chordates. Journal of Molecular Evolution. 2023. https://doi.org/10.1007/s00239-022-10091-z

Author contribution: Filipa Moreira conceived and designed the experiments, collected data, performed data analysis and interpretation, prepared figures and tables, drafted the work and critically revised the final work to be published.

Acknowledgments

I am extraordinarily thankful to my supervisor, Dr. Filipe Pereira, as his undying love and passion for Science made me believe that I, too, might strive to appreciate this PhD journey. His unwavering determination, competence, and everlasting dedication were paramount in helping me overcome the obstacles I encountered along the way. He is the greatest mentor I could have asked for, and I would welcome the opportunity to continue learning from him.

I additionally want to express my sincere gratitude to my co-supervisor, Professor Dr. Arnaldo Videira, for his guidance and support. He was always quick and decisive in the moments I needed. His knowledge and commitment were crucial in guiding me in the right direction. I also thank him for his hospitality in welcoming me in his research group.

I also thank Dr. Miguel Arenas for his effort and valuable insights in the preparation of all my PhD publications. His knowledge was instrumental to the quality of the work delivered.

A special thank you goes to my family and friends. They were relentless in pushing me to succeed and believing in my capabilities, even at times when I did not do it for myself. Their support was very important during this period of my life.

Table of Contents

Abstract 1
Resumo 3
1. Introduction 6
  1.1. The origin of DNA topoisomerases 6
  1.2. DNA topological problems solved by topoisomerases 7
  1.3 Types of DNA topoisomerases 8
  1.4. Type I Topoisomerases 10
    1.4.1. Type IA topoisomerases 11
      1.4.1.1. Topoisomerase IIIα (TOP3A) 14
      1.4.1.2. Topoisomerase IIIβ (TOP3B) 15
    1.4.2. Type IB topoisomerases 15
      1.4.2.1. TOP1 17
      1.4.2.2. TOP1MT 17
  1.5. Type II Topoisomerases 18
    1.5.1. Type IIA 19
      1.5.1.1 TOP2A and TOP2B 20
  1.6. Resistance to topoisomerase-targeted drugs 22
  1.7. The study of topoisomerase genes throughout the animal kingdom 23
  1.8. References 25
2. Objectives 34
3. Molecular evolution of DNA topoisomerase III beta (TOP3B) in Metazoa (Scientific Publication nº 1) 36
  3.1. Abstract 37
  3.2. Introduction 38
  3.3. Material and Methods 40
  3.4. Results and Discussion 43
  3.5. Conclusions 54
  3.6. References 55
  3.7. Supplementary Material 58
4. Evolutionary history of TOPIIA topoisomerases in animals (Scientific Publication nº 2) 67
  4.1. Abstract 68
  4.2. Introduction 69
  4.3. Material and Methods 73
  4.4. Results and Discussion 76
  4.5. Conclusions 91
  4.6. References 92
  4.7. Supplementary Material 98
5. Evolution of *TOP1* and *TOP1MT* topoisomerases in Chordata (Scientific Publication nº 3) 105
  5.1. Abstract 106
  5.2. Introduction 107
  5.3. Material and Methods 110
  5.4. Results and Discussion 112
  5.5. Conclusions 121
  5.6 References 122
  5.7. Supplementary Material 126
6. General Discussion 132
  6.1. The use of topoisomerases for phylogenetic inferences 132
  6.2. Topoisomerases evolve under strong purifying selection 136
  6.3. Variation in topoisomerases amongst modern and archaic humans 138

6.4. Evolution of topoisomerase domains 141
6.5. Mutations in topoisomerases causing disease 143
6.6. References 147
7. Conclusion 152

Abstract

Topoisomerases have long been regarded as fascinating components of the DNA world by opening, moving and closing the DNA strands without leaving a trace. They introduce transient breaks in the DNA phosphodiester backbones to avoid DNA entanglements such as supercoils, knots and catenanes. The activities of topoisomerases are crucial for controlling DNA topology during replication, transcription, recombination and chromatin remodelling, among other cellular transactions. DNA topoisomerases are also of great relevance in pharmacology and clinical medicine by being the molecular targets of many antimicrobial and anticancer agents. Despite their functional and clinical relevance, the evolution of topoisomerases is still poorly understood. Previous works have focused on the origin of the different types of topoisomerases in bacteria and viruses, with only a few studies dedicated to eukaryotes. In this dissertation, we provide a detailed investigation of the molecular evolution of DNA topoisomerases in animals. The Type IA (TOP3B), Type IIA (TOP2A and TOP2B) and Type IB (TOP1 and TOP1MT) topoisomerases were studied by a combination of phylogenetic, comparative genomics and structural analyses. We found that topoisomerases are conserved across Metazoa and can be used as informative genetic markers for deep phylogenetic inferences. We confirmed that Type IIA and Type IB paralogues are exclusive of vertebrates and possibly had different origins early in the radiation of this group, associated with whole-genome duplication events. The *TOP2A* and *TOP2B* paralogues from jawed and jawless vertebrates form separate clusters, while TOP1 and TOP1MT from both groups branch close to each other. We found strong signs of purifying selection acting in all topoisomerases, expected due to their essential biological roles. The only topoisomerase (TOP1MT) believed to act exclusively in mitochondria showed the highest nucleotide diversity and substitution rate among all topoisomerases. We also detected signs of positive selection in a few *TOP3B* and *TOP2B* sites located in peripheral regions of the protein that may be relevant for interaction with other cellular components. The comparison of the topoisomerases from archaic and modern humans revealed four mutations in *TOP2A* and *TOP1MT* that result in amino acids replacements. These mutations may have contributed to differences among the species in the cell division complex of some tissues. Despite an overall high conservation, topoisomerases can have domains and motifs that vary considerably. We found a high conservation in catalytic regions and some linker regions. The N- or C-terminal domains and localization signals varied considerably among topoisomerases, suggesting different evolutionary trajectories across metazoans. Critical active sites were found particularly conserved,

including those associated with resistance to anticancer agents. Some of these drug-resistant mutations were found in sites that differ among paralogues, which may be used to design specific inhibitors to only one of the paralogues, avoiding the side effects of blocking the other. Altogether, this dissertation helps to clarify the origin, evolution and selection pressures governing the evolution of animal topoisomerases.

Keywords

DNA topoisomerase, Metazoa phylogeny, purifying selection, archaic humans, functional divergence, drug-resistance mutations

Resumo

As topoisomerases têm sido consideradas desde há muito tempo como componentes fascinantes do mundo do DNA pelo facto de abrirem, moverem e fecharem as cadeias de DNA sem deixar vestígios. Elas introduzem quebras transitórias nas ligações fosfodiéster do DNA para evitar enrolamentos no DNA, tais como super-espirais, nós e catenanos. As atividades das topoisomerases são cruciais para controlar a topologia do DNA durante a replicação, transcrição, recombinação e remodelação da cromatina, entre outras transações celulares. As topoisomerases são também de grande relevância em farmacologia e medicina clínica por serem os alvos moleculares de muitos agentes antimicrobianos e anticancerígenos. Apesar da sua relevância funcional e clínica, a evolução das topoisomerases é ainda mal compreendida. Os trabalhos anteriores focaram-se na origem dos diferentes tipos de topoisomerases em bactérias e vírus, com apenas alguns estudos dedicados aos eucariotas. Nesta dissertação, investigámos detalhadamente a evolução molecular das topoisomerases nos animais. As topoisomerases do Tipo IA (TOP3B), Tipo IIA (TOP2A e TOP2B) e Tipo IB (TOP1 e TOP1MT) foram estudadas através de uma combinação de análises filogenéticas, genómica comparativa e análises estruturais. Verificámos que as topoisomerases são conservadas nos Metazoa e podem ser usadas como marcadores genéticos informativos para inferências filogenéticas. Confirmámos que os parálogos de Tipo IIA e Tipo IB são exclusivos dos vertebrados e possivelmente tiveram diferentes origens no início da radiação deste grupo, associados a eventos de duplicação de todo o genoma. Os parálogos de TOP2A e TOP2B de vertebrados com e sem mandíbula formam grupos separados, enquanto os parálogos de TOP1 e TOP1MT de ambos os grupos se ramificam juntamente. Encontrámos fortes sinais de seleção purificadora atuando em todas as topoisomerases, de acordo com os seus papéis biológicos essenciais. A única topoisomerase (TOP1MT) que se acredita atuar exclusivamente nas mitocôndrias foi a que revelou uma maior diversidade nucleotídica e taxa de substituição entre todas as topoisomerases. Foram também detetados sinais de seleção positiva em alguns locais da TOP3B e TOP2B, nomeadamente em regiões periféricas da proteína, que podem ser relevantes para interações com outros componentes celulares. A comparação das topoisomerases de humanos arcaicos e modernos revelou quatro mutações na TOP2A e TOP1MT que resultam em substituições de aminoácidos. Estas mutações podem ter contribuído para diferenças entre as espécies no complexo de divisão celular de alguns tecidos. Apesar de uma conservação global elevada, as topoisomerases podem ter domínios e motivos que variam consideravelmente. Encontrámos uma elevada

conservação em regiões catalíticas e em algumas regiões de ligação. Os domínios N- ou C- terminais e os sinais de localização variam consideravelmente entre topoisomerases, sugerindo diferentes trajetórias evolutivas nos metazoários. Foram encontrados sítios críticos ativos particularmente conservados, incluindo os associados à resistência a fármacos anticancerígenos. Algumas destas mutações resistentes aos medicamentos foram encontradas em locais que diferem entre parálogos, os quais podem ser utilizados para conceber inibidores específicos apenas para um dos parálogos, evitando os efeitos secundários de bloquear o outro. No conjunto, esta dissertação ajuda a esclarecer a origem, evolução e pressões seletivas que regeram a evolução das topoisomerases nos animais.

Palavras-chave

Topoisomerase de DNA, filogenia dos Metazoa, seleção purificante, humanos arcaicos, divergência funcional, mutações de resistência a drogas

Chapter 1
Introduction

1. Introduction

1.1. The origin of DNA topoisomerases

Since the discovery of the first DNA topoisomerase in 1971, *Escherichia coli* DNA topoisomerase I (Wang 1971), several forms of topoisomerases have been found in all taxonomic groups. All bacteria, archaea, and eukaryotic cells have a set of DNA topoisomerases that are capable of releasing negative and positive supercoils, as well as intermolecular catenation, thus attempting to resolve topological by-products of DNA metabolism (Bizard and Hickson 2020; Champoux 2001).

The presence of different topoisomerase families within the three domains of life do not follow the usual distribution of other informative proteins, such as ribosomal proteins or ATP synthases (Da Cunha et al. 2017). For example, the origin of one eukaryotic type of topoisomerases (IIA) remains enigmatic considering they have no orthologues in Archaea and are very different from the bacterial homologues (Forterre et al. 2007). According to one evolutionary hypothesis, life arose from self-replicating RNAs, and an RNA world with an RNA genome existed before the current DNA world (Poole and Logan 2005). According to other theories, the Last Universal Common Ancestor (LUCA) may also include organisms with RNA genomes, RNA-DNA hybrids, or DNA with an RNA replication intermediate (Forterre 2006; Forterre et al. 2007; Leipe et al. 1999). Whatever the case, it is clear that different types of topoisomerases already existed in the most recent common ancestors of each of the three cellular domains of life (Archaea, Bacteria and Eukarya). Some topoisomerases types were only acquired in a particular taxonomic group, possibly by horizontal gene transfers (Forterre and Gadelle 2009; Forterre et al. 2007). It has been also suggested that topoisomerases originated in ancestral viruses and were subsequently transferred independently to different ancient cellular lineages (Forterre and Gadelle 2009; Forterre et al. 2007).

Novel topoisomerases were also created more recently in evolutionary terms by gene or genome duplications. For instance, vertebrates have paralogues for two main types of topoisomerases (Rosa et al. 2009; Vos et al. 2011). Gene duplications have been commonly thought to be the most important step for the origin of genetic innovations, because it creates gene copies whose functions can subsequently evolve in divergent directions (Ohno 2013). The duplication of genes could be the result of whole genome duplications. The most commonly used model to explain the evolution of the vertebrate genome is the ‘one-two-four’ (or 1-2-4) rule. It assumes that the genome underwent two rounds of duplication (1R and 2R) leading from a single ancestral deuterostome genome

to two after the first duplication, predating the Cambrian explosion, and then to four genomes after the second genome duplication, possibly dating back to the early Devonian (Meyer and Schartl 1999; Meyer and Van de Peer 2005). Later in the Devonian period, the fish genome was duplicated for a third time to produce up to eight copies of the original deuterostome genome (3R) (Meyer and Van de Peer 2005). This last duplication took place after the two major radiations of jawed vertebrate life, the ray-finned fish (Actinopterygii) and the sarcopterygian lineage, occurred. Therefore, the sarcopterygian fish, which includes the coelacanth, lungfish and all land vertebrates such as amphibians, reptiles, birds and mammals, tend to have only half the number of genes compared with actinopterygian fish (Amores et al. 1998).

The evolution of topoisomerases is still an area of intense research and the origin of some families remains mysterious. The recent emergence of paralogues can be more easily understood by applying a phylogenetic approach using the growing number of genomic sequences currently available.

1.2. DNA topological problems solved by topoisomerases

Controlling the DNA topology is extremely important for most biochemical processes. The genome spatial architecture is intricately tied to its biological function (Dixon et al. 2012). For example, the formation of 'topological domains' (large-scale local chromatin interaction domains) requires topologically constrained DNA, being subjected to torsional tension or supercoiling. These constraints can occur at various levels of the genome organization (Vinograd and Lebowitz 1966). The DNA right-handed double helix polymer is folded and significantly compacted into the cell nucleus, thus preventing the separation of the complementary strands in order to safeguard the protection of the genetic material. Moreover, the functionality of the DNA must also be guaranteed, by making it accessible to RNA and DNA polymerases (Bizard and Hickson 2020). Diverse cellular processes, including DNA replication, transcription, recombination, and chromatin remodelling, require the collective action of multiple protein-DNA interactions (Siggers and Gordân 2014). Once a protein binds to DNA, it can serve as a foundation for further protein-DNA complexes setting up a combined network of topologically constrained DNA (Whyte et al. 2013). The proteins that bind and manipulate the DNA inevitably convey torsional stress, which is reallocated over constrained DNA regions and has a significant influence on its global conformation (Liu and Wang 1987).

The topological problems in DNA molecules that result from all these cellular processes are resolved by the action of DNA topoisomerases. Without these important enzymes,

positive supercoiling ahead of the DNA template quickly stalls replication and transcription, and negative supercoiling behind the DNA template favours the formation of abnormal DNA structures including D loops (invasion of a DNA duplex by a complementary single-stranded DNA segment), R loops (persistent annealing of RNA with its DNA template behind RNA polymerase), guanosine quartets, and Z-DNA, all of which interfere with normal DNA metabolism (Pommier et al. 2010).

All topoisomerases act by cleaving and re-joining the nucleic acid backbone using a tyrosine nucleophilic residue. Each break results from the formation of a covalent bond between the topoisomerase catalytic Tyr residue and one end of the broken nucleic acid. The covalent catalytic intermediates are referred to as TOP cleavage complexes or TOPccs. This bond can be self-reversed by another transesterification reaction, as the deoxyribose hydroxyl ends of the cleaved molecule act as nucleophiles towards the tyrosyl–DNA phosphodiester bonds, allowing the cut to be resealed and the enzyme to be released (Bizard and Hickson 2020; Champoux 2001; Pommier et al. 2016). If the TOPccs fail to reseal, they become topoisomerase DNA–protein crosslinks (TOP-DPCs), which require DNA repair pathways for their resolution. Trapping of TOP-DPCs is the mechanism of action of widely used anticancer and antibacterial chemotherapies (Nitiss 2009; Thomas and Pommier 2019; Vann et al. 2021).

1.3 Types of DNA topoisomerases

As the plethora of organisms evolved different strategies to manage topological problems, there is a wide variety of topoisomerases that reflect their specialization to particular DNA transactions (Takahashi et al. 2022). Topoisomerases are crucial enzymes to support cell viability and chromosome topology, as they have the ability to cut, rearrange and reconnect DNA strands, by the addition or removal of DNA supercoils and the disentanglement of DNA segments (Champoux 2001). They can preferentially distinguish between different types of DNA conformations and collaborate with various factors, such as protein expression levels and subcellular localization to direct topoisomerase action (Vos et al. 2011).

There are two principal classes that all topoisomerases can be allocated to, Type I and Type II, depending on whether they correspondingly cleave one or two strands of DNA. Topoisomerase subtypes - A, B or C - are then used to differentiate across enzyme families with different amino acid sequences and/or global structures (Forterre et al. 2007). Based on their sequence comparison, these enzymes can be divided into five evolutionary distinct families - IA, IB, IC, IIA, and IIB - with various folds and reaction

mechanisms (Forterre and Gadelle 2009). Humans encode six topoisomerases in different chromosomes (Table 1.1): TOP3A and TOP3B (Type IA topoisomerases), TOP1 and TOP1MT (Type IB topoisomerases) and TOP2A and TOP2B (Type IIA topoisomerases) (Pommier et al. 2010; Wang 2002).

Table 1.1. Name and genomic location of human topoisomerase genes and some of their orthologues.

<table>
<tr><th rowspan="2">Type</th><th rowspan="2">Abbreviated name</th><th rowspan="2">Full name</th><th rowspan="2">Chromosome band</th><th colspan="3">Orthologues</th></tr>
<tr><th>Yeast</th><th>Fruit fly</th><th>E. coli</th></tr>
<tr><td rowspan="2">IB</td><td>TOP1</td><td>DNA topoisomerase I</td><td>20q12</td><td>Top1</td><td>Top1</td><td>N.A</td></tr>
<tr><td>TOP1MT</td><td>DNA topoisomerase I mitochondrial</td><td>8q24.3</td><td>N.A</td><td>N.A</td><td>N.A</td></tr>
<tr><td rowspan="2">IIA</td><td>TOP2A</td><td>DNA topoisomerase II alpha</td><td>17q21.2</td><td rowspan="2">Top2</td><td rowspan="2">Top2</td><td rowspan="2">Gyrase, Topo IV</td></tr>
<tr><td>TOP2B</td><td>DNA topoisomerase II beta</td><td>3p24.2</td></tr>
<tr><td rowspan="2">IA</td><td>TOP3A</td><td>DNA topoisomerase III alpha</td><td>17p11.2</td><td rowspan="2">Top3</td><td>Top3a</td><td rowspan="2">Topo I, TopoIII</td></tr>
<tr><td>TOP3B</td><td>DNA topoisomerase III beta</td><td>22q11.22</td><td>Top3b</td></tr>
</table>

The polarity (3′ versus 5′ tyrosyl linkage); substrate specificity (single-stranded versus double-stranded DNA or RNA); nucleic acid relaxation mechanism (DNA strand rotation versus DNA crossover inversion or RNA crossover inversion); and cofactor requirements (ATP and/or magnesium) are the four biochemical features that distinguish these topoisomerases (Pommier et al. 2022), as shown in Table 1.2.

Table 1.2. Main features of all human topoisomerases.

	TYPE IA		TYPE IB		TYPE IIA	
	TOP3A	**TOP3B**	**TOP1**	**TOP1MT**	**TOP2A**	**TOP2B**
Cellular localization	Nucleus, Mitochondria	Nucleus, Cytosol	Nucleus	Mitochondria	Nucleus, Mitochondria	Nucleus, Mitochondria
5´P-Y	✓		-		✓	
3´P-Y	-		✓		-	
Strand breaks	1		1		2	
Substrates	ssDNA	ssDNA ssRNA	dsDNA		dsDNA crossovers	dsDNA
DNA relaxation	Underwound DNA		Underwound DNA Overwound DNA		Underwound DNA Overwound DNA	
ΔLk	+1		+/- 1		+/- 2	
DNA strand passage	Enzyme-bridging		DNA rotation Replication swivel		One intact DNA double helix through another	
Co-factors	Mg2+		None		Mg2+, ATP	
Partners	RMI1	TDRD3	POL2	?	?	
Biochemical Activities	Similar		Different		Similar	
Role	- Maintaining genomic stability - Recombinational repair and chromosome segregation		- Remove supercoiling stress - Elongation step in DNA synthesis	mtDNA stability	- Catenation and decatenation of dsDNA rings - Indispensable in chromosome condensation and segregation	
***In vivo* functions**	Nonoverlapping and distinct		Nonoverlapping and distinct		Overlapping	
Core Region	Conserved		Conserved		Conserved	
C-terminal region	Different		Similar		Different	
						Truncated variant
N-terminal region	More divergent		More divergent		More divergent	
	NLS+MTS	NLS	NLS	MTS	NLS (+MTS?)	NLS+MTS
Type of protein	Monomeric		Monomeric		Homodimers	
Homology	Top IA and Top IB are non-homologous, sharing neither sequence nor structure similarities				Top IIA and Top IIB share a homologous ATP-binding domain ('Bergerat fold')	

1.4. Type I Topoisomerases

Type I topoisomerases are monomeric enzymes that do not require ATP for strand catalysation, being classified into Type IA and Type IB enzymes (Table 1.2). They differ in their substrate preference for single-stranded DNA versus double-stranded DNA and the type of covalent connection formed with the DNA strand (Champoux 2001). These two families present no sequence nor structural similarity, indicating independent origins (Forterre et al. 2007). Type IA enzymes have a preference for single-stranded DNA, focusing to deal with underwound DNA, which accumulates negative supercoils and has single-strand character (Kim and Wang 1992). Type IB enzymes only cleave one strand in the double-stranded DNA (Been and Champoux 1984) and can resolve both positive (overwound DNA) and negative supercoils (underwound DNA).

The activities of these enzymes are based on a series of events that alter the substrate's topological state by changing the linking number (the number of times the two strands of

a duplex are wrapped around each other) in steps of one (Champoux 2002; Dekker et al. 2002). Type IA enzymes make a connection to the 5' end of the DNA strand, whereas in Type IB the tyrosine residue links to the 3' end. Following DNA cleavage, the enzyme alters the design of the DNA by passing another strand through the split DNA (enzyme bridging), in the case of Type IA, or by enabling the cut strand to rotate around the uncut strand (enzyme swivelase), in the case of Type IB (Capranico et al. 2017). The protein's brief covalent bond with the broken DNA avoids the unintentional release of broken DNA strands, which may otherwise harm the genome.

1.4.1. Type IA topoisomerases

Type IB and Type II topoisomerases have a well-characterized relaxation and decatenation functions that appear to be sufficient to address the topological restrictions associated with DNA metabolism in eukaryotes (Austin et al. 2018; Baxter and Diffley 2008; Bermejo et al. 2007; Brill et al. 1987; Lee and Berger 2019; Pommier et al. 2022; Pommier et al. 2016). Type IA topoisomerases are conserved in practically all organisms (Bugreev and Nevinsky 2009; Garnier et al. 2018; Viard and de la Tour 2007), being involved in a variety of cellular activities that necessitate the manipulation of particular topological configurations that Type IB and Type II topoisomerases are unable to resolve.

Once connected to the DNA, they only cut one strand (defined as the G-segment or Gated). Both DNA extremities are connected to the protein throughout the cleavage process, where the 5´end is covalently attached to the catalytic tyrosine, and the other chain is non-covalently but strongly bound to the protein (Viard et al. 2004; Zhang et al. 1996). The second strand (defined as T-segment or Transported) is then passed through the gap, using the torsional stress stored in the DNA molecule as energy (Viard and de la Tour 2007). This reaction mechanism, in which the DNA ends that are created in the DNA breakage reaction are bridged by the topoisomerase, is called ‘enzyme-bridging’.

A similar architecture is shared by all Type IA topoisomerases (Figure 1.1), encompassing one topoisomerase-primase subdomain (TOPRIM; subdomain I) and two catabolite activator protein subdomains (CAP; subdomains III and IV), which are connected by two topo-folds (subdomain II) (Bocquet et al. 2014; Capranico et al. 2017; Goto-Ito et al. 2017; Hansen et al. 2006; Lima et al. 1994; Mondragón and DiGate 1999; Rodríguez and Stock 2002). Subdomains I, III, and IV are connected at the base of an arc formed by subdomain II, resembling a toroidal clamp. Upon association or dissociation of a gate created at the interface between subdomain III and subdomains I and IV, it can adopt either a closed or open conformation. The transesterification reaction

requires ten very conserved residues distributed throughout subdomains I, III, and IV, in addition to the catalytic tyrosine, which is found in subdomain III (Garnier et al. 2018).

In spite of having the same structure and catalytic mechanism, Type IA topoisomerases are known to catalyse distinct events *in vitro*, implying potentially different *in vivo* activities (Viard and de la Tour 2007). DNA topoisomerases seem to be specialized for specific topological alterations, with some appearing to be better in relaxing supercoiled DNA *in vitro* (topoisomerases I or Relaxases), whereas others appear to be better at catenation/decatenation (topoisomerases III or Decatenases) (Viard and de la Tour 2007). However, the distinction between these two functions can sometimes be difficult to define as some enzymes are efficient in both activities (Viard et al. 2001).

Type IA enzymes are categorized into five subfamilies: (1) bacterial topoisomerase I, likely involved in the control of DNA supercoiling (Drlica 1992); (2) bacterial topoisomerase III, presumed to play a prominent role in episome segregation (DiGate and Marians 1988); (3) reverse gyrase, present in thermophilic archaea and bacteria (Bouthier De La Tour et al. 1991); (4) archaeal topoisomerase III (Slesarev et al. 1991), and (5) eukaryal topoisomerase III or TOP3 (Forterre et al. 2007).

Higher eukaryotic genomes have two genes for Type IA topoisomerases (Table 1.1): topoisomerase IIIα (*TOP3A*) and topoisomerase IIIβ (*TOP3B*) (Figure 1.1). They have similar enzymatic activity but different *in vivo* roles, suggesting they are non-redundant topoisomerases (Plank et al. 2005; Wilson et al. 2000). TOP3A and TOP3B catalyse a unique set of topological changes as they only relax hyper-negative supercoiling, forming cleavage complexes in single-stranded nucleic acids (Pommier et al. 2022). According to phylogenetic studies, these two topoisomerases arose from an early duplication event in the eukaryotic lineage (Forterre et al. 2007).

TOP3s are more related to chromosomal integrity maintenance than to global or local control of the genome´s topological status, being known for their capacity to work in tandem with DNA translocases to catalyse intricate topological changes that would be impossible for a single topoisomerase to achieve (Viard and de la Tour 2007). TOP3A and TOP3B enzymes are dependent on protein scaffolding cofactors (Figure 1.1 and Table 1.2), namely RecQ-mediated genome instability protein 1 (RMI1) and RMI2 for TOP3A, and Tudor domain-containing protein 3 (TDRD3) for TOP3B (Ahmad et al. 2017; Bizard and Hickson 2020; Yang et al. 2014).

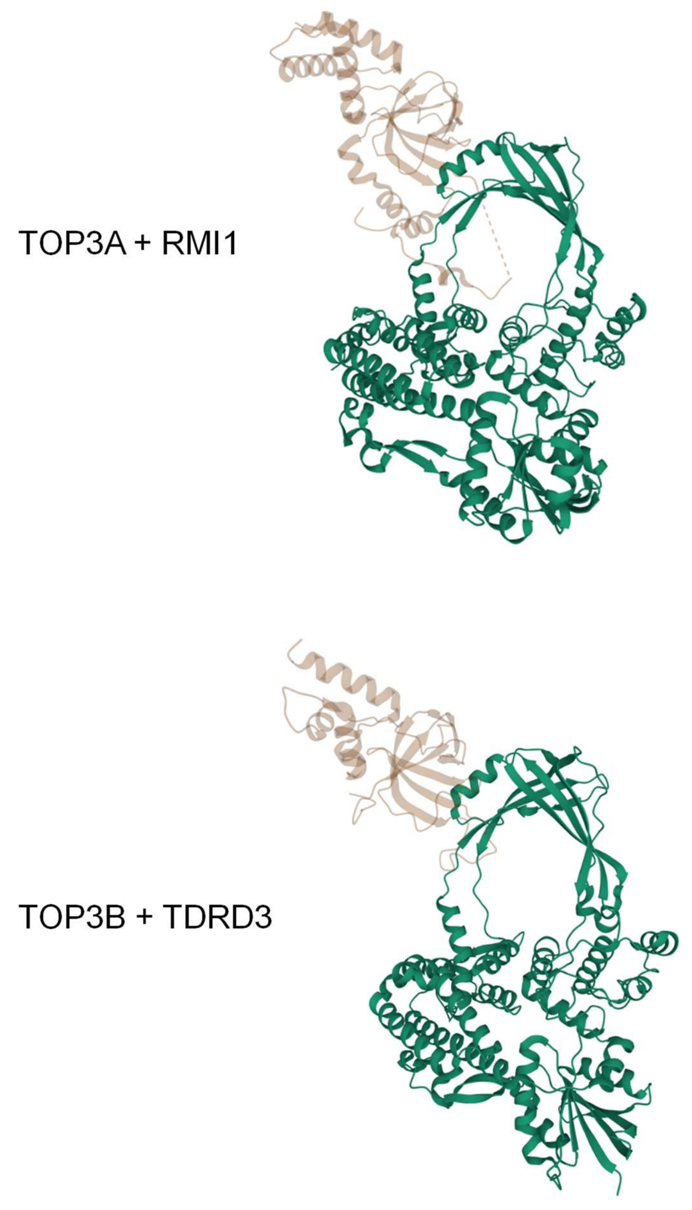


Figure 1.1. Cartoon representation of human Type IA topoisomerases. The TOP3A (PDB 4CGY) is in complex with RMI1 (light brown) and TOP3B (PDB 5GVE) is in complex with TDRD3 (light brown). Images from RCSB PDB (https://www.rcsb.org/).

1.4.1.1. Topoisomerase IIIα (TOP3A)

Inactivation of *TOP3A* in *Mus musculus*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Arabidopsis thaliana* causes embryonic death, indicating that *TOP3A* has an essential function (Kim et al. 2000; Li and Wang 1998; Plank et al. 2005). Mutations in *TOP3A* have been reported in individuals with combined Bloom and mitochondrial syndromes characterized by dilated cardiomyopathy, mitochondrial (mtDNA) depletion in muscles and progressive external ophthalmoplegia syndrome (Jiang et al. 2021; Martin et al. 2018). Humans that exhibit a reduced level of TOP3A activity are characterized by short stature and microcephaly (Martin et al. 2018).

In higher eukaryotes, the *TOP3A* gene encodes both a nuclear and a mitochondrial topoisomerase, with the latter being expressed from an alternative start codon that creates a mitochondrial targeting sequence at the enzyme´s N terminal (Nicholls et al. 2018; Wang et al. 2002). This mitochondrial version of TOP3A is essential for mtDNA replication and segregation (Jiang et al. 2021; Nicholls et al. 2018).

TOP3A is the only topoisomerase capable of eliminating hemicatenanes from converging replication forks (Lee et al. 2019) and recombination intermediates (Bizard and Hickson 2020; Pommier et al. 2016), making it a critical enzyme for the process of replication. TOP3A associates with RecQ-mediated genome instability protein 1 (RMI1) and RMI2 to form the heterotrimeric dissolvasome complex (Bloom syndrome protein (BLM)-TOP3A-RMI1/2 = BTR complex) for homologous recombination (Wright et al. 2018) and resolution of double Holliday junctions associated with DNA replication (Bizard and Hickson 2020). It can also associate its activity with FANCM (Fanconi anaemia group M helicase) at halted replication forks, to prevent sister chromatid swaps and enhance replication restart (Hemphill et al. 2009; Hoadley et al. 2012), and PICH (Plk1-interacting check-point helicase) during mitosis anaphase (Bizard et al. 2019).

1.4.1.2. Topoisomerase IIIβ (TOP3B)

Among the human topoisomerases, TOP3B is unique by having a dual role as DNA and RNA topoisomerase (Ahmad et al. 2017; Ahmad et al. 2016). TOP3B is presumably not essential for life, however, mice lacking TOP3B have a shortened lifespan, a higher incidence of aneuploidy in germ cells, increased autoimmunity (Kwan et al. 2007; Kwan et al. 2003; Kwan and Wang 2001), abnormal synapse formation (Xu et al. 2013) and behavioural impairments (Joo et al. 2020).

The preferred cellular substrates for TOP3B are RNA knots and catenanes (Ahmad et al. 2016). The inclusion of an RNA binding domain, RGG-box, in TOP3B but not in TOP3A is largely responsible for the difference in RNA topoisomerase activity, as ablation of this domain reduced TOP3B´s RNA topoisomerase activity (Xu et al. 2013). Indeed, it was discovered that the RGG-box is conserved in TOP3B from animals, plants, and fungi (Ahmad et al. 2016). TOP3B forms a complex with TDRD3 (Stoll et al. 2013; Xu et al. 2013) which interacts with a known mRNA-binding protein (RBP), FMRP (Fragile-X mental retardation protein) (Linder et al. 2008). In addition to its role in TOP3B recruitment, TDRD3 has also been shown to increase the proficiency of TOP3B (Siaw et al. 2016; Yang et al. 2014).

In eukaryotes with nuclear and cytoplasmic compartments, specialization of topoisomerase activity for each nucleic acid may be advantageous because it prevents DNA topoisomerases from binding to mRNAs and becoming mis-localized in the cytoplasm, where mRNA translation occurs (Ahmad et al. 2017).

1.4.2. Type IB topoisomerases

Type IB topoisomerases share no sequence or structural similarities with Type IA, and promote changes in DNA topology through a very different method (Figure 1.2 and Table 1.2). They act by a DNA rotation mechanism, rather than by enzyme-bridging (Champoux 2001). When they transiently cleave one strand of the dsDNA, only the side containing the protein-linked 3′ end of the broken strand is securely attached to the enzyme, facilitating the broken strand to rotate around the topoisomerase-DNA bond, dissipating torsional stress to a fully relaxed product, only using the free energy stored in DNA supercoils (Forterre et al. 2007; Pommier et al. 2016). This rotation/swivelling movement resets the DNA to its most stable topological shape, resulting in very efficient release of both negative and positive supercoils (Kim and Jinks-Robertson 2017).

Type IB topoisomerases are ubiquitous in eukaryotes and are also present in certain viruses and bacteria (Krogh and Shuman 2002). These topoisomerases share structural and functional properties with the tyrosine recombinases that include the bacteriophage P1 Cre, and *E. coli* XerD recombinases, and certain phage integrases (Sherratt and Wigley 1998). The structural similarities between the two classes of enzymes include the lack of Mg2+ dependence and the formation of a covalent intermediate involving attachment of the enzyme to the 3´phosphate of the cleaved strand (Table 1.2).

Type IB topoisomerases are well fitted to relieve torsional stress, having been linked to fundamental DNA-dependent processes like replication, transcription, DNA repair, chromatin alterations and nucleosome assembly (Capranico et al. 2007). There is a high degree of homology between TOP1 and TOP1MT (Figure 1.2), at both the amino acid and nucleotide levels (Zhang et al. 2001). The exon structure of the two genes is especially conserved, namely the last 13 exons, suggesting that the *TOP1MT* gene is not derived from any of the prokaryotic *top1* genes (Zhang et al. 2001).

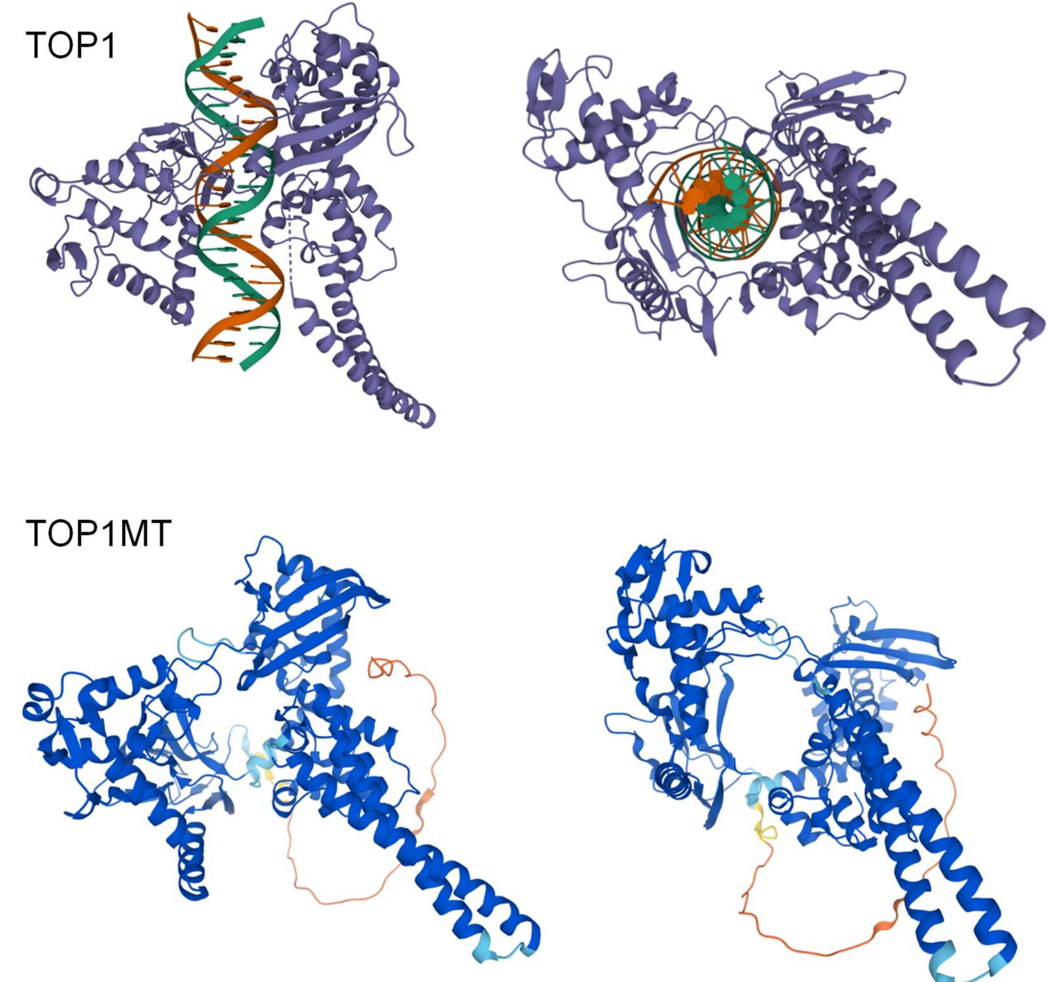


Figure 1.2. Cartoon representation of human Type IB topoisomerases. The TOP1 (PDB 1A36) is in complex with DNA (image from RCSB PDB). The TOP1MT (AF-Q969P6-F1) is coloured according to the confidence of prediction (image from AlphaFold).

1.4.2.1. TOP1

The human topoisomerase I enzyme gene (*TOP1*) is located on chromosome 20, and consists of 21 exons encoding a 765-amino-acid protein (Table 1.1). The domain organization of this gene has been established by crystallographic investigations (Redinbo et al. 1998; Stewart et al. 1998).

The enzyme is composed by four main domains: N-terminal (NTD), Core domain (CAP and CAT), Linker, and C-terminal (CTD) (Redinbo et al. 2000; Redinbo et al. 1998). The eukaryotic N-terminal is not required for relaxation activity *in vitro* and constitute a hydrophilic, unstructured, and highly protease-sensitive region of the protein (Stewart et al. 1996a). The N-terminal domain includes four nuclear localization signals (NLS) and one acidic NLS (Stewart et al. 1996b). This domain is where interactions with other cellular proteins occur, including nucleolins, SV40 T-antigen, certain transcription factors, p53, and the WRN protein (Albor et al. 1998; Bharti et al. 1996; Simmons et al. 1996). The N-terminal domain is followed by a highly conserved Core domain that is involved in the catalytic process, being important for the preferential binding of the enzyme to the supercoiled DNA (Madden et al. 1995). The Core domain contains four of the five catalytic residues (Arg488, Lys532, Arg590, and His632) (Redinbo et al. 2000; Redinbo et al. 1998). This domain is further subdivided into three subdomains based on its structure, namely subdomains I and II, that form a “CAP” region containing a pair of α-helices called the “nose cone”; and subdomain III that forms the “CAT” region (Schoeffler and Berger 2008). At the top of subdomain III there is also a region called “Hinge”, and opposite to that there are two loops (called “lips”) that interact with each other to close the enzyme around the DNA (Redinbo et al. 1998). A protease-sensitive and weakly conserved Linker domain, in effect dispensable for its *in vitro* activity (Stewart et al. 1997), connects the Core domain to the C-terminal domain. Even though the Linker domain is not directly involved in the enzyme catalysis, it plays an important role in the process of controlled rotation of the cleavage complex (D'Annessa et al. 2014). The active site tyrosine (Tyr723) is located in the C-terminal domain (Redinbo et al. 2000; Redinbo et al. 1998).

1.4.2.2. TOP1MT

TOP1MT, a type IB enzyme encoded in the nuclear genome but acting on mtDNA, is the most recent human topoisomerase to be found, localized on human chromosome 8q24 (Zhang et al. 2001). It binds to the region flanking the end of the replication D-loop in the mtDNA control region, at the putative attachment site of nucleoids to the mitochondrial

inner membrane (Zhang and Pommier 2008). In contrast to TOP1, which contains several NLSs, this mitochondrial enzyme lacks an amino acid sequence corresponding to an NLS. Instead, it revealed the presence of positively charged residues that fold in a positively charged amphiphilic helix suggestive of a mitochondrial targeting signal (MTS) (Zhang et al. 2001). The *TOP1MT* presumably arose by duplication and modification of an early nuclear *TOP1* gene. The size and sequence of the introns in nuclear and mitochondrial genes are vastly different, indicating that this duplication event is ancient (Zhang et al. 2001). TOP1MT seems to be dispensable in mice, despite its conservation in all vertebrates, indicating that other topoisomerases can complement it (Zhang et al. 2007).

1.5. Type II Topoisomerases

All Type II topoisomerases use an ATP-coupled DNA transport reaction, in which the enzyme cleaves a "gate" DNA (G-segment), transports a second duplex (T-segment) through the break and subsequently reconnects the cleaved DNA. During the cleavage the enzyme generates a pair of phosphotyrosine links to each of the 5´ends of the DNA (Table 1.2). The transport activity allows this type of topoisomerases to catenate or decatenate duplex DNA rings and to alter the superhelical density of the DNA substrate (Berger 1998; Wang 1996). This topological process changes the Lk by ±2 (Brown and Cozzarelli 1979). If the G and T-segments come from the same molecule, supercoiling is altered; if they come from distinct molecules, topoisomerase action causes catenation or decatenation of DNAs (Sutormin et al. 2021).

Based on sequence and structural similarities, Type II topoisomerases are classified into two families, IIA and IIB (Forterre and Gadelle 2009; Forterre et al. 2007; Gadelle et al. 2003). Type IIA topoisomerases are widespread and can be found in all domains of life, as well as in some viruses, being further subdivided into three homologous subfamilies - eukaryotic top II, bacterial top IV, and bacterial and archaeal gyrase - that exhibit distinct functional properties (Forterre et al. 2007; Gadelle et al. 2003; Schoeffler and Berger 2008). The taxonomic distribution of Type IIB topoisomerases is more restricted than that of Type IIA. Type IIB family includes the archaeal Topoisomerase VI and the homologs in plants, red and green algae, some protists and a few bacteria (Forterre and Gadelle 2009; Forterre et al. 2007; Gadelle et al. 2003). More recently, a new Type IIB topoisomerase, Top VIII, has been identified exclusively on the corresponding genes to mobile genetic elements (MGE) being distributed to nine different bacteria phyla and one archaeon superphylum (Gadelle et al. 2014; Takahashi et al. 2020).

Several studies on the yeast Type II DNA topoisomerases (Goto et al. 1984; Goto and Wang 1982) have demonstrated that the yeast enzymes are very similar to other eukaryotic Type II topoisomerases, in terms of its strict ATP dependence in catalysing the transient breakage and passage of duplex DNA and its inability to catalyse the supercoiling of DNA, thus, the structural insights from the yeast enzyme are also likely to apply to the human enzyme, making it a suitable system for studying the biological functions of the DNA topoisomerases in eukaryotic organisms.

1.5.1. Type IIA

Type IIA topoisomerases are always twofold symmetric (Figure 1.3). The active Type IIA holoenzyme can take the form of a homodimer, in the case of eukaryotic and viral top II; a heterotetramer, in bacterial and archaeal gyrase and bacterial top IV; or a heterohexamer in bacteriophage T4 top II (Schoeffler and Berger 2008). The quaternary structure of Type IIA enzymes results in the formation of three major gates (N-terminal gate, DNA gate and C-gate), which encompass two separate cavities within the enzyme (Schoeffler and Berger 2008), and although structural studies of Type IIA topoisomerases have revealed the existence of more than two physical gates, the enzymes are nonetheless believed to operate by the general principles of the two-gate model (Roca et al. 1996; Roca and Wang 1994).

The N-gate includes the conserved ATP-hydrolysis GHKL (Gyrase, Hsp90, Histidine Kinase, MutL) domain (Dutta and Inouye 2000). The DNA gate includes the TOPRIM and WHD (Winged-helix domain) domains (Broeck et al. 2019). The G-segment of DNA binds to the DNA-gate region of the enzyme and is cleaved by active site tyrosyl residues of the WHD domain (Schmidt et al. 2010). The third dimerization interface, the C-gate, is formed by the coiled-coil (CC) domain, which is only present in type IIA enzymes (Berger et al. 1996).

In the reaction cycle, one double-stranded DNA (termed the G segment) is bound and cleaved by the enzyme, while a second duplex (the 'T segment') is transported through the break (Roca and Wang 1994). G-segment breakage is catalysed by a pair of symmetrically related tyrosines (Tse et al. 1980; Worland and Wang 1989), in conjunction with a Mg2+ ion-binding topoisomerase-primase (TOPRIM) fold (Aravind et al. 1998), to form a transient covalent topoisomerase–DNA cleavage complex. Strand passage is coordinated by the ATPase domains (Brown et al. 1979; Gellert et al. 1979; Lindsley and Wang 1993), which use ATP binding and hydrolysis to promote T-segment capture, stimulate G-segment cleavage and coordinate successive opening and closing

of the gates (Roca and Wang 1992; Roca and Wang 1994; Williams and Maxwell 1999a; Williams and Maxwell 1999b).

1.5.1.1 TOP2A and TOP2B

Eukaryotic topoisomerase IIA (TOP2) is common to all known eukaryotes. It is encoded by one *TOP2* gene in most species; vertebrates, however, exhibit two paralogues, *Top2α* and *Top2β* (Chung et al. 1989), designated by *TOP2A* and *TOP2B* (Figure 1.3). Human topoisomerase IIα (TOP2A) is encoded by the *TOP2A* gene on chromosome 17q21-22 (Tsai-Pflugfelder et al. 1988). TOP2A is found in proliferating cell types, and expression peaks during the G2 and M phases of the cell cycle (Turley et al. 1997). Human topoisomerase IIβ (TOP2B) is encoded by the *TOP2B* on chromosome 3p24 (Jenkins et al. 1992). TOP2B is found in all cell types, and its expression is constant throughout the cell cycle (Turley et al. 1997).

These paralogues appear to be expressed at different times in the cell cycle and in different cell types (Capranico et al. 1992; Gonzalez et al. 2011). The TOP2 CTD is the least conserved region. The CTD undergoes post-translational modifications, most prominently, phosphorylation, which changes in a cell cycle-dependent manner. The divergence between TOP2A and TOP2B CTDs determines the functional differences between the paralogs and their regulation (Gilroy and Austin 2011; Meczes et al. 2008).

The differences between the C-terminal regions of TOP2A and TOP2B suggests that these regions may mediate different cellular functions and making them more prone to different DNA topologies. In fact, the C terminus region of TOP2A moves the activity of the enzyme towards the preferential relaxation of positive supercoils, whereas the equivalent region of TOP2B does not appear to impart any supercoil preference (McClendon et al. 2005). These differences may be linked to particular cellular functions and biological roles (Linka et al. 2007), as TOP2A is essential in proliferating cells and assists in chromosome condensation, segregation and replication (Grue et al. 1998; Ye et al. 2010), whereas TOP2B is associated with DNA repair, transcription and development (Bollimpelli et al. 2017; Ju et al. 2006; McNamara et al. 2008; Yang et al. 2000).

With the exception for regions encoding the extreme N-terminal and C-terminal domains of the protein, there is a high degree of sequence organization and conservation between TOP2A and TOP2B. Also, the amino-acid sequence of human TOP2A is more similar to TOP2A of other vertebrate species than it is to human TOP2B and vice versa. The two paralogue genes derived from the duplication of an ancestral gene, duplication event

that in turn occurred prior to the speciation event that gave rise to the different Chordates (Lang et al. 1998).

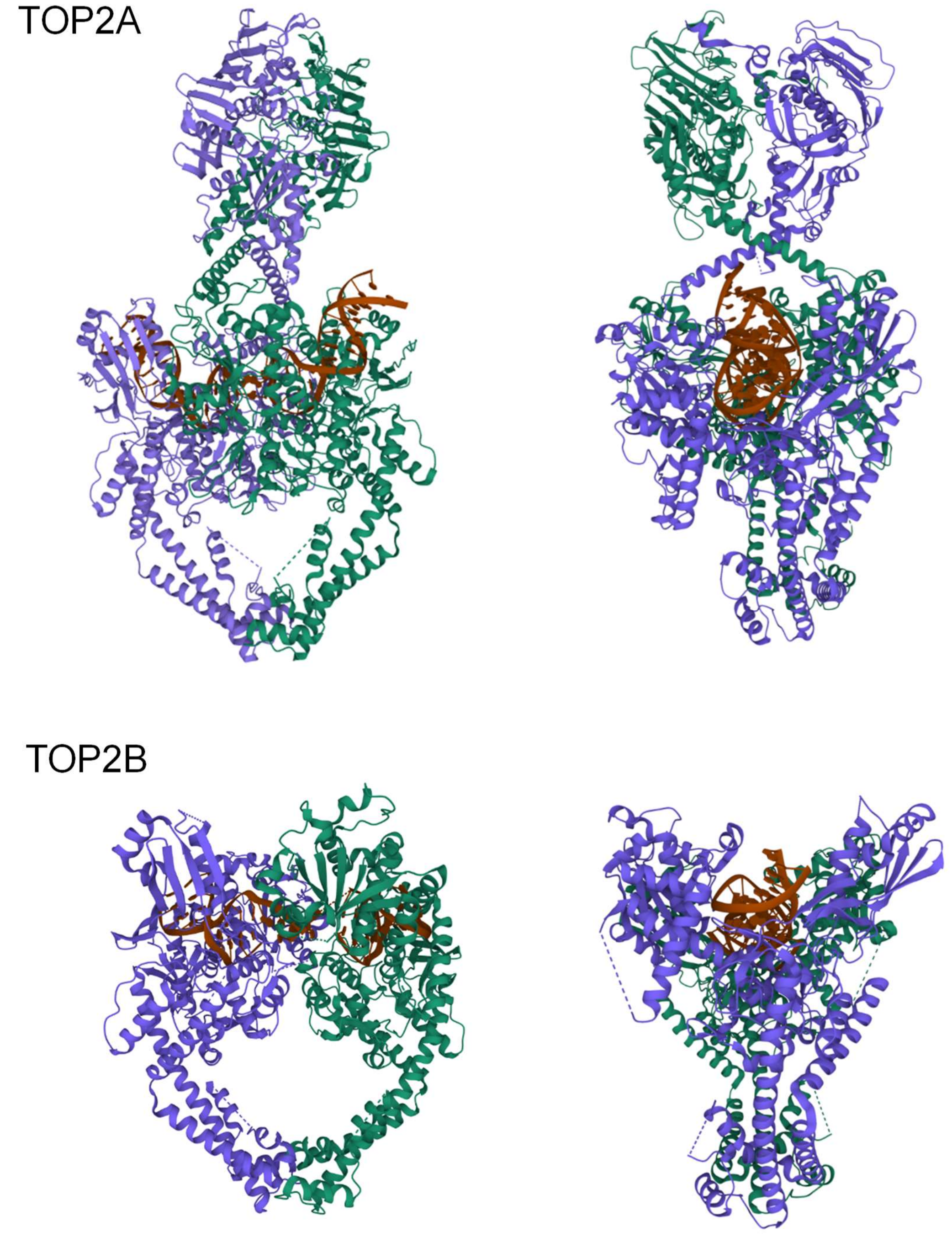


Figure 1.3. Cartoon representation of human Type IIA topoisomerases. The TOP2A (PDB 6ZY7) and TOP2B (PDB 4J3N) are in complex with DNA (images from RCSB PDB). The dimers are coloured differently.

1.6. Resistance to topoisomerase-targeted drugs

Topoisomerases are an important target of anticancer and antibacterial therapeutics (Pommier et al. 2010). If properly induced by drugs, irreversible topoisomerase cleavage complexes (TOPccs) can be used in cancer treatment. The irreversible TOPccs lead to the accumulation of double stranded DNA breaks, causing cells to undergo apoptosis, being beneficial when occurring in cancer cells, which often overexpress topoisomerase genes and are highly dependent on topoisomerase activities. The wide spectrum of antimicrobial and anticancer agents that target topoisomerases either act to stabilize TOPccs (defined as 'poisons') or inhibit enzyme catalysis to induce DNA damage (Cuya et al. 2017; Pommier 2013). Inhibition of the DNA topoisomerases occur essentially by the formation of the topoisomerase poisons, a ternary protein-DNA-drug complex that prevents DNA re-ligation and locks the enzyme into a 'cleavage complex'. Type IA topoisomerases (TOP3A and TOP3B) are not yet clinical therapeutic targets, but Type IB topoisomerases and Type IIA topoisomerases are important targets for anticancer and antibacterial drugs (Bailly 2000; Pommier 2006).

Type IB topoisomerases (TOP1 and TOP1MT) are targeted by camptothecins and noncamptothecins (indenoisoquinolines and ARC-111). The targeting of TOP1 for the stabilization of TOP1ccs by camptothecins during replication is an effective strategy for treating different tumours (Pommier 2006). The TOP1ccs cause DNA strand breaks, disruption of DNA uncoiling and unstable RNA transcripts, being highly detrimental to cells. The TOP1 inhibitor camptothecin is no longer used due to its side effects, being replaced by derived forms such as topotecan and irinotecan. TOP1MT is sensitive to camptothecin, however, the drug is unlikely to target the enzyme in mitochondria since the mitochondrial matrix is alkaline, which renders camptothecin inactive (Zhang et al. 2001).

Type IIA topoisomerases (TOP2A and TOP2B) are targets for etoposide, anthracyclines (doxorubicin, daunorubicin), and mitoxantrone. The drugs targeting these topoisomerases have been divided into two types. The most relevant type includes etoposide, doxorubicin and mitoxantrone (TOP2 poisons), leading to the formation of TOP2-DNA covalent complexes. The other type of compounds inhibits the TOP2 catalytic activity, but do not increase the amounts of TOP2-DNA covalent complexes.

A significant obstacle to the efficacy of chemotherapy is the occurrence of topoisomerase mutations that confer resistance of the cancer cells to the treatment, resulting in refractory cases (Chrencik et al. 2004; Cretaio et al. 2007; Saleem et al. 2000). Cancer cells are positively selected if they have a mutation that will block the action of the drug,

increasing the replicative advantage of the cell and derived clones. The structures of drug-resistant forms of topoisomerases have been used to better understand the mechanisms of drug action (Chrencik et al. 2004; Staker et al. 2005; Staker et al. 2002).

1.7. The study of topoisomerase genes throughout the animal kingdom

The chronological order of early animal diversification events is a matter of debate for a very long time. The evolutionary relationships of the five major metazoan lineages (Porifera, Placozoa, Ctenophora, Cnidaria and Bilateria) are controversial when only considering morphological characters (Collins et al. 2005). The use of molecular technologies has altered our knowledge of metazoan phylogeny, however, data inaccuracy and scarcity from many major taxa hinders the process (Dunn et al. 2008; Roure et al. 2013). Phylogenomic techniques demonstrate that early animal evolution resulted in considerably more flexibility in phenotypic evolution than initially expected (Adoutte et al. 2000; Giribet 2016; Wanninger 2016). These techniques employ data from many genes to resolve deep animal connections, allowing the placement of taxa whose morphology and embryology have proven inconclusive or deceptive.

According to molecular and morphological data, bilaterian animals are divided in three main clades - Deuterostomia, Lophotrochozoa, and Ecdysozoa (Nosenko et al. 2013). Deuterostomia is formed by the phyla Hemichordata, Echinodermata and Chordata. Lophotrochozoa has the greatest diversity of body plans of the three bilaterian supergroups. It is a well-supported clade of invertebrates, grouping animals with a lophophore feeding mechanism, trochophore larvae, as well as several other recognized phyla: Annelida, Brachiopoda, Bryozoa, Cycliophora, Dicyemida, Entoprocta, Gastrotricha, Gnathostomulida, Micrognathozoa, Mollusca, Nemertea, Orthonectida, Phoronida, Platyhelminthes and Rotifera (Kocot et al. 2017). Ecdysozoa includes moulting animals (Arthropoda, Nematoda, Tardigrada, Priapulida). Sponges (Porifera), ctenophorans (Ctenophora), cnidarians (Cnidaria), and placozoans (Placozoa) diverged from the main animal lineage before bilateral animals appeared, although the order of divergence is still subject of debate (Halanych 2004).

Despite the relevance of topoisomerases for basic biological processes, these enzymes are poorly studied across the animal kingdom. For almost all animal phyla, there is no work that addresses the genetics of topoisomerases or their molecular features. Some authors used topoisomerase genes as a target for real-time PCR assay for species detection [e.g., pine wood nematode, (Huang et al. 2010)], taking advantage of the high conservation of these proteins. But almost all published works were done in model

organisms. Within protostomes, nematodes and arthropods stood out mainly because of their model organisms. The topoisomerases from *C. elegans* are relatively well characterized (Jaramillo-Lambert et al. 2016; Kim et al. 1996; Kim et al. 2000; Lee et al. 1998; Lee et al. 2001). In arthropods, the Topoisomerase I of *Spodoptera exigua* (Lepidoptera: Noctuidae) has been studied as a potential target of pesticides for insect control (Zhang et al. 2017). Similarly, topoisomerase II has been studied as a target in vector control of the yellow fever mosquito, *Aedes aegypti* (Diptera: Culicidae) (Santos et al. 2021). Obviously, topoisomerases are better characterized in the fruit fly, *D. melanogaster*. Several studies have determined the cellular roles of topoisomerases [e.g., (Lee et al. 1993; Lee et al. 2018; Osheroff et al. 1983; Sander and Hsieh 1983; Yang et al. 2021)] and genetic features [e.g., (Nolan et al. 1986; Wyckoff et al. 1989)] of topoisomerases in this model species.

Concerning deuterostomes, studies on the biological roles of topoisomerases in echinoderms are rare (Poccia et al. 1978). Similarly, no relevant work on the genetics of topoisomerase has been conducted in Cephalochordata and Tunicata (Urochordata). The Cyclostomata (jawless fishes) group, including the lampreys and hagfishes, are particularly important for the evolutionary history of topoisomerases. It has been shown that Type IIA and Type IB topoisomerases have paralogues in vertebrates (Forterre and Gadelle 2009; Forterre et al. 2007). Unfortunately, the molecular features of Cyclostomata topoisomerases remains largely unknown, despite the ongoing or completed genome sequencing projects in species of this group (Mehta et al. 2013; Smith et al. 2013; Yamaguchi et al. 2020).

Vertebrates are obviously the animal group where topoisomerases have been better studied. With exception of the zebrafish, fishes are rarely studied, and without going into details on the molecular evolution of these enzymes (Patel et al. 2019; Postlethwait et al. 2003). Regarding zebrafish, several studies investigate the functions of topoisomerases, most related with embryonic development and the nervous system [e.g., (Doolittle 2017; Dovey et al. 2009; Sapetto-Rebow et al. 2011; Zaksauskaite et al. 2021)]. Although not so recent, a few works have been done in *Xenopus laevis* [e.g., (Benedetti et al. 1983; Pandit et al. 1996; Richard and Bogenhagen 1991)]. Aves or reptiles have been rarely used as models to study these proteins (Niimi et al. 2001; Petruti-Mot and Earnshaw 2000). When not considering humans, studies on mammals are focused on the model organisms, mainly *Mus musculus* [e.g., (Baechler et al. 2019; Fragola et al. 2020; Khiati et al. 2015; Morham et al. 1996)], which contributed significantly for our knowledge on topoisomerases.

1.8. References

Adoutte A, Balavoine G, Lartillot N, Lespinet O, Prud'homme B, De Rosa R (2000) The new animal phylogeny: reliability and implications. Proceedings of the National Academy of Sciences 97:4453-4456.

Ahmad M, Xu D, Wang W (2017) Type IA topoisomerases can be "magicians" for both DNA and RNA in all domains of life. RNA biology 14:854-864.

Ahmad M et al. (2016) RNA topoisomerase is prevalent in all domains of life and associates with polyribosomes in animals. Nucleic acids research 44:6335-6349.

Albor A, Kaku S, Kulesz-Martin M (1998) Wild-type and mutant forms of p53 activate human topoisomerase I: a possible mechanism for gain of function in mutants. Cancer research 58:2091-2094.

Amores A et al. (1998) Zebrafish hox clusters and vertebrate genome evolution. Science 282:1711-1714.

Aravind L, Leipe DD, Koonin EV (1998) Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Nucleic acids research 26:4205-4213.

Austin CA et al. (2018) TOP2B: the first thirty years. International Journal of Molecular Sciences 19:2765.

Baechler S et al. (2019) The mitochondrial type IB topoisomerase drives mitochondrial translation and carcinogenesis. Nature communications 10:1-13.

Bailly C (2000) Topoisomerase I poisons and suppressors as anticancer drugs. Current medicinal chemistry 7:39-58.

Baxter J, Diffley JF (2008) Topoisomerase II inactivation prevents the completion of DNA replication in budding yeast. Molecular cell 30:790-802.

Been MD, Champoux JJ (1984) Breakage of single-stranded DNA by eukaryotic type 1 topoisomerase occurs only at regions with the potential for base-pairing. Journal of molecular biology 180:515-531.

Benedetti P, Baldi MI, Mattoccia E, Tocchini-Valentini GP (1983) Purification and characterization of Xenopus laevis topoisomerase II. The EMBO Journal 2:1303-1308.

Berger JM (1998) Type II DNA topoisomerases. Current opinion in structural biology 8:26-32.

Berger JM, Gamblin SJ, Harrison SC, Wang JC (1996) Structure and mechanism of DNA topoisomerase II. Nature 379:225-232.

Bermejo R, Doksani Y, Capra T, Katou Y-M, Tanaka H, Shirahige K, Foiani M (2007) Top1-and Top2-mediated topological transitions at replication forks ensure fork progression and stability and prevent DNA damage checkpoint activation. Genes & development 21:1921-1936.

Bharti AK, Olson MO, Kufe DW, Rubin EH (1996) Identification of a Nucleolin Binding Site in Human Topoisomerase I. Journal of Biological Chemistry 271:1993-1997.

Bizard AH, Allemand J-F, Hassenkam T, Paramasivam M, Sarlos K, Singh MI, Hickson ID (2019) PICH and TOP3A cooperate to induce positive DNA supercoiling. Nature Structural & Molecular Biology 26:267-274.

Bizard AH, Hickson ID (2020) The many lives of type IA topoisomerases. Journal of Biological Chemistry 295:7138-7153.

Bocquet N et al. (2014) Structural and mechanistic insight into Holliday-junction dissolution by topoisomerase IIIα and RMI1. Nature structural & molecular biology 21:261-268.

Bollimpelli VS, Dholaniya PS, Kondapi AK (2017) Topoisomerase IIβ and its role in different biological contexts. Archives of Biochemistry and Biophysics 633:78-84.

Bouthier De La Tour C, Portemer C, Huber R, Forterre P, Duguet M (1991) Reverse gyrase in thermophilic eubacteria. Journal of bacteriology 173:3921-3923.

Brill SJ, DiNardo S, Voelkel-Meiman K, Sternglanz R (1987) Need for DNA topoisomerase activity as a swivel for DNA replication for transcription of ribosomal RNA. Nature 326:414-416.

Broeck AV, Lotz C, Ortiz J, Lamour V (2019) Cryo-EM structure of the complete E. coli DNA gyrase nucleoprotein complex. Nature communications 10:1-12.

Brown PO, Cozzarelli NR (1979) A sign inversion mechanism for enzymatic supercoiling of DNA. Science 206:1081-1083.

Brown PO, Peebles CL, Cozzarelli NR (1979) A topoisomerase from Escherichia coli related to DNA gyrase. Proceedings of the National Academy of Sciences 76:6110-6114.

Bugreev D, Nevinsky G (2009) Structure and mechanism of action of type IA DNA topoisomerases. Biochemistry (Moscow) 74:1467-1481.

Capranico G, Ferri F, Fogli MV, Russo A, Lotito L, Baranello L (2007) The effects of camptothecin on RNA polymerase II transcription: roles of DNA topoisomerase I. Biochimie 89:482-489.
Capranico G, Marinello J, Chillemi G (2017) Type i DNA topoisomerases. Journal of medicinal chemistry 60:2169-2192.
Capranico G, Tinelli S, Austin CA, Fisher ML, Zunino F (1992) Different patterns of gene expression of topoisomerase II isoforms in differentiated tissues during murine development. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1132:43-48.
Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. Annual review of biochemistry 70:369-413.
Champoux JJ (2002) Type IA DNA topoisomerases: strictly one step at a time. Proceedings of the National Academy of Sciences 99:11998-12000.
Chrencik JE, Staker BL, Burgin AB, Pourquier P, Pommier Y, Stewart L, Redinbo MR (2004) Mechanisms of camptothecin resistance by human topoisomerase I mutations. Journal of molecular biology 339:773-784.
Chung T, Drake FH, Tan K, Per SR, Crooke ST, Mirabelli CK (1989) Characterization and immunological identification of cDNA clones encoding two human DNA topoisomerase II isozymes. Proceedings of the National Academy of Sciences 86:9431-9435.
Collins AG, Cartwright P, McFadden CS, Schierwater B (2005) Phylogenetic context and basal metazoan model systems. Integrative and Comparative Biology 45:585-594.
Cretaio E, Pattarello L, Fontebasso Y, Benedetti P, Losasso C (2007) Human DNA topoisomerase IB: structure and functions. The Italian journal of biochemistry 56:91-102.
Cuya SM, Bjornsti M-A, van Waardenburg RC (2017) DNA topoisomerase-targeting chemotherapeutics: what's new? Cancer chemotherapy and pharmacology 80:1-14.
D'Annessa I, Coletta A, Sutthibutpong T, Mitchell J, Chillemi G, Harris S, Desideri A (2014) Simulations of DNA topoisomerase 1B bound to supercoiled DNA reveal changes in the flexibility pattern of the enzyme and a secondary protein–DNA binding site. Nucleic acids research 42:9304-9312.
Da Cunha V, Gaia M, Gadelle D, Nasir A, Forterre P (2017) Lokiarchaea are close relatives of Euryarchaeota, not bridging the gap between prokaryotes and eukaryotes. PLoS genetics 13:e1006810.
Dekker N, Rybenkov V, Duguet M, Crisona N, Cozzarelli N, Bensimon D, Croquette V (2002) The mechanism of type IA topoisomerases. Proceedings of the National Academy of Sciences 99:12126-12131.
DiGate RJ, Marians KJ (1988) Identification of a potent decatenating enzyme from Escherichia coli. Journal of Biological Chemistry 263:13366-13373.
Dixon JR et al. (2012) Topological domains in mammalian genomes identified by analysis of chromatin interactions. Nature 485:376-380.
Doolittle S (2017) The Influence of Autism Linked Gene Topoisomerase 3B (Top3B) on Neural Development in Zebrafish.
Dovey M, Patton EE, Bowman T, North T, Goessling W, Zhou Y, Zon LI (2009) Topoisomerase IIα is required for embryonic development and liver regeneration in zebrafish. Molecular and cellular biology 29:3746-3753.
Drlica K (1992) Control of bacterial DNA supercoiling. Molecular microbiology 6:425-433.
Dunn CW et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-749.
Dutta R, Inouye M (2000) GHKL, an emergent ATPase/kinase superfamily. Trends in biochemical sciences 25:24-28.
Forterre P (2006) Three RNA cells for ribosomal lineages and three DNA viruses to replicate their genomes: a hypothesis for the origin of cellular domain. Proceedings of the National Academy of Sciences 103:3669-3674.
Forterre P, Gadelle D (2009) Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. Nucleic Acids Research 37:679-692.
Forterre P, Gribaldo S, Gadelle D, Serre M-C (2007) Origin and evolution of DNA topoisomerases. Biochimie 89:427-446.
Fragola G et al. (2020) Deletion of Topoisomerase 1 in excitatory neurons causes genomic instability and early onset neurodegeneration. Nature communications 11:1-13.
Gadelle D, Filée J, Buhler C, Forterre P (2003) Phylogenomics of type II DNA topoisomerases. Bioessays 25:232-242.

Gadelle D, Krupovic M, Raymann K, Mayer C, Forterre P (2014) DNA topoisomerase VIII: a novel subfamily of type IIB topoisomerases encoded by free or integrated plasmids in Archaea and Bacteria. Nucleic acids research 42:8578-8591.
Garnier F, Debat H, Nadal M (2018) Type IA DNA topoisomerases: A universal core and multiple activities. In: DNA Topoisomerases. Springer, pp 1-20
Gellert M, Fisher LM, O'Dea MH (1979) DNA gyrase: purification and catalytic properties of a fragment of gyrase B protein. Proceedings of the National Academy of Sciences 76:6289-6293.
Gilroy KL, Austin CA (2011) The impact of the C-Terminal domain on the interaction of human DNA topoisomerase II α and β with DNA. PLoS One 6:e14693.
Giribet G (2016) New animal phylogeny: future challenges for animal phylogeny in the age of phylogenomics. Organisms Diversity & Evolution 16:419-426.
Gonzalez RE et al. (2011) Effects of conditional depletion of topoisomerase II on cell cycle progression in mammalian cells. Cell Cycle 10:3505-3514.
Goto-Ito S, Yamagata A, Takahashi TS, Sato Y, Fukai S (2017) Structural basis of the interaction between Topoisomerase IIIβ and the TDRD3 auxiliary factor. Scientific reports 7:1-10.
Goto T, Laipis P, Wang JC (1984) The purification and characterization of DNA topoisomerases I and II of the yeast Saccharomyces cerevisiae. Journal of Biological Chemistry 259:10422-10429.
Goto T, Wang JC (1982) Yeast DNA topoisomerase II. An ATP-dependent type II topoisomerase that catalyzes the catenation, decatenation, unknotting, and relaxation of double-stranded DNA rings. Journal of Biological Chemistry 257:5866-5872.
Grue P et al. (1998) Essential mitotic functions of DNA topoisomerase IIα are not adopted by topoisomerase IIβ in human H69 cells. Journal of Biological Chemistry 273:33660-33666.
Halanych KM (2004) The new view of animal phylogeny. Annu. Rev. Ecol. Evol. Syst. 35:229-256.
Hansen G, Harrenga A, Wieland B, Schomburg D, Reinemer P (2006) Crystal structure of full length topoisomerase I from Thermotoga maritima. Journal of molecular biology 358:1328-1340.
Hemphill A et al. (2009) Topo IIIα and BLM act within the Fanconi anemia pathway in response to DNA-crosslinking agents. Cytogenetic and genome research 125:165-175.
Hoadley KA, Xue Y, Ling C, Takata M, Wang W, Keck JL (2012) Defining the molecular interface that connects the Fanconi anemia protein FANCM to the Bloom syndrome dissolvasome. Proceedings of the National Academy of Sciences 109:4437-4442.
Huang L, Ye J-r, Wu X-q, Xu X-l, Sheng J-m, Zhou Q-x (2010) Detection of the pine wood nematode using a real-time PCR assay to target the DNA topoisomerase I gene. European Journal of Plant Pathology 127:89-98.
Jaramillo-Lambert A, Fabritius AS, Hansen TJ, Smith HE, Golden A (2016) The identification of a novel mutant allele of topoisomerase II in Caenorhabditis elegans reveals a unique role in chromosome segregation during spermatogenesis. Genetics 204:1407-1422.
Jenkins JR et al. (1992) Isolation of cDNA clones encoding the β isozyme of human DNA topoisomerase II and localisation of the gene to chromosome 3p24. Nucleic acids research 20:5587-5592.
Jiang W, Jia N, Guo C, Wen J, Wu L, Ogi T, Zhang H (2021) Predominant cellular mitochondrial dysfunction in the TOP3A gene-caused Bloom syndrome-like disorder. Biochimica et Biophysica Acta (BBA)-Molecular Basis of Disease 1867:166106.
Joo Y et al. (2020) Topoisomerase 3β knockout mice show transcriptional and behavioural impairments associated with neurogenesis and synaptic plasticity. Nature communications 11:1-18.
Ju B-G, Lunyak VV, Perissi V, Garcia-Bassets I, Rose DW, Glass CK, Rosenfeld MG (2006) A topoisomerase IIß-mediated dsDNA break required for regulated transcription. science 312:1798-1802.
Khiati S et al. (2015) Lack of mitochondrial topoisomerase I (TOP1mt) impairs liver regeneration. Proceedings of the National Academy of Sciences 112:11282-11287.
Kim J, Kim YC, Lee JH, Jang YJ, Chung IK, Koo HS (1996) cDNA cloning, expression, and chromosomal localization of Caenorhabditis elegans DNA topoisomerase I. European journal of biochemistry 237:367-372.
Kim N, Jinks-Robertson S (2017) The Top1 paradox: Friend and foe of the eukaryotic genome. DNA repair 56:33-41.

Kim RA, Wang JC (1992) Identification of the yeast TOP3 gene product as a single strand-specific DNA topoisomerase. Journal of Biological Chemistry 267:17178-17185.
Kim Y-C, Lee J, Koo H-S (2000) Functional characterization of Caenorhabditis elegans DNA topoisomerase IIIα. Nucleic acids research 28:2012-2017.
Kocot KM et al. (2017) Phylogenomics of Lophotrochozoa with consideration of systematic error. Systematic biology 66:256-282.
Krogh BO, Shuman S (2002) A poxvirus-like type IB topoisomerase family in bacteria. Proceedings of the National Academy of Sciences 99:1853-1858.
Kwan KY, Greenwald RJ, Mohanty S, Sharpe AH, Shaw AC, Wang JC (2007) Development of autoimmunity in mice lacking DNA topoisomerase 3β. Proceedings of the National Academy of Sciences 104:9242-9247.
Kwan KY, Moens PB, Wang JC (2003) Infertility and aneuploidy in mice lacking a type IA DNA topoisomerase IIIβ. Proceedings of the National Academy of Sciences 100:2526-2531.
Kwan KY, Wang JC (2001) Mice lacking DNA topoisomerase IIIβ develop to maturity but show a reduced mean lifespan. Proceedings of the National Academy of Sciences 98:5717-5721.
Lang A, Mirski S, Cummings H, Yu Q, Gerlach J, Cole S (1998) Structural organization of the human TOP2A and TOP2B genes. Gene 221:255-266.
Lee CM, Wang G, Pertsinidis A, Marians KJ (2019) Topoisomerase III acts at the replication fork to remove precatenanes. Journal of bacteriology 201:e00563-00518.
Lee JH, Berger JM (2019) Cell cycle-dependent control and roles of DNA topoisomerase II. Genes 10:859.
Lee MH, Jang YJ, Koo H-S (1998) Alternative splicing in the Caenorhabditis elegans DNA topoisomerase I gene. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1396:207-214.
Lee MH, Park H, Shim G, Lee J, Koo HS (2001) Regulation of gene expression, cellular localization, and in vivo function of Caenorhabditis elegans DNA topoisomerase I. Genes to Cells 6:303-312.
Lee MP, Brown SD, Chen A, Hsieh T-s (1993) DNA topoisomerase I is essential in Drosophila melanogaster. Proceedings of the National Academy of Sciences 90:6656-6660.
Lee SK et al. (2018) Topoisomerase 3β interacts with RNAi machinery to promote heterochromatin formation and transcriptional silencing in Drosophila. Nature communications 9:1-17.
Leipe DD, Aravind L, Koonin EV (1999) Did DNA replication evolve twice independently? Nucleic acids research 27:3389-3401.
Li W, Wang JC (1998) Mammalian DNA topoisomerase IIIα is essential in early embryogenesis. Proceedings of the National Academy of Sciences 95:1010-1013.
Lima CD, Wang JC, Mondragón A (1994) Three-dimensional structure of the 67K N-terminal fragment of E. coli DNA topoisomerase I. Nature 367:138-146.
Linder B et al. (2008) Tdrd3 is a novel stress granule-associated protein interacting with the Fragile-X syndrome protein FMRP. Human molecular genetics 17:3236-3246.
Lindsley JE, Wang JC (1993) Study of allosteric communication between protomers by immunotagging. Nature 361:749-750.
Linka RM, Porter AC, Volkov A, Mielke C, Boege F, Christensen MO (2007) C-terminal regions of topoisomerase II α and II β determine isoform-specific functioning of the enzymes in vivo. Nucleic acids research 35:3810-3822.
Liu LF, Wang JC (1987) Supercoiling of the DNA template during transcription. Proceedings of the National Academy of Sciences 84:7024-7027.
Madden K, Stewart L, Champoux J (1995) Preferential binding of human topoisomerase I to superhelical DNA. The EMBO journal 14:5399-5409.
Martin C-A et al. (2018) Mutations in TOP3A cause a Bloom syndrome-like disorder. The American Journal of Human Genetics 103:221-231.
McClendon AK, Rodriguez AC, Osheroff N (2005) Human topoisomerase IIα rapidly relaxes positively supercoiled DNA: implications for enzyme action ahead of replication forks. Journal of Biological Chemistry 280:39337-39345.
McNamara S, Wang H, Hanna N, Miller Jr WH (2008) Topoisomerase IIβ negatively modulates retinoic acid receptor α function: a novel mechanism of retinoic acid resistance. Molecular and cellular biology 28:2066-2077.
Meczes EL, Gilroy KL, West KL, Austin CA (2008) The impact of the human DNA topoisomerase II C-terminal domain on activity. PloS one 3:e1754.

Mehta TK et al. (2013) Evidence for at least six Hox clusters in the Japanese lamprey (Lethenteron japonicum). Proceedings of the National Academy of Sciences 110:16044-16049.
Meyer A, Schartl M (1999) Gene and genome duplications in vertebrates: the one-to-four (-to-eight in fish) rule and the evolution of novel gene functions. Current opinion in cell biology 11:699-704.
Meyer A, Van de Peer Y (2005) From 2R to 3R: evidence for a fish-specific genome duplication (FSGD). Bioessays 27:937-945.
Mondragón A, DiGate R (1999) The structure of Escherichia coli DNA topoisomerase III. Structure 7:1373-1383.
Morham SG, Kluckman KD, Voulomanos N, Smithies O (1996) Targeted disruption of the mouse topoisomerase I gene by camptothecin selection. Molecular and cellular biology 16:6804-6809.
Nicholls TJ et al. (2018) Topoisomerase 3α is required for decatenation and segregation of human mtDNA. Molecular cell 69:9-23. e26.
Niimi A, Suka N, Harata M, Kikuchi A, Mizuno S (2001) Co-localization of chicken DNA topoisomerase IIα, but not β, with sites of DNA replication and possible involvement of a C-terminal region of α through its binding to PCNA. Chromosoma 110:102-114.
Nitiss JL (2009) Targeting DNA topoisomerase II in cancer chemotherapy. Nature Reviews Cancer 9:338-350.
Nolan JM, Lee MP, Wyckoff E, Hsieh T-S (1986) Isolation and characterization of the gene encoding Drosophila DNA topoisomerase II. Proceedings of the National Academy of Sciences 83:3664-3668.
Nosenko T et al. (2013) Deep metazoan phylogeny: when different genes tell different stories. Molecular phylogenetics and evolution 67:223-233.
Ohno S (2013) Evolution by gene duplication. Springer Science & Business Media,
Osheroff N, Shelton E, Brutlag D (1983) DNA topoisomerase II from Drosophila melanogaster. Relaxation of supercoiled DNA. Journal of Biological Chemistry 258:9536-9543.
Pandit SD, Richard RE, Sternglanz R, Bogenhagen DF (1996) Cloning and characterization of the gene for the somatic form of DNA topoisomerase I from Xenopus laevis. Nucleic acids research 24:3593-3600.
Patel DM, Bhide K, Bhide M, Iversen MH, Brinchmann MF (2019) Proteomic and structural differences in lumpfish skin among the dorsal, caudal and ventral regions. Scientific reports 9:1-13.
Petruti-Mot AS, Earnshaw WC (2000) Two differentially spliced forms of topoisomerase IIα and β mRNAs are conserved between birds and humans. Gene 258:183-192.
Plank JL, Chu SH, Pohlhaus JR, Wilson-Sali T, Hsieh T-s (2005) Drosophila melanogaster topoisomerase IIIα preferentially relaxes a positively or negatively supercoiled bubble substrate and is essential during development. Journal of Biological Chemistry 280:3564-3573.
Poccia D, LeVine D, Wang J (1978) Activity of a DNA topoisomerase (nicking-closing enzyme) during sea urchin development and the cell cycle. Developmental Biology 64:273-283.
Pommier Y (2006) Topoisomerase I inhibitors: camptothecins and beyond. Nature Reviews Cancer 6:789-802.
Pommier Y (2013) Drugging topoisomerases: lessons and challenges. ACS chemical biology 8:82-95.
Pommier Y, Leo E, Zhang H, Marchand C (2010) DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. Chemistry & biology 17:421-433.
Pommier Y, Nussenzweig A, Takeda S, Austin C (2022) Human topoisomerases and their roles in genome stability and organization. Nature Reviews Molecular Cell Biology:1-21.
Pommier Y, Sun Y, Shar-yin NH, Nitiss JL (2016) Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nature reviews Molecular cell biology 17:703-721.
Poole AM, Logan DT (2005) Modern mRNA proofreading and repair: clues that the last universal common ancestor possessed an RNA genome? Molecular biology and evolution 22:1444-1455.
Postlethwait JH, Amores A, Yan Y-L, Austin C (2003) Duplication of a portion of human chromosome 20q containing Topoisomerase (Top1) and Snail genes provides evidence on genome expansion and the radiation of teleost fish. In: Aquatic genomics. Springer, pp 20-34

Redinbo MR, Champoux JJ, Hol WG (2000) Novel insights into catalytic mechanism from a crystal structure of human topoisomerase I in complex with DNA. Biochemistry 39:6832-6840.
Redinbo MR, Stewart L, Kuhn P, Champoux JJ, Hol WG (1998) Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. Science 279:1504-1513.
Richard RE, Bogenhagen DF (1991) The 165-kDa DNA topoisomerase I from Xenopus laevis oocytes is a tissue-specific variant. Developmental biology 146:4-11.
Roca J, Berger JM, Harrison SC, Wang JC (1996) DNA transport by a type II topoisomerase: direct evidence for a two-gate mechanism. Proceedings of the National Academy of Sciences 93:4057-4062.
Roca J, Wang JC (1992) The capture of a DNA double helix by an ATP-dependent protein clamp: a key step in DNA transport by type II DNA topoisomerases. Cell 71:833-840.
Roca J, Wang JC (1994) DNA transport by a type II DNA topoisomerase: evidence in favor of a two-gate mechanism. Cell 77:609-616.
Rodríguez AC, Stock D (2002) Crystal structure of reverse gyrase: insights into the positive supercoiling of DNA. The EMBO journal 21:418-426.
Rosa ID et al. (2009) Adaptation of topoisomerase I paralogs to nuclear and mitochondrial DNA. Nucleic acids research 37:6414-6428.
Roure B, Baurain D, Philippe H (2013) Impact of missing data on phylogenies inferred from empirical phylogenomic data sets. Molecular biology and evolution 30:197-214.
Saleem A, Edwards TK, Rasheed Z, Rubin EH (2000) Mechanisms of resistance to camptothecins. Annals of the New York Academy of Sciences 922:46-55.
Sander M, Hsieh T-S (1983) Double strand DNA cleavage by type II DNA topoisomerase from Drosophila melanogaster. Journal of Biological Chemistry 258:8421-8428.
Santos D et al. (2021) Inhibition of Aedes aegypti DNA topoisomerase II by etoposide: Impact on survival and morphology of larvae and pupae. Comparative Biochemistry and Physiology Part C: Toxicology & Pharmacology 247:109066.
Sapetto-Rebow B et al. (2011) Maternal topoisomerase II alpha, not topoisomerase II beta, enables embryonic development of zebrafish top2a-/-mutants. BMC developmental biology 11:1-18.
Schmidt BH, Burgin AB, Deweese JE, Osheroff N, Berger JM (2010) A novel and unified two-metal mechanism for DNA cleavage by type II and IA topoisomerases. Nature 465:641-644.
Schoeffler AJ, Berger JM (2008) DNA topoisomerases: harnessing and constraining energy to govern chromosome topology. Quarterly reviews of biophysics 41:41-101.
Sherratt DJ, Wigley DB (1998) Conserved themes but novel activities in recombinases and topoisomerases. Cell 93:149-152.
Siaw GE-L, Liu I-F, Lin P-Y, Been MD, Hsieh T-s (2016) DNA and RNA topoisomerase activities of Top3β are promoted by mediator protein Tudor domain-containing protein 3. Proceedings of the National Academy of Sciences 113:E5544-E5551.
Siggers T, Gordân R (2014) Protein–DNA binding: complexities and multi-protein codes. Nucleic acids research 42:2099-2111.
Simmons DT, Melendy T, Usher D, Stillman B (1996) Simian virus 40 large T antigen binds to topoisomerase I. Virology 222:365-374.
Slesarev A, Zaitzev D, Kopylov V, Stetter K, Kozyavkin S (1991) DNA topoisomerase III from extremely thermophilic archaebacteria. ATP-independent type I topoisomerase from Desulfurococcus amylolyticus drives extensive unwinding of closed circular DNA at high temperature. Journal of Biological Chemistry 266:12321-12328.
Smith JJ et al. (2013) Sequencing of the sea lamprey (Petromyzon marinus) genome provides insights into vertebrate evolution. Nature genetics 45:415-421.
Staker BL, Feese MD, Cushman M, Pommier Y, Zembower D, Stewart L, Burgin AB (2005) Structures of three classes of anticancer agents bound to the human topoisomerase I–DNA covalent complex. Journal of medicinal chemistry 48:2336-2345.
Staker BL, Hjerrild K, Feese MD, Behnke CA, Burgin AB, Stewart L (2002) The mechanism of topoisomerase I poisoning by a camptothecin analog. Proceedings of the National Academy of Sciences 99:15387-15392.
Stewart L, Ireton GC, Champoux JJ (1996a) The Domain Organization of Human Topoisomerase I. Journal of Biological Chemistry 271:7602-7608.
Stewart L, Ireton GC, Champoux JJ (1997) Reconstitution of human topoisomerase I by fragment complementation. Journal of molecular biology 269:355-372.

Stewart L, Ireton GC, Parker LH, Madden KR, Champoux JJ (1996b) Biochemical and Biophysical Analyses of Recombinant Forms of Human Topoisomerase I. Journal of Biological Chemistry 271:7593-7601.
Stewart L, Redinbo MR, Qiu X, Hol WG, Champoux JJ (1998) A model for the mechanism of human topoisomerase I. Science 279:1534-1541.
Stoll G et al. (2013) Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. Nature neuroscience 16:1228-1237.
Sutormin D, Galivondzhyan A, Polkhovskiy A, Kamalyan S, Severinov K, Dubiley S (2021) Diversity and Functions of Type II Topoisomerases. Acta Naturae 13:59.
Takahashi DT et al. (2022) Topoisomerase I (TOP1) dynamics: conformational transition from open to closed states. Nature communications 13:1-11.
Takahashi TS, Da Cunha V, Krupovic M, Mayer C, Forterre P, Gadelle D (2020) Expanding the type IIB DNA topoisomerase family: identification of new topoisomerase and topoisomerase-like proteins in mobile genetic elements. NAR genomics and bioinformatics 2:lqz021.
Thomas A, Pommier Y (2019) Targeting topoisomerase I in the era of precision medicine. Clinical Cancer Research 25:6581-6589.
Tsai-Pflugfelder M et al. (1988) Cloning and sequencing of cDNA encoding human DNA topoisomerase II and localization of the gene to chromosome region 17q21-22. Proceedings of the National Academy of Sciences 85:7177-7181.
Tse Y, Kirkegaard K, Wang J (1980) Covalent bonds between protein and DNA. Formation of phosphotyrosine linkage between certain DNA topoisomerases and DNA. Journal of Biological Chemistry 255:5560-5565.
Turley H et al. (1997) The distribution and expression of the two isoforms of DNA topoisomerase II in normal and neoplastic human tissues. British journal of cancer 75:1340-1346.
Vann KR, Oviatt AA, Osheroff N (2021) Topoisomerase II poisons: converting essential enzymes into molecular scissors. Biochemistry 60:1630-1641.
Viard T, Cossard R, Duguet M, de La Tour CB (2004) Thermotoga maritima-Escherichia coli chimeric topoisomerases: Answers about involvement of the carboxyl-terminal domain in DNA topoisomerase I-mediated catalysis. Journal of Biological Chemistry 279:30073-30080.
Viard T, de la Tour CB (2007) Type IA topoisomerases: a simple puzzle? Biochimie 89:456-467.
Viard T, Lamour V, Duguet M, de la Tour CB (2001) Hyperthermophilic topoisomerase I from Thermotoga maritima: A very efficient enzyme that functions independently of zinc binding. Journal of Biological Chemistry 276:46495-46503.
Vinograd J, Lebowitz J (1966) Physical and topological properties of circular DNA. The Journal of General Physiology 49:103-125.
Vos SM, Tretter EM, Schmidt BH, Berger JM (2011) All tangled up: how cells direct, manage and exploit topoisomerase function. Nature reviews Molecular cell biology 12:827-841.
Wang JC (1971) Interaction between DNA and an Escherichia coli protein ω. Journal of molecular biology 55:523-IN516.
Wang JC (1996) DNA topoisomerases. Annual review of biochemistry 65:635-692.
Wang JC (2002) Cellular roles of DNA topoisomerases: a molecular perspective. Nature reviews Molecular cell biology 3:430-440.
Wang Y, Lyu YL, Wang JC (2002) Dual localization of human DNA topoisomerase IIIα to mitochondria and nucleus. Proceedings of the National Academy of Sciences 99:12114-12119.
Wanninger A (2016) Twenty years into the “new animal phylogeny”: Changes and challenges vol 16. Springer,
Whyte WA et al. (2013) Master transcription factors and mediator establish super-enhancers at key cell identity genes. Cell 153:307-319.
Williams NL, Maxwell A (1999a) Locking the DNA gate of DNA gyrase: investigating the effects on DNA cleavage and ATP hydrolysis. Biochemistry 38:14157-14164.
Williams NL, Maxwell A (1999b) Probing the two-gate mechanism of DNA gyrase using cysteine cross-linking. Biochemistry 38:13502-13511.
Wilson TM, Chen AD, Hsieh T-s (2000) Cloning and Characterization of DrosophilaTopoisomerase IIIβ: relaxation of hypernegatively supercoiled DNA. Journal of Biological Chemistry 275:1533-1540.

Worland S, Wang J (1989) Inducible overexpression, purification, and active site mapping of DNA topoisomerase II from the yeast Saccharomyces cerevisiae. Journal of Biological Chemistry 264:4412-4416.
Wright WD, Shah SS, Heyer W-D (2018) Homologous recombination and the repair of DNA double-strand breaks. Journal of Biological Chemistry 293:10524-10535.
Wyckoff E, Natalie D, Nolan JM, Lee M, Hsieh T-s (1989) Structure of the Drosophila DNA topoisomerase II gene: nucleotide sequence and homology among topoisomerases II. Journal of molecular biology 205:1-13.
Xu D et al. (2013) Top3β is an RNA topoisomerase that works with fragile X syndrome protein to promote synapse formation. Nature neuroscience 16:1238-1247.
Yamaguchi K, Hara Y, Tatsumi K, Nishimura O, Smith JJ, Kadota M, Kuraku S (2020) Inference of a genome-wide protein-coding gene set of the inshore hagfish Eptatretus burgeri. bioRxiv.
Yang D-W, Mok J-W, Telerman SB, Amson R, Telerman A, Choi K-W (2021) Topoisomerase II is regulated by translationally controlled tumor protein for cell survival during organ growth in Drosophila. Cell death & disease 12:1-10.
Yang X, Li W, Prescott ED, Burden SJ, Wang JC (2000) DNA topoisomerase IIβ and neural development. Science 287:131-134.
Yang Y, McBride KM, Hensley S, Lu Y, Chedin F, Bedford MT (2014) Arginine methylation facilitates the recruitment of TOP3B to chromatin to prevent R loop accumulation. Molecular cell 53:484-497.
Ye J et al. (2010) TRF2 and Apollo cooperate with topoisomerase 2α to protect human telomeres from replicative damage. Cell 142:230-242.
Zaksauskaite R, Thomas RC, van Eeden F, El-Khamisy SF (2021) Tdp1 protects from topoisomerase 1–mediated chromosomal breaks in adult zebrafish but is dispensable during larval development. Science Advances 7:eabc4165.
Zhang H, Barceló JM, Lee B, Kohlhagen G, Zimonjic DB, Popescu NC, Pommier Y (2001) Human mitochondrial topoisomerase I. Proceedings of the National Academy of Sciences 98:10608-10613.
Zhang H, Meng L-H, Pommier Y (2007) Mitochondrial topoisomerases and alternative splicing of the human TOP1mt gene. Biochimie 89:474-481.
Zhang H, Pommier Y (2008) Mitochondrial topoisomerase I sites in the regulatory D-loop region of mitochondrial DNA. Biochemistry 47:11196-11203.
Zhang HL, Malpure S, Li Z, Hiasa H, DiGate RJ (1996) The Role of the Carboxyl-terminal Amino Acid Residues in Escherichia coli DNA Topoisomerase III-mediated Catalysis. Journal of Biological Chemistry 271:9039-9045.
Zhang P, Zhang L, Zhang Y, Mao L, Jiang H (2017) Substitutions in Spodoptera exigua topoisomerase I modulate its relaxation activity and camptothecin sensitivity. Pest management science 73:1179-1186.

Chapter 2
Objectives

2. Objectives

DNA topoisomerases are of vital importance for life and their meticulous activities are a testament to the power of evolution. Most works have focused on the biochemical and structural features of topoisomerases and how they mediate DNA replication, transcription, chromatin dynamics, DNA repair and genomic stability. The role of topoisomerases in neurodegenerative diseases, immune disorders and as target of anticancer treatments has also been extensively studied. However, our current knowledge on topoisomerases is restricted to the information gained by studying a few model organisms. Most works fail to incorporate an evolutionary perspective in the interpretation of their experiments and findings.

The main objective of this PhD dissertation is to study the molecular evolution of DNA topoisomerases in animals. We aim to provide new insights into the origin and evolution of topoisomerase in Metazoa, taking advantage of the increasing number of sequenced genomes and improved computational and theoretical resources. In particular, this PhD dissertation has the following specific objectives:

1) Reconstruct the phylogeny of topoisomerases in animals, with focus on the origin of vertebrate paralogues;

2) Determine the selective forces governing the evolution of animal topoisomerases, including the detection of positively selected sites;

3) Identify polymorphisms amongst the topoisomerases of archaic and modern humans, including missense mutations of putative biological relevance;

4) Uncover the evolutionary history of topoisomerase domains and motifs in the light of their biological roles and

5) Examine the topoisomerase sites associated with diseases or conferring resistance to anticancer drugs.

Chapter 3
Publication nº1

Journal of Molecular Evolution (2021) 89:384–395
https://doi.org/10.1007/s00239-021-10011-7

ORIGINAL ARTICLE

Molecular Evolution of DNA Topoisomerase III Beta (TOP3B) in Metazoa

Filipa Moreira[1,2] · Miguel Arenas[3,4,5] · Arnaldo Videira[2,6,7] · Filipe Pereira[8,9]

Received: 5 December 2020 / Accepted: 30 April 2021 / Published online: 17 May 2021

3. Molecular evolution of DNA topoisomerase III beta (TOP3B) in Metazoa (Scientific Publication nº 1)

Filipa Moreira[1,2], Miguel Arenas[3,4,5], Arnaldo Videira[2,6,7] and Filipe Pereira[8,9*]

[1] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208, Matosinhos, Portugal
[2] Institute of Biomedical Sciences Abel Salazar (ICBAS), University of Porto, Rua Jorge de Viterbo Ferreira 228, 4050-313, Porto, Portugal
[3] Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain
[4] CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[5] Galicia Sur Health Research Institute (IIS Galicia Sur), 36310 Vigo, Spain
[6] IBMC-Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal
[7] i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
[8] IDENTIFICA genetic testing, Rua Simão Bolívar 259 3º Dir Tras. 4470-214 Maia, Portugal
[9] Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal

* Corresponding author. E-mail address: fpereirapt@gmail.com (F. Pereira)

3.1. Abstract

DNA topoisomerase III beta (TOP3B) is unique by operating on both DNA and RNA substrates to regulate gene expression and genomic stability. Mutations in human *TOP3B* are linked to neurodevelopmental and cognitive disorders, highlighting its relevance for human health. Despite the emerging importance of TOP3B, its precise cellular functions and evolutionary history remain poorly understood. Here we show that TOP3B is conserved across main metazoan groups and evolved under strong purifying selection. Subdomain IV was identified as the most conserved TOP3B region, in agreement with its role in providing the structural foundation of the protein. On the contrary, subdomain II is the less conserved, possibly because it is the most structurally flexible region of all TOP3B regions. Interestingly, TOP3B residue at position 472, previously associated with schizophrenia, is highly variable across animals, suggesting a more specific role in humans and related species. Finally, we show that all TOP3B CXXC zinc finger motifs previously identified at the protein C-terminal region are retained across metazoans. We also found that the two major methylation sites known to regulate TOP3B activity are located in the most conserved region of the C-terminal arginine-glycine-glycine (RGG) box, suggesting that a similar regulatory mechanism may operate throughout animals. Overall, our results provide a better understanding of the evolution and functional roles of TOP3B.

Keywords: Topoisomerase IA; animal phylogeny; protein domains; zinc finger; schizophrenia

3.2. Introduction

Organisms must have a strict control of the degree of DNA intertwining in order to ensure the protection and the functionality of their genetic material (Bizard and Hickson 2020). DNA topoisomerases are the enzymes that regulate and modify the DNA topological state by catalysing different types of interconversions between DNA topological isomers (topoisomers). This phenomenon happens by the transient cleavage of DNA, accompanied by the simultaneous formation of a transient phosphodiester bond between a tyrosine residue in the protein and one of the ends of the broken DNA strand (Bizard and Hickson 2020; Champoux 2001; Wang 1985).

Type I topoisomerases introduce transient single breaks into the DNA and are organized into three different families (TopIA, TopIB, and TopIC) based on their amino acid sequences and reaction mechanisms (Capranico et al. 2017). The TopIA family includes topoisomerases that link to a 5’ phosphate, while TopIB and TopIC family members make a covalent complex with the 3' end of the broken DNA strand (Garnier et al. 2018). TopIA form a vast and well-defined superfamily present in Bacteria, Archaea and Eukarya, constituting the most ubiquitous protein among all the topoisomerases (Champoux 2001). All TopIA share a similar design within their core domain, which is constituted by one Topoisomerase-Primase subdomain (TOPRIM; subdomain I) and two Catabolite Activator Protein subdomains (CAP-Y and CAP; Subdomains III and IV, respectively), being connected by two Topo-folds (Subdomain II) (Capranico et al. 2017; Goto-Ito et al. 2017; Hansen et al. 2006; Mondragón and DiGate 1999; Rodríguez and Stock 2002).

TopIA has five subfamilies, four of them exclusively found in Bacteria and Archaea (Duguet et al. 2006). Topoisomerase III or TOP3 is the only TopIA found in eukaryotes, with most species having two related genes, *TOP3A* and *TOP3B*. Phylogenetic analysis indicates that the presence of these two topoisomerases resulted from an early duplication event in the eukaryotic lineage (Forterre et al. 2007). Both TOP3 are involved in an extensive range of cellular processes that need the manipulation of specific topological structures that cannot be resolved by Type IB and Type II topoisomerases. Their catalytic versatility has been well-characterized *in vitro,* however the knowledge of their physiological roles is still scarce, despite years of intense research (Bizard and Hickson 2020).

TOP3A and TOP3B exhibit different functions *in vivo*, suggesting they are not redundant enzymes, despite being biochemically similar (Bizard and Hickson 2020; Kwan and Wang 2001). TOP3B appears to be the only nuclear eukaryotic topoisomerase that is not essential for life (Kwan and Wang 2001), however it is highly relevant for human

health. In humans, the gene *TOP3B* resides on chromosome 22q11.2, a region that is frequently affected by deletions or duplications, leading to a variety of problems, such as cognitive dysfunctions, congenital heart disease and facial malformation (Ahmad et al. 2017; Kaufman et al. 2016; Pires et al. 2014; Stoll et al. 2013; Tarsitano et al. 2014). Furthermore, mice lacking TOP3B are deficient in fertility and immunity, reinforcing the importance of this protein (Kwan et al. 2007). TOP3B localizes to both the nucleus and cytoplasm where it interacts with its specific auxiliary factor, Tudor domain–containing 3 protein (TDRD3) (Goto-Ito et al. 2017; Yang et al. 2010). In the nucleus, the TOP3B-TDRD3 complex facilitates transcription by unwinding negative supercoiled DNA and resolving R-loops (Yang et al. 2014). In the cytoplasm, TOP3B-TDRD3 interacts with the Fragile-X syndrome protein (FMRP) to regulate mRNA topological stress and translation (Goulet et al. 2008; Linder et al. 2008; Xu et al. 2013).

TOP3B has a large N-terminal catalytic region (TOPO domain), similar to what is observed in all other Type IA topoisomerases (Champoux 2001). The TOPO domain is further divided into four subdomains (I or TOPRIM, II, III and IV) that assume the configuration of a flattened torus. The structure forms a positively charged hole large enough to accommodate double-stranded DNA. Subdomains II and IV result from two separated regions in the protein primary sequence, which come together in the three-dimensional structure. Interestingly, TOP3B possesses a dual activity in both RNA and DNA, whereas its paralog, TOP3A, has activity only in DNA (Xu et al. 2013). The activity in RNA is fundamentally due to a conserved RNA-binding motif, the arginine-glycine-glycine (RGG) box, located in the C-terminal domain, since deletion of this motif disrupts the RNA topoisomerase activity (Huang et al. 2018; Xu et al. 2013).

Here, we provide a deep analysis of the evolutionary and structural features of TOP3B proteins in metazoans. A comparative genomic analysis of TOP3B was used to support prevailing hypothesis on metazoan phylogeny and evolution. We also identified the most conserved TOP3B residues both in human populations and across animals, which may have functional relevance.

3.3. Material and Methods

TOP3B sequences

TOP3B gene sequences were obtained from the Ensembl Genome Server (Hunt et al. 2018) and NCBI Nucleotide database. TOP3B protein sequences were obtained from the Ortho DB v10 (https://www.orthodb.org), a comprehensive database with evolutionary and functional annotations of orthologues (Kriventseva et al. 2019). We searched for TOP3B orthologues in Metazoa using the database default parameters. Some species with available genomes lack a TOP3B orthologue or large sections of the protein. Although it is possible that such absences are real, they most likely result from gaps in the sequencing project. We have also removed duplicated sequences from the same species that showed 100% identity and most likely represent different entries of the same sequence in databases.

Denisova and Neanderthal *TOP3B* sequences were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/) (Kent et al. 2002). All BAM reads for tracks *Denisova* and *Neanderthal Cntgs* matching the Human Mar. 2006 (NCBI36/hg18) chr22:20,641,403-20,667,147 region were downloaded. The BAM reads from each track were then reassembled against the human TOP3B reference sequence (NC_000022.11) using the Geneious v2019.0.4 program (http://www.geneious.com). We only considered a position as polymorphic in archaic hominids when: 1) at least two reads overlap in that position; 2) the polymorphism represents more than 75% of all reads and 3) the polymorphism is not at the end of a read.

Data on human *TOP3B* missense variants were retrieved from the TOP3B (ENSG00000100038) Variation Table available on the Ensembl Genome Server (Hunt et al. 2018). The frequency of reference and alternate alleles for amino acid position 472 were obtained from the Ensembl and the NCBI dbSNP database (Sherry et al. 2001).

Sequence alignments

The multiple sequence alignment of *TOP3B* orthologue genes from Chordates (*n* = 234), Actinopteri (*n* = 77), Reptilia (*n* = 16), Aves (*n* = 45) and Mammalia (*n* = 93) were obtained from the Ensembl Genome Server. The alignments from Arthropoda (*n* = 111) and Nematoda (*n* = 7) were performed in the TranslatorX server (Abascal et al. 2010) using parameters by default.

The protein sequence alignments were performed using the default settings of the MUSCLE 3.8.425 software (Edgar 2004) implemented in the Geneious program. The

Gblocks server was used to remove the poorly aligned positions and divergent regions of the protein alignments using default parameters except for the 'with-half' gaps option (Castresana 2000; Talavera and Castresana 2007). Three main protein sequence alignments were analysed: Metazoa (*n* = 265), Chordata (*n* = 144) and Arthropoda (*n* = 105).

The level of sequence conservation was measured with the Geneious program as a percentage of pairwise identity. This measure gives the average percent identity over part or the complete alignment by looking at all pairs of bases at the same column and scoring a hit (one) when they are identical, divided by the total number of pairs. The sequence alignments are available at Mendeley Data (https://data.mendeley.com/datasets/3t6sp4vw2v/1).

Estimation of selection and substitution saturation

We analysed selection in the multiple sequence alignments of *TOP3B* orthologue genes through the nonsynonymous/synonymous substitution rate ratio (dN/dS), which is a common metric to identify selection in protein-coding sequences (Del Amparo et al. 2021; Jeffares et al. 2015). For every multiple sequence alignment, we identified the best-fitting substitution model of DNA evolution with jModelTest2 (Darriba et al. 2012) under the Bayesian Information Criterion (BIC) following (Luo et al. 2010), as shown in Table 3.1. Next, we reconstructed a maximum likelihood (ML) phylogenetic tree with the framework RAxML-NG (Kozlov et al. 2019) under the previously selected substitution model. Finally, we estimated the dN/dS with the well-established evolutionary framework Hyphy (Kosakovsky Pond and Frost 2005; Kosakovsky Pond et al. 2020) and accounting for the phylogenetic tree. In particular, we applied the SLAC (single-likelihood ancestor counting) method, which provides dN/dS estimation with accuracy similar to that obtained with other likelihood-based methods (Kosakovsky Pond and Frost 2005). We estimated global (entire sequence) genetic signatures of selection but also the presence of positively selected sites (PSSs). Next, we explored the presence of saturation of substitution events in every studied multiple sequence alignment since saturation could affect evolutionary analyses. We applied the well-established test of saturation based on the *index of substitutional saturation* (Iss) (Xia and Lemey 2009; Xia et al. 2003) implemented in DAMBE7 (Xia 2018) to the first and second codon positions and to the third codon position, separately.

Phylogenetic analyses

The phylogenies were built using the protein alignment of TOP3B from all metazoan species (*n* = 265) after removing poorly aligned positions using Gblocks, resulting in an alignment with 706 amino acid positions. The *Arabidopsis thaliana* TOP3B was used as outgroup in all phylogenetic trees. The best-fitting amino acid substitution models of evolution were estimated from the alignments with ProtTest 3.4.2 software with a gamma distribution that consider four rate categories. The LG model was selected as the best-fitting substitution model to build the phylogenetic trees. Bayesian analyses were performed with MrBayes v3.2.6 software (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) running on the CIPRES Science Gateway v3.3 (Miller et al. 2010). The Metropolis-coupled Markov chain Monte Carlo process was set with two runs, each with four independent chains that ran simultaneously during 4,000,000 iterations. The average standard deviation of split frequencies of the final Metazoa tree was 0.018541. A burn-in value of 0.25 was applied. Trees were edited with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree/).

TOP3B structure

We used the Human Topoisomerase IIIb topo domain (PDB ID: 5GVC) structure (Goto-Ito et al. 2017) available at The Protein Data Bank (http://www.rcsb.org/) (Berman et al. 2000) and the SWISS-MODEL workspace (Waterhouse et al. 2018). The structure was coloured according to the subdomains described by Goto-Ito *et al.*: subdomains I (1-168), IV (169-234), II (235-295), III (296-422), II (423-494), IV (495-612). The structure of C-terminal protein (613-862) has not been determined experimentally. The C-terminal region was named here as “ZF and RGG” due to the presence of CXXC zinc finger (ZF) motifs (Bizard and Hickson 2020) and an arginine-glycine-glycine (RGG) box (Bizard and Hickson 2020; Xu et al. 2013). The residues interacting with the insertion loop and core region of TDRD3 were retrieved from Goto-Ito *et al.* (2017) and the R472Q mutation found in schizophrenia patients from Xu *et al*. (2012). The amino acids positions are those of the reference TOP3B sequence (GenBank accession number EAW59487.1; Ensemble ENST00000357179.10).

3.4. Results and Discussion

TOP3B are widespread and its phylogeny supports prevailing hypotheses on animal evolution

We were able to recover TOP3B protein sequences for most metazoans with available complete genome sequences, supporting their widespread distribution (Figure 3.1; Supplementary Figures S3.1 to S3.5). The TOP3B phylogenetic analysis placed Porifera, Placozoa and Cnidaria at the base of the tree, supporting their ancient origin within Metazoa (Collins et al. 2005). The data agrees with the placement of placozoans before the separation of cnidarians and bilaterians, but after the divergence of Porifera from other animals, with Bayesian posterior probabilities (PP, in percentage) of 100. This topology is in line with the phylogeny generated using a concatenation of 104 slowly evolving single-copy nuclear genes (Srivastava et al. 2008). The two cnidarian representatives were not retrieved as monophyletic, although in branches with a low statistical support (PP = 69). The TOP3B phylogeny yielded Annelida as the sister-group of Mollusca (PP = 96), supporting the Lophotrochozoa hypothesis and previous molecular and morphological studies (Nosenko et al. 2013).

Our analyses placed arthropods and the nematodes clustering together (PP = 80) as predicted by the Ecdysozoan hypothesis (Aguinaldo et al. 1997), although our tree did not include some of the Ecdysozoa phyla. The phylogeny of the Class Insecta revealed a major split into two clades (PP = 100). One of the major clades (top clade of Supplementary Figure S3.2) includes three monophyletic orders (Lepidoptera, Coleoptera and Hymenoptera) from the Endopterygota or Holometabola superorder, which comprises the insects that undergo complete metamorphosis (Wiegmann et al. 2009). The root of this clade consists of Neoptera insects from the Paraneoptera (Thysanoptera, Psocodea and Hemiptera) and Polyneoptera (Blattodea) superorders. The bottom insect clade (Supplementary Figure S3.2) includes the monophyletic Diptera order and three Hemiptera species. This order was polyphyletic in our analyses, since *Acyrthosiphon pisum* branched in the top clade.

Our phylogenetic inference supports the positioning of the Cephalochordate *Branchiostoma floridae* at the base of chordates (PP = 100; Figure 3.1). The Actinopteri form a well-supported monophyletic group (PP = 96; Supplementary Figure S3.3). The coelacanth (*Latimeria chalumnae*) is at the root of the tetrapods (Amemiya et al. 2013), with *Xenopus tropicalis* branching before the coelacanth, although with a weak statistical support (PP = 68). The Aves clade forms a monophyletic group (PP = 100), clustering

with Reptilia (Supplementary Figure S3.4). The Mammalia class is also monophyletic (PP = 99), with two major branches representing the infraclasses Marsupialia and Placentalia (PP = 100), as shown in Supplementary Figure S3.5.

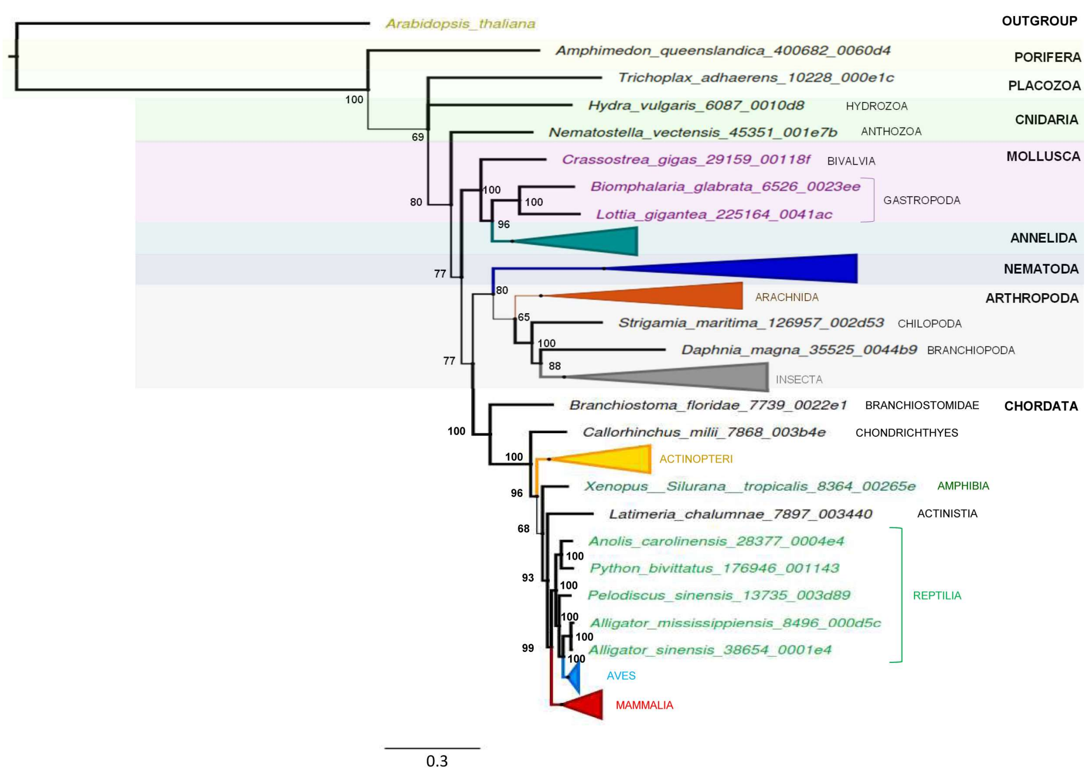


Figure 3.1. Phylogenetic analysis of TOP3B. Bayesian phylogenetic tree built with an alignment of 265 TOP3B protein sequences from metazoan and *Arabidopsis thaliana* as outgroup, including 706 amino acid positions. Bayesian posterior probabilities (PP, in percentage) are shown on the internal nodes. The scale bar indicates substitutions per site. Major monophyletic clades are collapsed for visualization purposes. The complete tree can be accessed in the supplementary material.

Strong purifying selection drives the evolution of TOP3B

Overall, we found a strong selective pressure to conserve the TOP3B sequence (Table 3.1). The ratio of fixed nonsynonymous to synonymous differences (dN/dS) among orthologues suggests purifying selection for functional constraint in all analysed taxonomic groups (dN/dS < 0.1). The analysis of saturation of substitution events showed saturation only in the third codon position of Arthropoda and Nematoda (Supplementary Figure S3.6). This could lead to some underestimation of dN/dS in these two groups, while dN/dS estimates in the other groups would not be biased due to saturation. Despite this possible bias, the data suggests that negative selection is present in all the groups given their extremely small dN/dS estimates (Table 3.1). The results suggest that TOP3B is an essential gene in different taxonomic groups and

redundancy with TOP3A is unlikely. Moreover, the strong purifying selection confirms the rigorous functional or structural requirements of TOP3B when interacting with DNA. We found that the amino acid site 284 is positively selected within the complete Chordata and Actinopteri dataset (Table 3.1). The human TOP3B position 284 has a Gln located in the arc formed by the topo-fold subdomain II, near the binding sites to TDRD3 (Figure 3.2).

Table 3.1. *TOP3B* estimation of selection. The Table includes the nucleotide diversity (π), best fitting substitution model of DNA evolution, the estimated global (entire sequences) *dN/dS* (including the 95% confidence interval) and the number of detected positively selected sites (PSSs) considering a cutoff of 0.05 for the *p*-value (for each significant PSS it includes the position, *dN-dS* and *p*-value).

Dataset	**Number of sequences**	**Nucleotide diversity**	**Substitution model**	**Global *dN/dS***	**Positively selected sites**
Chordata	235	0.21	SYM+I+G	0.076 [0.073–0.078]	1 (position 284, *dN-dS*=11.158, *p*-value=0.015)
Actinopteri	77	0.18	SYM+I+G	0.087 [0.084–0.090]	1 (position 284, *dN-dS*=7.546, *p*-value=0.037)
Reptilia	16	0.16	K80+I+G	0.054 [0.049–0.059]	0
Aves	45	0.06	SYM+I+G	0.043 [0.037–0.050]	0
Mammalia	93	0.11	GTR+I+G	0.051 [0.048–0.054]	0
Arthropoda	111	0.31	GTR+I+G	0.074 [0.072–0.076]*	0
Nematoda	7	0.33	GTR+G	0.093 [0.087–0.098]*	0

*The datasets of Arthropoda and Nematoda displayed saturation of substitution events in the third codon position, which could produce an underestimation of dN/dS.

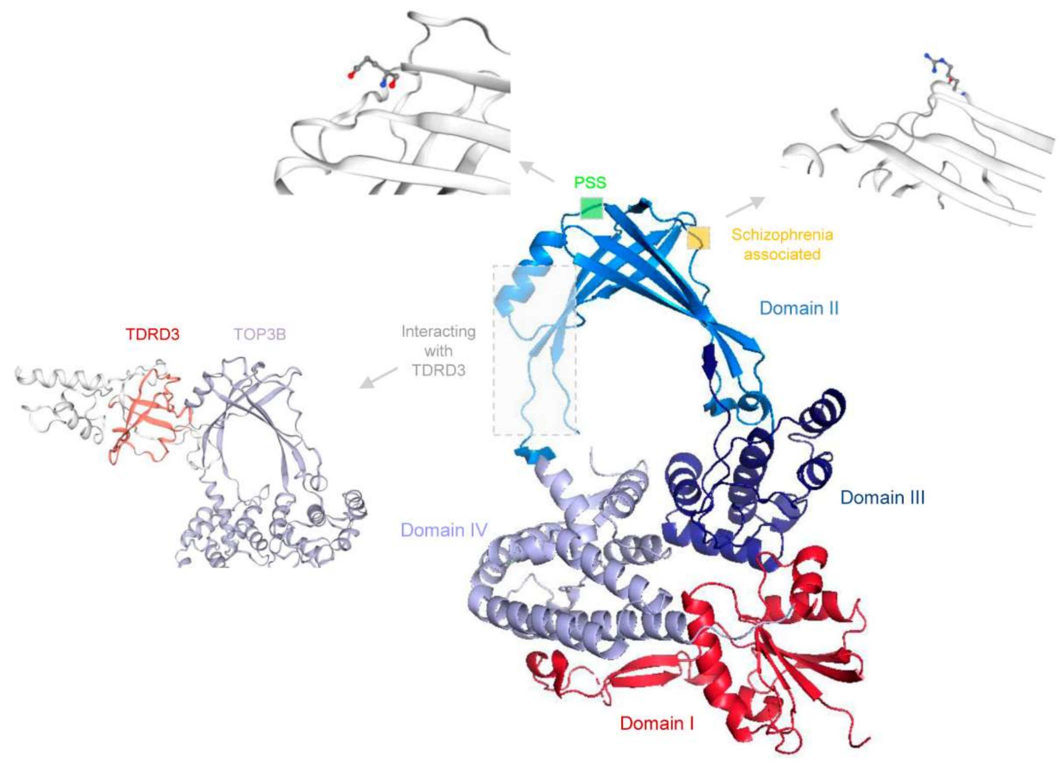
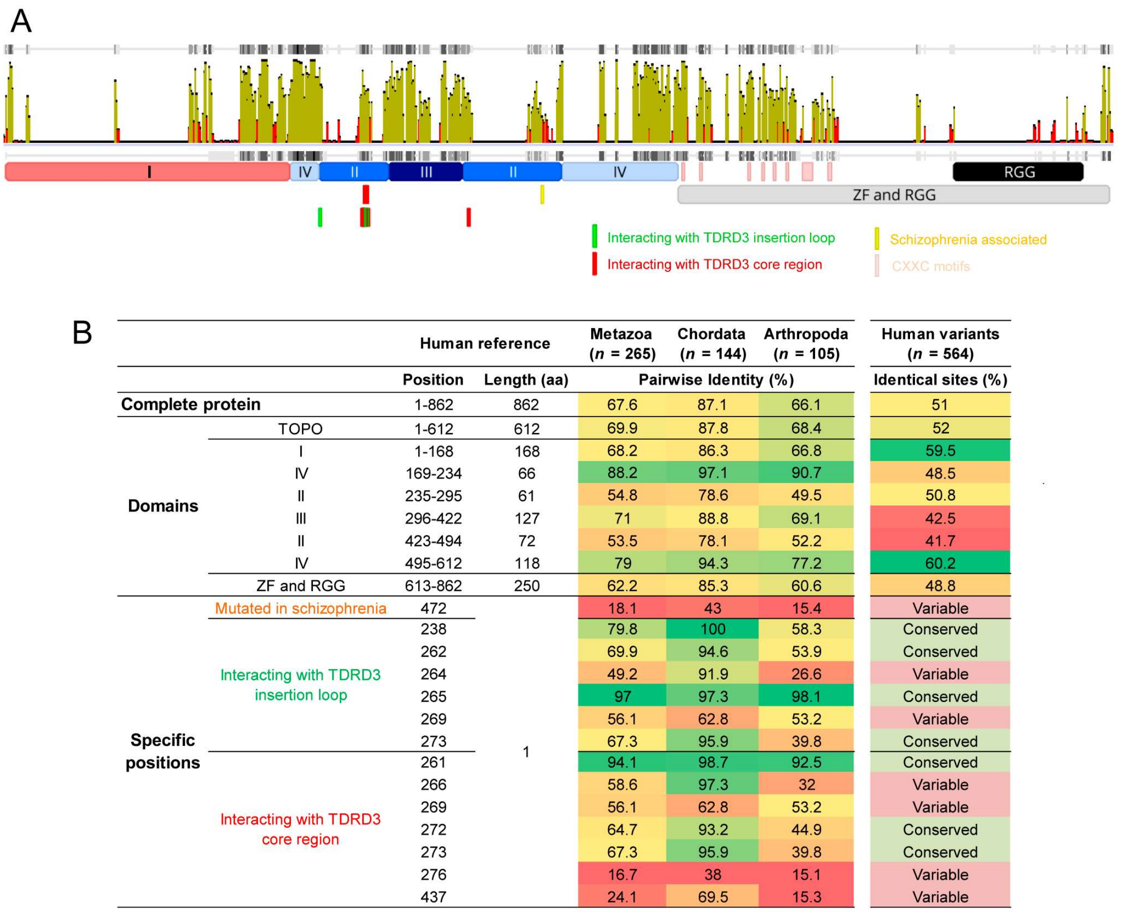


Figure 3.2. TOP3B structural organization. Overall structure of the TOP3B with major subdomains highlighted. The locations of the 472 site that is associated with schizophrenia, the 284 positively selected site (PSS) and interaction with TDRD3 are shown with extra panels.

The alignment of 265 metazoans TOP3Bs revealed a pattern of conserved regions interspersed with variable domains, both with amino acid substitutions and large insertion/deletions (Figure 3.3A). The overall pairwise identity of the metazoan alignment was 67.6%, being higher when only chordates were aligned (87.1%) in comparison with arthropods (66.1%) (Figure 3.3B). We found that subdomain IV is the most conserved subdomain across metazoan species, with a pairwise identity of 88.2% and 79% for its two segments (Figure 3.3B). The conservation is retained within chordates and arthropods, varying from 97.1% identity in chordates subdomain IV segment 169-234 to 77.2% in arthropods segment 495-612. When comparing different TOP3B variants in humans, subdomain IV segment 495-612 was also the most conserved with the highest percentage of invariable sites (Figure 3.3B). It has been suggested that subdomain IV provides a structural foundation for the protein (Figure 3.3), which may explain its high sequence conservation (Lima et al. 1994). Although the three-dimensional structure remains to be determined for most metazoan species, the observed conservation at the primary sequence level points towards a conserved protein structure.



		Human reference		Metazoa (*n* = 265)	Chordata (*n* = 144)	Arthropoda (*n* = 105)	Human variants (*n* = 564)
		Position	Length (aa)	Pairwise Identity (%)			Identical sites (%)
Complete protein		1-862	862	67.6	87.1	66.1	51
Domains	TOPO	1-612	612	69.9	87.8	68.4	52
	I	1-168	168	68.2	86.3	66.8	59.5
	IV	169-234	66	88.2	97.1	90.7	48.5
	II	235-295	61	54.8	78.6	49.5	50.8
	III	296-422	127	71	88.8	69.1	42.5
	II	423-494	72	53.5	78.1	52.2	41.7
	IV	495-612	118	79	94.3	77.2	60.2
	ZF and RGG	613-862	250	62.2	85.3	60.6	48.8
Specific positions	Mutated in schizophrenia	472	1	18.1	43	15.4	Variable
	Interacting with TDRD3 insertion loop	238		79.8	100	58.3	Conserved
		262		69.9	94.6	53.9	Conserved
		264		49.2	91.9	26.6	Variable
		265		97	97.3	98.1	Conserved
		269		56.1	62.8	53.2	Variable
		273		67.3	95.9	39.8	Conserved
	Interacting with TDRD3 core region	261		94.1	98.7	92.5	Conserved
		266		58.6	97.3	32	Variable
		269		56.1	62.8	53.2	Variable
		272		64.7	93.2	44.9	Conserved
		273		67.3	95.9	39.8	Conserved
		276		16.7	38	15.1	Variable
		437		24.1	69.5	15.3	Variable

Figure 3.3. TOP3B diversity and structural organization. A) Identity plot for the alignments of 265 TOP3B protein sequences from metazoan species. The most conserved positions are indicated by brown bars, the less conserved by red bars. The main TOP3B (sub)domains and motifs are indicated. The main 5-terminal TOPO domain includes the subdomains I (or TOPRIM), II, III and IV. The C-terminal region harbours eight highly conserved CXXC zink finger motifs (pink bars) and an arginine-glycine-glycine (RGG) box (black annotation). Highlighted are also the TOP3B residues interacting with the insertion loop (green bars) and core region (red bars) of TDRD3 and the site of the R472Q mutation previously associated with schizophrenia. B) Percentage of pairwise identity across the full TOP3B protein, domains, subdomains, motifs and relevant sites. The identity was obtained in three different alignments, with all metazoans or with chordates or arthropods alone. The right column indicates the percentage of identical sites considering all human missense variants reported so far.

Subdomain II was the least conserved, with its two segments showing an identity of near or slightly above 50% in metazoan and in arthropods alone (Figure 3.3B). The only polymorphism in exons of archaic hominins DNA occurred in subdomain II, although the amino acid was not changed (Supplementary Figure S3.7). In humans, segment 423-494 of subdomain II has the largest number of missense variants of all domains (Figure 3.3B). Subdomain II was also found less conserved than other subdomains amongst members of the *E. coli* DNA topoisomerase I subfamily (Lima et al. 1994). This subdomain is where TOP3B interacts with the OB-fold domain of TDRD3 and it has been

suggested that it is more flexible than domains I, III and IV (Goto-Ito et al. 2017). It can be hypothesized that such flexibility can be achieved with different amino acid compositions, thus explaining the lower conservation. Subdomains I and III showed intermediate levels of conservation, either considering all metazoans or chordates and arthropods alone (Figure 3.3B).

Identification of relevant TOP3B residues for interaction with TDRD3

Goto-Ito *et al.* identified several TOP3B residues interacting with the insertion loop and core region of TDRD3 (Figures 3.2A, 3.4A and 3.4B). In Metazoa, the residues interacting with the TDRD3 insertion loop (69.9% on average) where slightly more conserved than those interacting with the core region (54.5% on average). Very different sequence identity values per site were observed (Figure 3.3B), with some sites being extremely conserved (e.g., site 265 with 97% identity or site 261 with 94.1%) and other highly variable (e.g., site 276 with 16.7% or 437 with 24.1%). In chordates, most interacting sites have a conservation above 90%. In arthropods, only sites 265 (98.1%) and 261 (92.5%) were found highly conserved. The low conservation in arthropods (and other metazoan groups) comes from replacements with functionally equivalent residues that would result in similar protein structures. Moreover, compensatory replacements can also occur, as observed in the human Arg96 (TDRD3) - Asp266 (TOP3B) hydrogen bond pair replaced by the Gln96 (TDRD3) - Lys266 (TOP3B) pair in *D. melanogaster* (Goto-Ito et al. 2017). Such replacement can explain the low conservation observed in residue 266 and may be common to other sites.

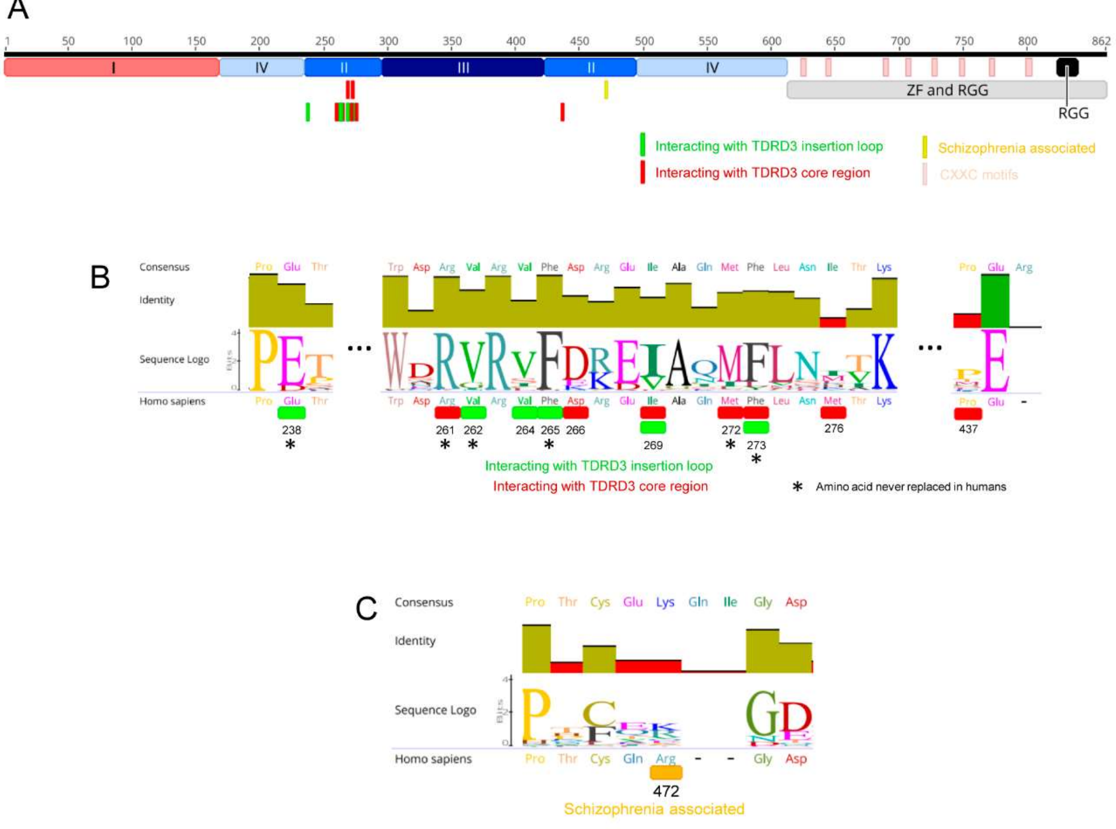


Figure 3.4. TOP3B interacting residues in TOPO domain. A) Localization of the TOP3B residues interacting with the insertion loop (green bars) and core region (red bars) of TDRD3 and the site of the R472Q mutation previously associated with schizophrenia. B) Identity plot and sequence logo for the TOP3B region harbouring the residues interacting with the insertion loop (green bars) and core region (red bars) of TDRD3. The protein sites that were always strictly conserved in humans are indicated by an asterisk. C) Identity plot and sequence logo for the TOP3B 472 site associated with schizophrenia.

The conservation pattern observed among species is supported by the variation observed in humans (Figures 3.3B and 3.4B). Two of the six sites interacting with TDRD3 insertion loop that display missense variants in humans coincide with the less conserved regions among metazoans (sites 264 and 269). Similarly, the four (out of seven) less conserved sites interacting with TDRD3 core region in metazoans were those found variable in humans. The selective pressures to maintain invariable some of these interacting sites seem to be the same in humans and across other animals. Although these sites could have conservative substitutions with amino acids displaying similar biochemical properties, our data suggest that sites 238, 261, 262, 265, 272 and 273 are conserved probably due to strong selective forces caused by their role in relevant protein interactions.

TOP3B residue 472 linked to schizophrenia is variable across Metazoa

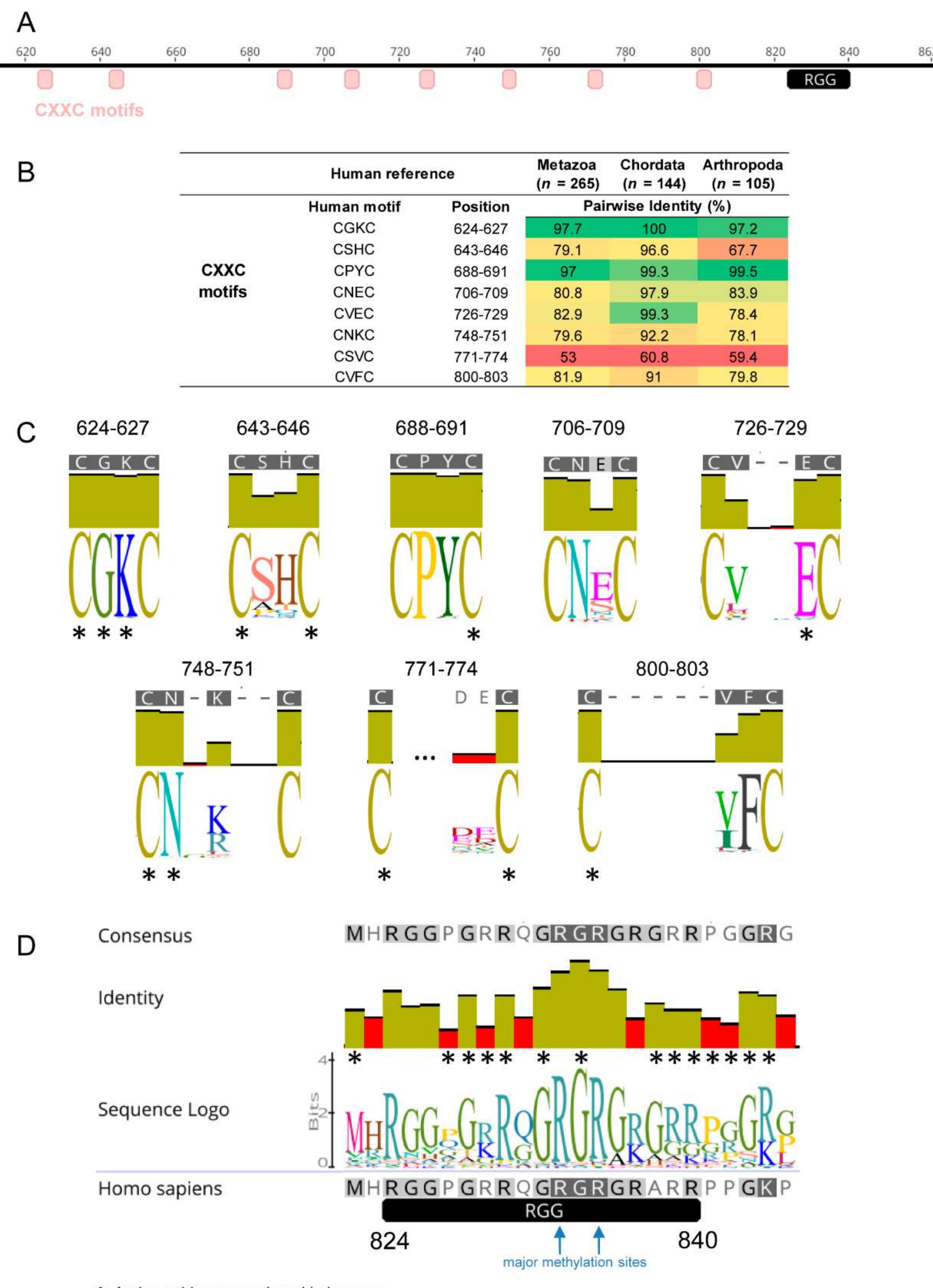
TOP3B *de novo* mutations and copy number variants have been associated to an increased risk of neurodevelopmental and cognitive disorders, namely schizophrenia and autism (Ahmad et al. 2017; Alemany et al. 2015; Daghsni et al. 2018; Rosato et al. 2019; Stoll et al. 2013). We assessed the degree of conservation of a TOP3B mutation (R472Q) previously associated with schizophrenia (Xu et al. 2012). Position 472 stands out as being highly variable across metazoan species (18.1% of identity), even when only chordates are considered (43%) (Figure 3.3B). The glutamine (G) variant (causing disease in humans) was found in 18 (6.9%) out of the 265 metazoan species, but not in species at the root of the tree: Porifera (Lys), Placozoa (Arg) and Cnidaria (Ile and Lys). It is therefore difficult to identify the ancestral state in this position. The flanking region of residue 472 is also variable, with the three upstream positions varying considerably (Figure 3.4C). Position 473 downstream is indeed more conserved (67.7%) than position 472. In humans, two SNPs that change the amino acid were reported (rs146766833 and rs116628543). The frequency of such alternate variants was always less than 0.004% in various studies and populations (Supplementary Figure S3.8). Overall, the data for position 472 contradicts the evolutionary conservation normally found in disease-causing amino acid changes (Miller and Kumar 2001).

The R472Q mutation does not seem to affect the TOP3B tertiary structure neither its binding to TDRD3 (Goto-Ito et al. 2017). Recent studies showed that TOP3B and TDRD3 form a ternary complex (named TTF complex) with the fragile X mental retardation protein (FMRP) in neurons (Linder et al. 2008; Stoll et al. 2013; Xu et al. 2013). In the context of TTF complex, TOP3B regulates the expression of mRNAs that are important in neurodevelopment. It is possible that TOP3B residue 472 is relevant for the proper activity of the TTF complex, whose disruption may affect neuronal development. It is unknown to which extent the TTF complex occurs across metazoan, but data suggests that it is also present in *Drosophila* (Xu et al. 2013) in addition to chordates. The low conservation observed at TOP3B position 472 could in part be related to the specificity of the TTF complex in only a subset of metazoan groups. In humans, alternate alleles that change the amino acid are rare, but have been detected in population studies (although some may occur in heterozygosity). It would be important to study such individuals to assess the possible existence of a predisposition for neurological disorders or to identify compensatory changes in other proteins interacting with TOP3B that may compensate the change and restore the proper function of the proteins. In any case, a strong selective pressure to maintain site 472 invariable would only exist if individuals were significantly affected in their reproductive fitness when this position is mutated. It is

not clear to which degree the alteration at site 472 affects individuals, particularly in species other than humans. If survival and reproduction are not severely affected, accumulation of genetic changes and amino acid replacements are expected, which are particularly visible when very divergent lineages are compared, as presented here.

TOP3B C-terminal CXXC zinc finger motifs are highly conserved in metazoans

The different C-terminal regions of TOP3B and TOP3A might explain their different functions and specialization (Bizard and Hickson 2020). TOP3B has been described to include multiple zinc finger motifs possibly involved in protein-DNA and protein-protein interactions (Wilson et al. 2000). Wilson *et al.* identified eight highly conserved CXXC motifs in the TOP3B C-terminal region (Figure 3.5A). CXXC zinc finger motifs are found in proteins with functions related to DNA or chromatin modification, in some cases mediating specific binding to double stranded DNA templates containing unmethylated CpG sites (Frauer et al. 2011; Lee et al. 2001; Thomson et al. 2010). Here we could confirm that the peripheral C residues are highly conserved in all eight motifs, with only a few exceptions across metazoans. When considering the four CXXC residues, seven of the eight motifs are conserved across metazoan with a percentage of identity above 79 (Figure 3.5B). The most conserved motif (% identity > 97) has a CGKC sequence and is the one located immediately downstream subdomain IV, at positions 624-627 of the human reference sequence (Figure 3.5C). In this motif, three out of the four residues were not found to vary in humans. A similarly high level of conservation was observed for motif CPYC (688-691), reaching more than 97% of identity in all alignments. Motif CSHC (643-646) is less conserved in arthropods (67.7%) than chordates (96.6%). Although conservation levels are higher in chordates overall, the difference to arthropods is more pronounced in motif CSHC. In this motif, the peripheral C amino acids were always conserved in humans. We found that motif CSVC (771-774) is poorly conserved in metazoans particularly when considering its two central residues, with identity values of 50 to 60% (Figures 3.5B and 3.5C). A large insertion of 20 amino acids occurs in *Anopheles darling.* Interestingly, only the C amino acids were found conserved in humans. Overall, although CXXC motifs vary in their conservation, the peripheral C residues tend to be maintained both in humans and other species, supporting their functional relevance.



	Human reference		Metazoa (*n* = 265)	Chordata (*n* = 144)	Arthropoda (*n* = 105)
	Human motif	Position	Pairwise Identity (%)		
CXXC motifs	CGKC	624-627	97.7	100	97.2
	CSHC	643-646	79.1	96.6	67.7
	CPYC	688-691	97	99.3	99.5
	CNEC	706-709	80.8	97.9	83.9
	CVEC	726-729	82.9	99.3	78.4
	CNKC	748-751	79.6	92.2	78.1
	CSVC	771-774	53	60.8	59.4
	CVFC	800-803	81.9	91	79.8

Figure 3.5. TOP3B C-terminal region. A) Location of the eight conserved CXXC zinc finger motifs and the arginine-glycine-glycine (RGG) box. B) Percentage of pairwise identity for the eight CXXC motifs in metazoans, chordates and arthropods. C) Identity plot and sequence logo for the eight CXXC motifs in the Metazoa alignment. Motif 771-774 has a large insertion removed for visualization purposes. The protein sites conserved in humans are indicated by an asterisk. D) Consensus sequence, identity plot and sequence logo of the RGG box (positions 824 to 840) obtained from the Metazoa alignment after removing insertions. Positions 833 and 835 are two major methylation sites.

Major TOP3B methylation sites at RGG box are retained across metazoans

The C-terminal domain of TOP3B harbours arginine-glycine-glycine (RGG) box motifs found in certain RNA-binding proteins, but absent from TOP3A and other mammalian topoisomerases (Bizard and Hickson 2020; Xu et al. 2013) (Figure 3.5A). The RGG box is important for TOP3B topoisomerase activity toward both DNA and RNA (Ahmad et al. 2017; Xu et al. 2013), being a target site for arginine methylation (Huang et al. 2018). We found that the RGG box is located in a protein region where some species present large insertions/deletions, which makes the alignment difficult to curate (Figure 3.3A). It is challenging to assess if the insertions are indeed part of the protein or result from poor database annotations. Whatever the case, we removed such insertions to better visualize the conservation pattern in the RGG box (Figure 3.5D). The RGG was defined between positions 824-840 of the human reference by Xu *et al.* (Xu et al. 2013). This RGG box has a pairwise identity of 47.1%, a value lower than that observed in all other protein motifs and domains. The most conserved segment of the RGG box was found near its centre, where a five amino acid GRGRG sequence (832-836) shows a 70.4% of identity. Interestingly, Huang *et al.* found R833 and R835 as the major methylation sites, which are located in the most conserved RGG box region. The authors showed that methylation-deficient TOP3B (R833/835K) exhibited weaker DNA topoisomerase activity, which led to accumulation of R-loops at target gene promoters and inhibition of gene transcription. It is therefore likely that a similar mechanism operates throughout metazoans, explaining the observed sequence conservation.

3.5. Conclusions

TOP3B is the only eukaryotic nuclear topoisomerase that appears not to be essential (Kwan and Wang 2001). Nevertheless, we found that TOP3B is conserved across major metazoan groups, allowing a robust reconstruction of the animal phylogeny. Our results suggest that TOP3B can be a good candidate for more detailed phylogenetic inferences in Metazoa as new genome sequences are becoming available. The relevance of TOP3B across animals is supported by a strong purifying selection to conserve its sequence, excluding any functional redundancy with TOP3A. In line with this observation, we confirmed the high conservation of CXXC ZF motifs and the central section of the RGG box in TOP3B C-terminal domain, a region not found in TOP3A and other mammalian topoisomerases. Interestingly, a mutated position (R472Q), known to be linked to schizophrenia, was found variable across metazoans, and detected at low frequencies in human populations. Overall, the data provided here can guide further research for a better comprehension of TOP3B evolution and function.

Acknowledgements

This work was partially supported by a research grant to FM (SFRH/BD/131584/2017) from Fundação para a Ciência e a Tecnologia. MA was supported by a research grant (RYC-2015-18241) from the “Ministerio de Ciencia, Innovación y Universidades” of the Spanish Government.

3.6. References

Abascal F, Zardoya R, Telford MJ (2010) TranslatorX: multiple alignment of nucleotide sequences guided by amino acid translations. Nucleic acids research 38:W7-W13.

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489-493.

Ahmad M, Shen W, Li W, Xue Y, Zou S, Xu D, Wang W (2017) Topoisomerase 3β is the major topoisomerase for mRNAs and linked to neurodevelopment and mental dysfunction. Nucleic acids research 45:2704-2713.

Alemany S et al. (2015) New suggestive genetic loci and biological pathways for attention function in adult attention-deficit/hyperactivity disorder. American Journal of Medical Genetics Part B: Neuropsychiatric Genetics 168:459-470.

Amemiya CT et al. (2013) The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311-316.

Berman HM et al. (2000) The protein data bank. Nucleic acids research 28:235-242.

Bizard AH, Hickson ID (2020) The many lives of type IA topoisomerases. Journal of Biological Chemistry 295:7138-7153.

Capranico G, Marinello J, Chillemi G (2017) Type i DNA topoisomerases. Journal of medicinal chemistry 60:2169-2192.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution 17:540-552.

Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. Annual review of biochemistry 70:369-413.

Collins AG, Cartwright P, McFadden CS, Schierwater B (2005) Phylogenetic context and basal metazoan model systems. Integrative and Comparative Biology 45:585-594.

Daghsni M et al. (2018) TOP3B: a novel candidate gene in juvenile myoclonic epilepsy? Cytogenetic and genome research 154:1-5.

Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature methods 9:772-772.

Del Amparo R, Branco C, Arenas J, Vicens A, Arenas M (2021) Analysis of selection in protein-coding sequences accounting for common biases. Briefings in Bioinformatics.

Duguet M, Serre M-C, de La Tour CB (2006) A universal type IA topoisomerase fold. Journal of molecular biology 359:805-812.

Edgar RC (2004) MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC bioinformatics 5:113.

Forterre P, Gribaldo S, Gadelle D, Serre M-C (2007) Origin and evolution of DNA topoisomerases. Biochimie 89:427-446.

Frauer C et al. (2011) Different binding properties and function of CXXC zinc finger domains in Dnmt1 and Tet1. PloS one 6:e16627.

Garnier F, Debat H, Nadal M (2018) Type IA DNA topoisomerases: A universal core and multiple activities. In: DNA Topoisomerases. Springer, pp 1-20

Goto-Ito S, Yamagata A, Takahashi TS, Sato Y, Fukai S (2017) Structural basis of the interaction between Topoisomerase IIIβ and the TDRD3 auxiliary factor. Scientific reports 7:42123.

Goulet I, Boisvenue S, Mokas S, Mazroui R, Côté J (2008) TDRD3, a novel Tudor domain-containing protein, localizes to cytoplasmic stress granules. Human Molecular Genetics 17:3055-3074.

Hansen G, Harrenga A, Wieland B, Schomburg D, Reinemer P (2006) Crystal structure of full length topoisomerase I from Thermotoga maritima. Journal of molecular biology 358:1328-1340.

Huang L, Wang Z, Narayanan N, Yang Y (2018) Arginine methylation of the C-terminus RGG motif promotes TOP3B topoisomerase activity and stress granule localization. Nucleic acids research 46:3061-3074.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

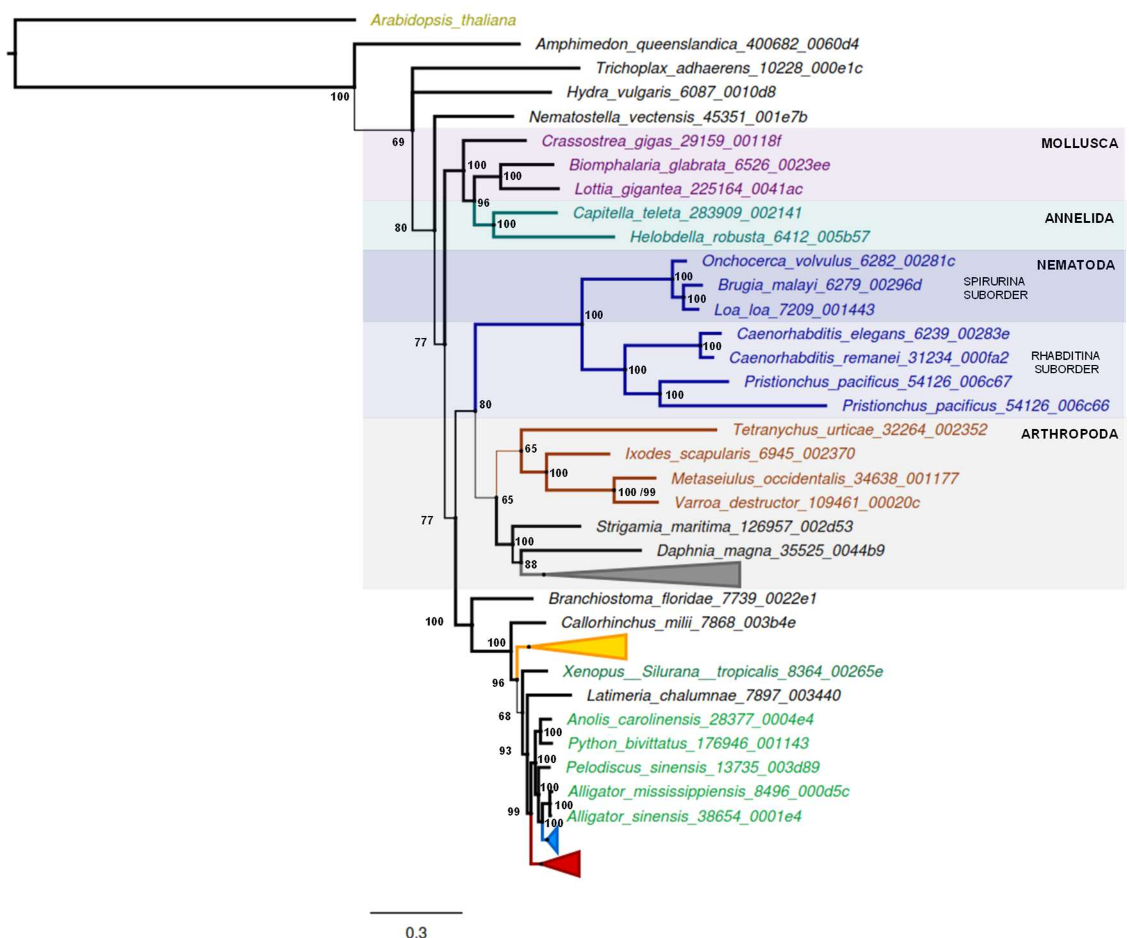
Hunt SE et al. (2018) Ensembl variation resources. Database 2018.

Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In: Parasite Genomics Protocols. Springer, pp 65-90

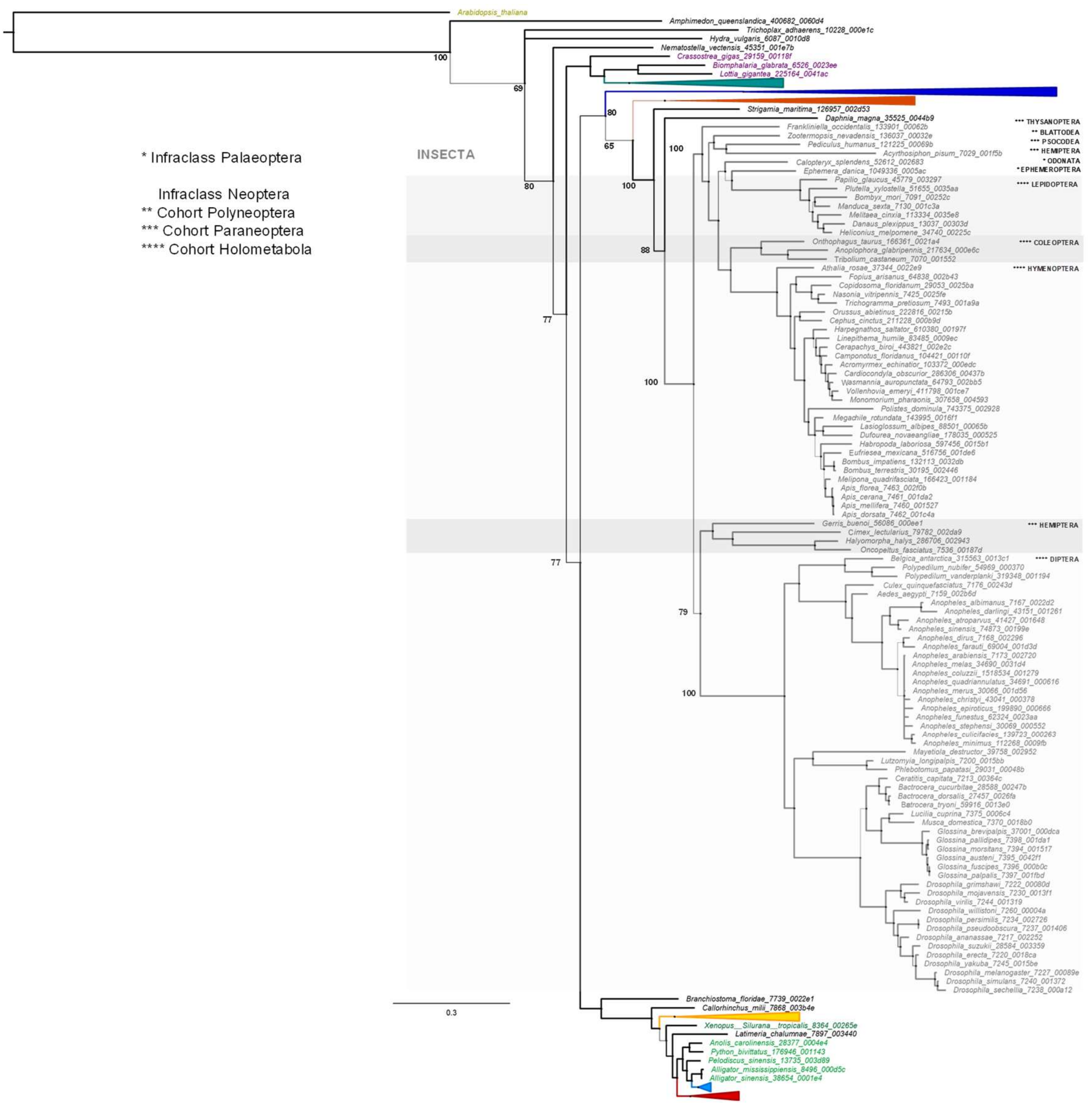
Kaufman CS, Genovese A, Butler MG (2016) Deletion of TOP3B is associated with cognitive impairment and facial dysmorphism. Cytogenetic and genome research 150:106-111.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome research 12:996-1006.
Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular biology and evolution 22:1208-1222.
Kosakovsky Pond SL et al. (2020) HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. Molecular biology and evolution 37:295-299.
Kozlov AM, Darriba D, Flouri T, Morel B, Stamatakis A (2019) RAxML-NG: a fast, scalable and user-friendly tool for maximum likelihood phylogenetic inference. Bioinformatics 35:4453-4455.
Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic acids research 47:D807-D811.
Kwan KY, Greenwald RJ, Mohanty S, Sharpe AH, Shaw AC, Wang JC (2007) Development of autoimmunity in mice lacking DNA topoisomerase 3β. Proceedings of the National Academy of Sciences 104:9242-9247.
Kwan KY, Wang JC (2001) Mice lacking DNA topoisomerase IIIβ develop to maturity but show a reduced mean lifespan. Proceedings of the National Academy of Sciences 98:5717-5721.
Lee J-H, Voo KS, Skalnik DG (2001) Identification and characterization of the DNA binding domain of CpG-binding protein. Journal of Biological Chemistry 276:44669-44676.
Lima CD, Wang JC, Mondragón A (1994) Three-dimensional structure of the 67K N-terminal fragment of E. coli DNA topoisomerase I. Nature 367:138-146.
Linder B et al. (2008) Tdrd3 is a novel stress granule-associated protein interacting with the Fragile-X syndrome protein FMRP. Human molecular genetics 17:3236-3246.
Luo A et al. (2010) Performance of criteria for selecting evolutionary models in phylogenetics: a comprehensive study based on simulated datasets. BMC Evolutionary Biology 10:1-13.
Miller MA, Pfeiffer W, Schwartz T Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 gateway computing environments workshop (GCE), 2010. Ieee, pp 1-8
Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Human molecular genetics 10:2319-2328.
Mondragón A, DiGate R (1999) The structure of Escherichia coli DNA topoisomerase III. Structure 7:1373-1383.
Nosenko T et al. (2013) Deep metazoan phylogeny: when different genes tell different stories. Molecular phylogenetics and evolution 67:223-233.
Pires R et al. (2014) Screening of copy number variants in the 22q11. 2 region of congenital heart disease patients from the São Miguel Island, Azores, revealed the second patient with a triplication. BMC genetics 15:1-11.
Rodríguez AC, Stock D (2002) Crystal structure of reverse gyrase: insights into the positive supercoiling of DNA. The EMBO journal 21:418-426.
Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.
Rosato M et al. (2019) Combined cellomics and proteomics analysis reveals shared neuronal morphology and molecular pathway phenotypes for multiple schizophrenia risk genes. Molecular Psychiatry:1-16.
Sherry ST, Ward M-H, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K (2001) dbSNP: the NCBI database of genetic variation. Nucleic acids research 29:308-311.
Srivastava M et al. (2008) The Trichoplax genome and the nature of placozoans. Nature 454:955-960.
Stoll G et al. (2013) Deletion of TOP3β, a component of FMRP-containing mRNPs, contributes to neurodevelopmental disorders. Nature neuroscience 16:1228-1237.
Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic biology 56:564-577.
Tarsitano M et al. (2014) Microduplications in 22q11. 2 and 8q22. 1 associated with mild mental retardation and generalized overgrowth. Gene 536:213-216.

Thomson JP et al. (2010) CpG islands influence chromatin structure via the CpG-binding protein Cfp1. Nature 464:1082-1086.
Wang JC (1985) DNA topoisomerases. Annual review of biochemistry 54:665-697.
Waterhouse A et al. (2018) SWISS-MODEL: homology modelling of protein structures and complexes. Nucleic acids research 46:W296-W303.
Wiegmann BM, Trautwein MD, Kim J-W, Cassel BK, Bertone MA, Winterton SL, Yeates DK (2009) Single-copy nuclear genes resolve the phylogeny of the holometabolous insects. BMC biology 7:1-16.
Wilson TM, Chen AD, Hsieh T-s (2000) Cloning and Characterization of *Drosophila*Topoisomerase IIIβ relaxation of hypernegatively supercoiled DNA. Journal of Biological Chemistry 275:1533-1540.
Xia X (2018) DAMBE7: New and improved tools for data analysis in molecular biology and evolution. Molecular biology and evolution 35:1550-1552.
Xia X, Lemey P (2009) Assessing substitution saturation with DAMBE. In: Vandamme A-M, Salemi M, Lemey P (eds) The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. 2 edn. Cambridge University Press, Cambridge, pp 615-630. doi:10.1017/CBO9780511819049.022
Xia X, Xie Z, Salemi M, Chen L, Wang Y (2003) An index of substitution saturation and its application. Molecular Phylogenetics and Evolution 26:1-7.
Xu B et al. (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nature genetics 44:1365-1369.
Xu D et al. (2013) Top3β is an RNA topoisomerase that works with fragile X syndrome protein to promote synapse formation. Nature neuroscience 16:1238-1247.
Yang Y, Lu Y, Espejo A, Wu J, Xu W, Liang S, Bedford MT (2010) TDRD3 is an effector molecule for arginine-methylated histone marks. Molecular cell 40:1016-1023.
Yang Y, McBride KM, Hensley S, Lu Y, Chedin F, Bedford MT (2014) Arginine methylation facilitates the recruitment of TOP3B to chromatin to prevent R loop accumulation. Molecular cell 53:484-497.

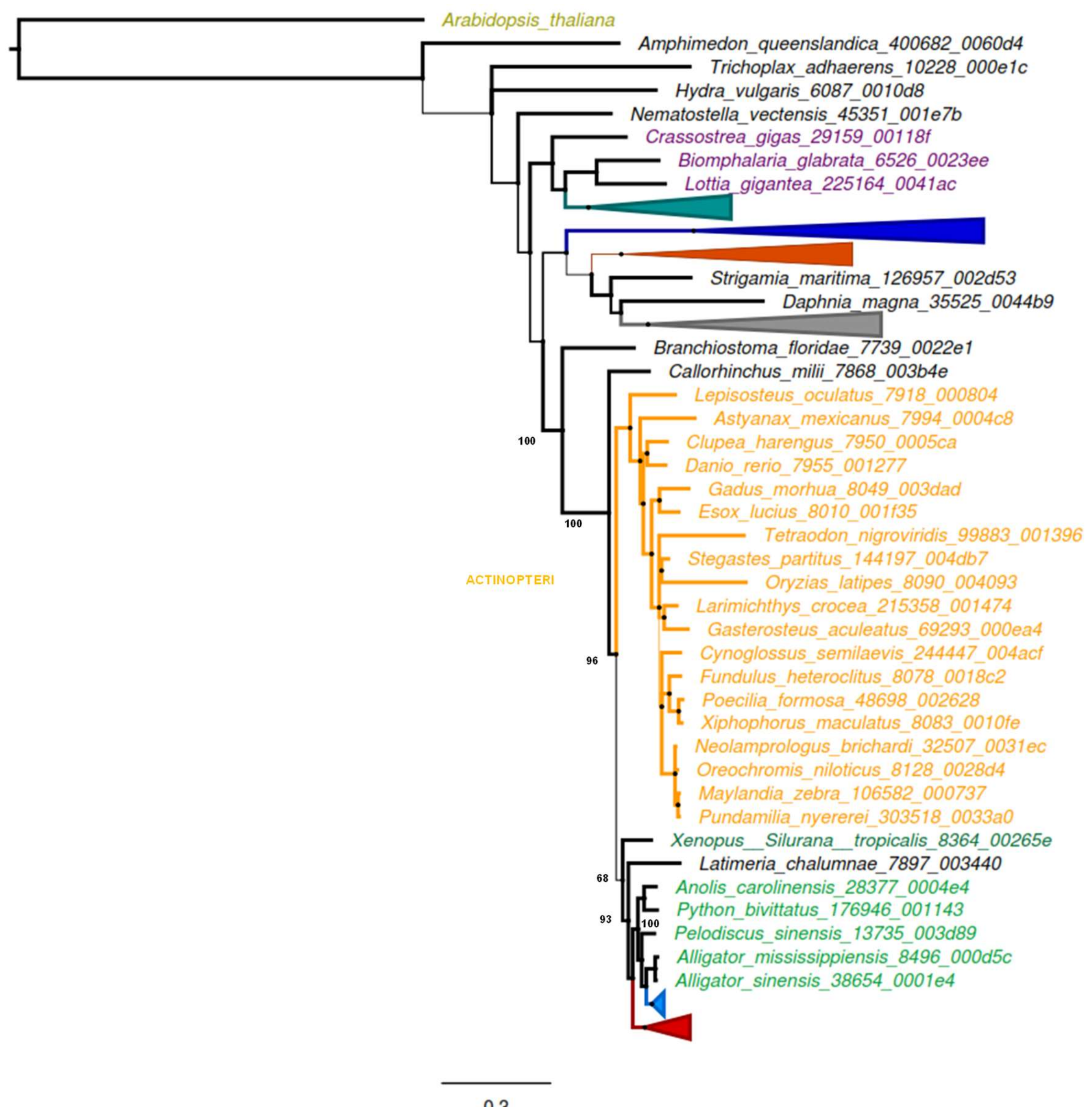
3.7. Supplementary Material



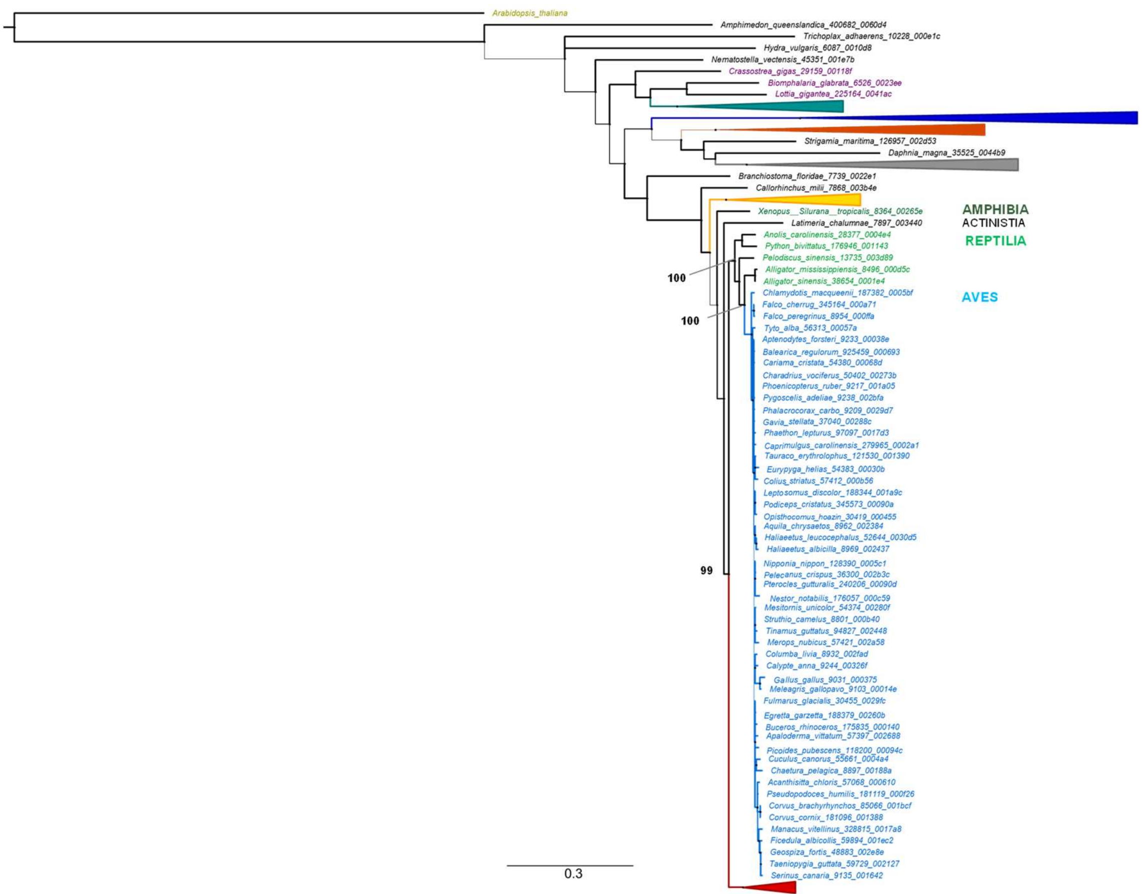
Supplementary Figure S3.1. Bayesian phylogenetic tree built with 265 TOP3B protein sequences from metazoan and Arabidopsis thaliana as outgroup. The Annelida and Nematoda clades are shown for all its species. Other monophyletic clades are collapsed for visualization purposes. Bayesian posterior probabilities (PP) are shown on the nodes. The scale bar indicates substitutions per site.



Supplementary Figure S3.2. Bayesian phylogenetic tree built with 265 TOP3B protein sequences from metazoan and Arabidopsis thaliana as outgroup. The Insecta clade is shown for all its species. Other monophyletic clades are collapsed for visualization purposes. Bayesian posterior probabilities (PP) are shown on the nodes. The scale bar indicates substitutions per site.



Supplementary Figure S3.3. Bayesian phylogenetic tree built with 265 TOP3B protein sequences from metazoan and Arabidopsis thaliana as outgroup. The Actinopteri clade is shown for all its species. Other monophyletic clades are collapsed for visualization purposes. Bayesian posterior probabilities (PP) are shown on the nodes. The scale bar indicates substitutions per site.



Supplementary Figure S3.4. Bayesian phylogenetic tree built with 265 TOP3B protein sequences from metazoan and Arabidopsis thaliana as outgroup. The Aves clade is shown for all its species. Other monophyletic clades are collapsed for visualization purposes. Bayesian posterior probabilities (PP) are shown on the nodes. The scale bar indicates substitutions per site.

Supplementary Figure S3.5. Bayesian phylogenetic tree built with 265 TOP3B protein sequences from metazoan and Arabidopsis thaliana as outgroup. The Mammalia clade is shown for all its species. Other monophyletic clades are collapsed for visualization purposes. Bayesian posterior probabilities (PP) are shown on the nodes. The scale bar indicates substitutions per site.

	First and second codon position		Third codon position	
Dataset	***Iss***	***Iss.c***	***Iss***	***Iss.c***
Chordata	0.12600	0.80825	0.59275	0.77375
Actinopteri	0.11975	0.80825	0.54575	0.80825
Reptilia	0.1833	0.7982	0.6800	0.7626
Aves	0.0500	0.80825	0.5275	0.77375
Mammalia	0.04575	0.80825	0.382	0.77375
Arthropoda	0.39775	0.81075	**0.88025**	**0.7775**
Nematoda	0.6243	0.8144	**1.0161**	**0.7823**

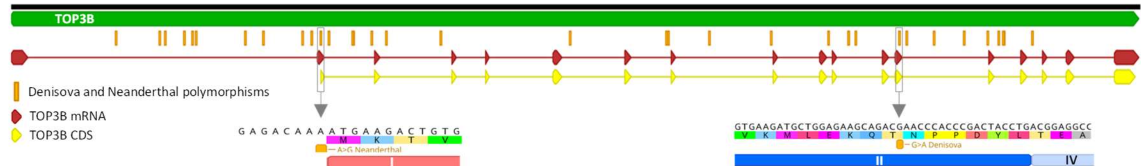
Supplementary Figure S3.6. Evaluation of saturation of substitution events in the studied datasets at the first and second codon positions and at the third codon position. For every dataset the table shows the index of substitutional saturation (Iss) and the critical Iss (Iss.c), see (Xia, et al. 2003; Xia and Lemey 2009; Xia 2018) for further details, for the first and second codon positions and at the third codon position of the studied datasets. Note that saturation of substitution events is detected if Iss > Iss.c (cases shown in bold).

Cited references

Xia X. 2018. DAMBE7: New and Improved Tools for Data Analysis in Molecular Biology and Evolution. Molecular Biology and Evolution 35:1550-1552.

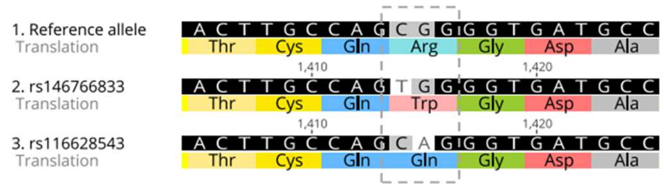
Xia X, Lemey P. 2009. Assessing substitution saturation with DAMBE. In: Vandamme A-M, Salemi M, Lemey P, editors. The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing. Cambridge: Cambridge University Press. p. 615-630.

Xia X, Xie Z, Salemi M, Chen L, Wang Y. 2003. An index of substitution saturation and its application. Molecular Phylogenetics and Evolution 26:1-7.



Human reference (NC_000022.11)				
Position	Region	Allele	Alternative allele	Species
21 980 398	Intron	G	A	Denisova & Neanderthal
21 979 405	Intron	C	T	Denisova
21 979 265	Intron	A	G	Neanderthal
21 978 846	Intron	G	A	Denisova
21 978 653	Intron	G	A	Neanderthal
21 978 571	Intron	G	A	Neanderthal
21 977 447	Intron	G	A	Denisova
21 977 033	Intron	T	C	Denisova
21 976 135	Intron	C	A	Denisova
21 975 920	Intron	C	T	Neanderthal
21 975 917	Intron	C	T	Neanderthal
21 975 710	Intron	A	G	Neanderthal
21 975 536	Intron	C	T	Neanderthal
21 974 997	Intron	C	T	Neanderthal
21 974 962	Intron	C	T	Denisova
21 974 559	Intron	C	T	Denisova
21 974 215	Intron	A	T	Denisova
21 972 970	Intron	G	A	Denisova
21 970 016	Intron	A	G	Denisova
21 967 820	Intron	G	T	Denisova
21 967 766	Intron	C	T	Denisova
21 966 831	Intron	C	T	Denisova
21 965 415	Intron	C	T	Neanderthal
21 964 108	Intron	G	T	Neanderthal
21 963 660	Intron	C	T	Denisova
21 963 659	Intron	G	A	Neanderthal
21 963 497	Intron	C	T	Denisova
21 962 490	Coding region	G	A	Denisova
21 962 319	Intron	C	T	Denisova
21 961 704	Intron	C	T	Denisova
21 961 022	Intron	G	A	Denisova
21 960 470	Intron	C	T	Neanderthal
21 960 232	Intron	C	T	Denisova
21 960 116	Intron	G	T	Denisova
21 960 091	Intron	A	G	Denisova
21 959 457	Intron	T	G	Denisova & Neanderthal

Supplementary Figure S3.7. Localization and list of polymorphisms identified in Denisova and Neanderthal TOP3B sequences.



SNP	rs146766833			rs116628543		
		Reference allele	Alternate allele		Reference allele	Alternate allele
Study	Sample size	C	T	Sample size	G	A
ALFA Allele Frequency	88414	0.99612	0.00388	11176	0.99991	0.00009
1000Genomes	5008	0.9996	0.0004	5008	0.9998	0.0002
ExAC	120346	0.997358	0.002642	120330	0.999934	0.000066
gnomAD - Exomes	250956	0.997701	0.002299	250894	0.999904	0.000096
gnomAD - Genomes	31386	0.99787	0.00213	n/a	n/a	n/a
GO Exome Sequencing Project	13006	0.99608	0.00392	13006	0.99992	0.00008

Supplementary Figure S3.8. Two SNPs change the amino acid at position 472 of TOP3B, associated in previous studies to an increased risk of neurodevelopmental and cognitive disorders. Data on global frequencies of reference and alternate alleles were obtained in the Ensembl and the NCBI dbSNP databases from different studies.

Chapter 4
Publication nº2

Journal of Molecular Evolution (2022) 90:149–165
https://doi.org/10.1007/s00239-022-10048-2

ORIGINAL ARTICLE

Evolutionary History of TOPIIA Topoisomerases in Animals

Filipa Moreira[1,2] · Miguel Arenas[3,4,5] · Arnaldo Videira[2,6,7] · Filipe Pereira[8,9]

Received: 17 November 2021 / Accepted: 26 January 2022 / Published online: 14 February 2022

4. Evolutionary history of TOPIIA topoisomerases in animals (Scientific Publication nº 2)

Filipa Moreira[1,2], Miguel Arenas[3,4,5], Arnaldo Videira[2,6,7] and Filipe Pereira[8,9*]

[1] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208, Matosinhos, Portugal
[2] ICBAS - Instituto de Ciências Biomédicas de Abel Salazar, Universidade do Porto, Rua Jorge de Viterbo Ferreira 228, 4050-313, Porto, Portugal
[3] Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain
[4] CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[5] Galicia Sur Health Research Institute (IIS Galicia Sur), 36310 Vigo, Spain
[6] IBMC-Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal
[7] i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
[8] IDENTIFICA genetic testing, Rua Simão Bolívar 259 3º Dir Tras. 4470-214 Maia, Portugal
[9] Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal

* Corresponding author. E-mail address: fpereirapt@gmail.com (F. Pereira)

4.1. Abstract

TOPIIA topoisomerases are required for the regulation of DNA topology by DNA cleavage and re-ligation and are important targets of antibiotic and anticancer agents. Humans possess two TOPIIA paralogue genes (*TOP2A* and *TOP2B*) with high sequence and structural similarity but distinct cellular functions. Despite their functional and clinical relevance, the evolutionary history of TOPIIA is still poorly understood. Here we show that TOPIIA is highly conserved in Metazoa. We also found that TOPIIA paralogues from jawed and jawless vertebrates had different origins related with tetraploidization events. After duplication, *TOP2B* evolved under a stronger purifying selection than *TOP2A*, perhaps promoted by the more specialized role of *TOP2B* in postmitotic cells. We also detected genetic signatures of positive selection in the highly variable C-terminal domain (CTD), possibly associated with adaptation to cellular interactions. By comparing TOPIIA from modern and archaic humans, we found two amino acid substitutions in the TOP2A CTD, suggesting that *TOP2A* may have contributed to the evolution of present-day humans, as proposed for other cell cycle-related genes. Finally, we identified six residues conferring resistance to chemotherapy differing between TOP2A and TOP2B. These six residues could be targets for the development of TOP2A-specific inhibitors that would avoid the side effects caused by inhibiting TOP2B. Altogether, our findings clarify the origin, diversification and selection pressures governing the evolution of animal TOPIIA.

Keywords: *TOP2A*, *TOP2B*, Metazoan phylogeny, selection, archaic humans, anticancer drugs

4.2. Introduction

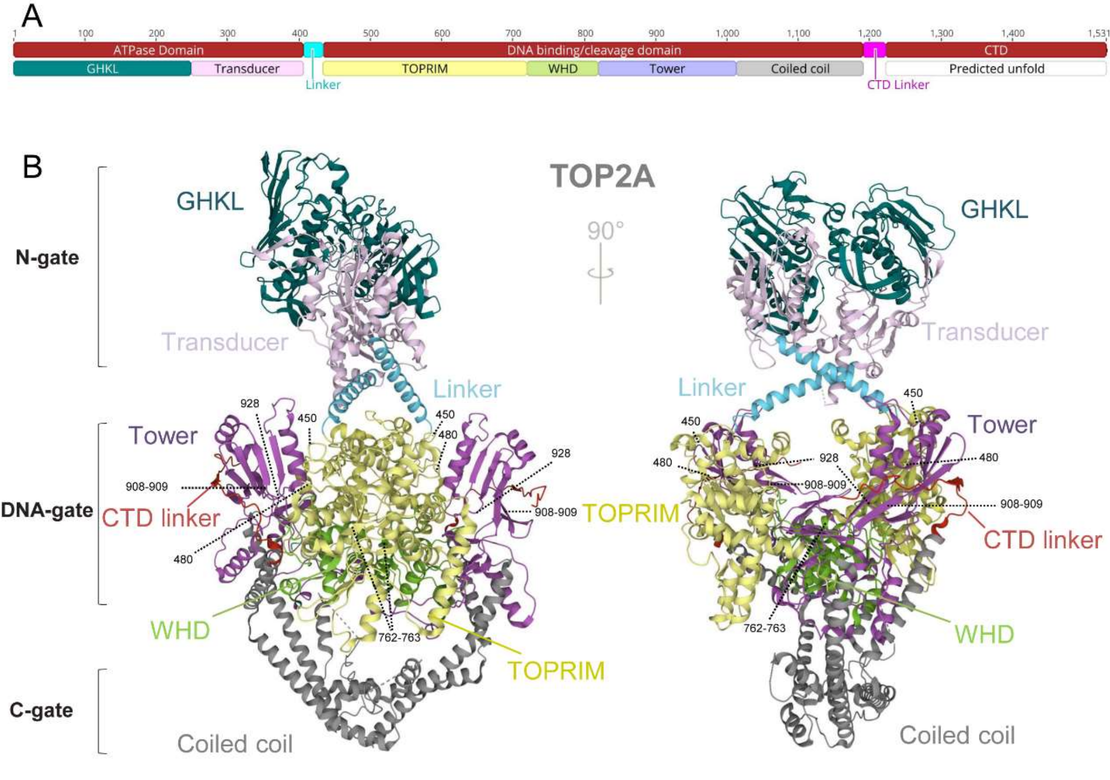
Topoisomerases are important enzymes that support cell viability and chromosome topology. They have the ability to cut, shuffle and reconnect DNA strands by adding or removing DNA supercoils and disentangle DNA segments (Champoux 2001; Wang 1996). Type IIA topoisomerases transiently cleave two strands of DNA and include bacterial and archaeal gyrase, bacterial topoisomerase IV and eukaryotic topoisomerase II. Some viruses also encode for type II topoisomerases, such as some Nucleocytoplasmic Large DNA Viruses (NCLDV) and T4-like bacterioviruses (Forterre et al. 2007; Gadelle et al. 2003; Schoeffler and Berger 2008). Type II topoisomerases have been identified in early diverging lineages of eukaryotes, such as kinetoplastid protozoans, *Giardia lamblia* and *Plasmodium falciparum* (Chakraborty and Majumder 1987; Cheesman et al. 1994; De et al. 2005; Strauss and Wang 1990). Previous studies have shown that most eukariotes have a single type IIA topoisomerase (TOPIIA), with the notable exception of vertebrates that have two paralogues, topoisomerase IIα (TOP2A) and topoisomerase IIβ (TOP2B) (Austin and Fisher 1990; Drake et al. 1987). In humans, TOP2A is encoded by the *TOP2A* gene on chromosome 17q21-22 (Tsai-Pflugfelder et al. 1988) and TOP2B by *TOP2B* gene on chromosome 3p24 (Austin et al. 1993; Jenkins et al. 1992; Tan et al. 1992). Both proteins display similar structures (Fig. 4.1) and biochemical activities but have different biological roles (Cornarotti et al. 1996; Drake et al. 1989; Leontiou et al. 2003; Marsh et al. 1996; Wang 2002).

TOPIIA activity relies on a mechanism that involves the controlled association and dissociation of three subunit-dimerization interfaces, or 'gates', termed the N gate, DNA gate and C gate, which guide the physical movement of one DNA duplex through another (Berger et al. 1996; Cabral et al. 1997; Roca et al. 1996; Roca and Wang 1992; Roca and Wang 1994; Wigley et al. 1991). TOPIIA can also be described as including three structural domains (Fig. 4.1): an N-terminal ATPase domain (NTD), a central catalytic DNA-binding/cleavage domain (CD) and a C-terminal domain (CTD). The N-terminal gate (ATPase domain) is composed by two elements: an ATP binding domain of the GHKL superfamily of proteins, including the Bergerat fold (Dutta and Inouye 2000), and an adjacent domain called the transducer (Classen et al. 2003; Corbett and Berger 2003; Wigley et al. 1991) that is thought to transmit signals (Corbett and Berger 2003; Kingma et al. 2000) from the N gate to the DNA gate (where DNA is bound and cleaved). The DNA gate is formed by a divalent metal-binding TOPRIM domain and a WHD similar to that typified by the catabolite activation protein (Aravind et al. 1998; Berger et al. 1996; Gajiwala and Burley 2000; McKay and Steitz 1981). The WHD contains the catalytic tyrosine and cooperates with the TOPRIM fold to cleave DNA, generating a pair of 5´cuts

staggered 4 bp apart from each other on opposite strands (Liu et al. 1983; Morrison and Cozzarelli 1979; Sander and Hsieh 1983). Adjacent to the WHD is a fold termed the “shoulder” or “tower” (Cabral et al. 1997), which also participates in DNA binding (Dong and Berger 2007). The third dimerization interface, the C-gate, is a coiled-coil element capped at its distal end with a small globular domain that extends from the Tower. This interface is well conserved and serves as the primary dimer interface of the protein (Berger et al. 1996; Cabral et al. 1997; Corbett et al. 2005; Fass et al. 1999; Laponogov et al. 2007).

TOP2A and TOP2B have very similar structures resulting from a high degree of sequence homology (~ 70 to 80 %) (Austin et al. 2018; Austin et al. 1993; Jenkins et al. 1992). The main differences reside in the C-terminal region, which is responsible for their different biological roles (Kozuki et al. 2017; Linka et al. 2007). The complete CTD crystal structure has not been determined, but secondary structure prediction suggests that it is structurally disordered (Broeck et al. 2021). The TOP2A CTD seems to act in the preferential relaxation of positive supercoils, whereas the equivalent region of TOP2B does not reveal a supercoil preference (McClendon et al. 2005). The C-terminal region contains nuclear localization signals and undergo extensive post-translational modifications (Lane et al. 2013; Lotz and Lamour 2020).

TOP2A is highly expressed during mitosis, being essential in proliferating cells and assists in chromosome segregation and replication (Akimitsu et al. 2003; Ali and Abd Hamid 2016; Cuvier and Hirano 2003; Grue et al. 1998; Niimi et al. 2001; Ye et al. 2010). On the other hand, TOP2B appears dispensable for cell proliferation, but regulates gene expression and is associated with developmental and differentiation events, in particular nerve growth and brain development (Austin et al. 2018; Bollimpelli et al. 2017; Ju et al. 2006; Lyu et al. 2006; Lyu and Wang 2003; McNamara et al. 2008; Tiwari et al. 2012; Yang et al. 2000). Mutations in *TOP2B* have been associated with B-cell immunodeficiency (Broderick et al. 2019; Papapietro et al. 2020), hearing loss (Xia et al. 2019) and intellectual disability (Lam et al. 2017). For instance, mice lacking TOP2A fail to develop beyond the 4–8-cell stage, while those without TOP2B exhibit a perinatal death due to defects in neuronal development (Akimitsu et al. 2003; Lyu and Wang 2003; Yang et al. 2000).



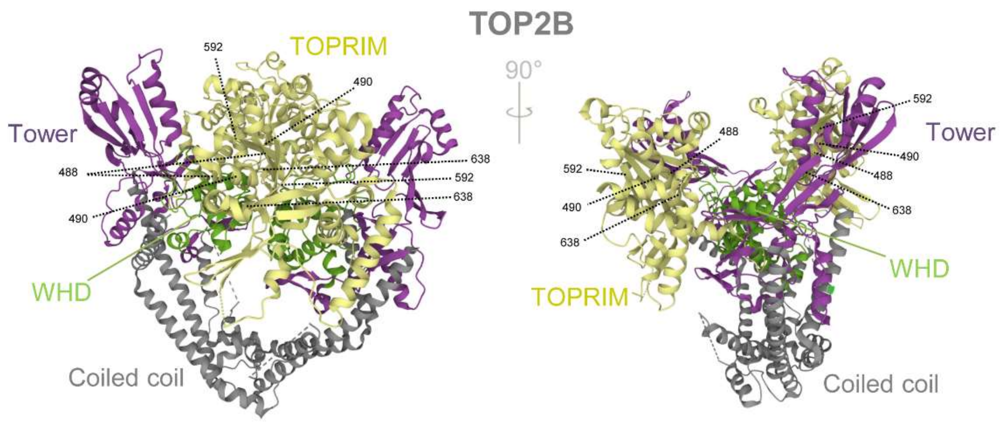


Figure 4.1. TOPIIA structural organization. A) Arrangement of human TOP2A with domains and other protein regions labelled and coloured. CTD, C-terminal domain. B) Cartoon representation of the human TOP2A and TOP2B structures with major domains highlighted.

In addition to their vital cellular functions, TOPIIA are a target for some of the most active anticancer agents (Nitiss 2009). TOP2A is responsible for the anticancer effects of TOP2 inhibitors due to its high activity in proliferating cells. TOP2 poisons (e.g., etoposide, doxorubicin) increase the levels of TOP2–DNA covalent complexes, resulting in double-strand DNA breaks that can cause cell death. TOP2B is believed to be responsible for most of the secondary malignancies and cardiotoxicity caused by TOP2-targeting drugs

due to its similar structure to TOP2A (Azarova et al. 2007; Chen et al. 2012; Haffner et al. 2010). Moreover, several mutations in *TOP2A* confer resistance to anticancer drugs (Nitiss and Beck 1996; Wu et al. 2011).

TOPIIA structure and function have been well studied in model organisms and humans, but often without including a comprehensive evolutionary analysis. Here, we address this limitation by providing a detailed evolutionary study of TOPIIA in animals. We provide a comprehensive phylogeny of TOPIIA in animals, including a detailed view of the duplication event that originated the *TOP2A* and *TOP2B* paralogues. We also identified the most conserved protein domains of functional relevance and assessed selective pressures governing the evolution of these important topoisomerases.

4.3. Material and Methods

TOPIIA sequences

We obtained TOPIIA protein sequences using the Ortho DB v10 (https://www.orthodb.org), which is a comprehensive catalogue of putative orthologues from more than 400 metazoan species (Kriventseva et al. 2019). We also used the protein-protein BLAST (blastp) to retrieve sequences from phyla that were not found in the Ortho DB, by using as query TOPIIA sequences from close phylogenetic groups. We excluded repeated sequences from the same species that showed 100% identity, which most likely represented different entries of the same sequence in the databases. We also removed short sequences with less than half of the average of TOPIIA length from further analyses as they can represent partial protein sequences derived from poorly assembled genomes. In some cases, contigs do not cover the complete genomic region where the protein is encoded, resulting in partial protein sequences. By this reason, some species lack one of the paralogues in our dataset, although it may be present in their genomes.

Denisovan and Neanderthal *TOP2A* and *TOP2B* sequences were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/) (Kent et al. 2002). We downloaded all BAM reads for tracks *Denisova* and *Neanderthal Cntgs* matching the Human Mar. 2006 (NCBI36/hg18) chr17:35,798,322-35,827,695 (*TOP2A*) and chr3:25,614,479-25,680,835 (*TOP2B*). The BAM reads from each track were then reassembled against the human *TOP2A* (NC_000017.11) and *TOP2B* (NC_000003.12) reference sequences using Geneious v2020.2.4 (http://www.geneious.com). We only considered a polymorphic position in archaic hominids when: 1) at least two reads overlap in that position; 2) the polymorphism represents more than 75% of all the reads and 3) the polymorphism is not at the end of a read.

TOPIIA multiple sequence alignments

We aligned the TOPIIA protein sequences with MAFFT version 7 (Katoh et al. 2019). The following three multiple alignments of protein sequences were used in subsequent analyses: Metazoa (n = 389), Chordata TOP2A (n = 105) and Chordata TOP2B (n = 125). The conservation across the alignments was estimated using the percentage of pairwise identity (PI) calculated in the Geneious program. Note that PI constitutes the average percent identity calculated by comparing the base pairs at every site.

TOPIIA coding domain sequences (CDS) from chordates were obtained from the Ensembl Genome Server (Hunt et al. 2018), as multiple sequence alignments of *TOP2A* (ENSG00000131747) and *TOP2B* (ENSG00000077097) human orthologues. The orthologues were organized in four different alignments for either *TOP2A* or *TOP2B:* Chordata (*n* = 159), Actinopteri (*n* = 51), Aves (*n* = 11) and Mammalia (*n* = 86). For coherence, we included the same species in the alignments of *TOP2A* and *TOP2B*.

The Ensembl server includes two long transcripts for *TOP2B*: TOP2B-201 (ENST00000264331.9) with 5814 nucleotides and 1626 amino acids and TOP2B-204 (ENST00000435706.6) with 5389 nucleotides and 1621 amino acids. Here we used the longest transcript (TOP2B-201) and resulting protein sequence unless stated otherwise.

The sequence alignments are available at Mendeley Data (https://data.mendeley.com//datasets/h2xfj5fsxw/1).

Phylogenetic analyses

The phylogenetic tree with all metazoan species (*n* = 389), considering *Arabidopsis thaliana* as outgroup, was obtained from the corresponding multiple alignment of protein sequences. We reconstructed a maximum likelihood (ML) phylogenetic tree with PhyML 3.0 (Guindon et al. 2010), implemented in the ATGC bioinformatics platform (http://www.atgc-montpellier.fr). The JTT +G +I substitution model of protein evolution was selected with the Smart Model Selection (SMS) v1.8.4 method implemented in PhyML (Lefort et al. 2017), under the Akaike Information Criterion (AIC). The branch support was evaluated with aBayes method (Anisimova et al. 2011). We analysed the duplication events in chordates with a phylogenetic tree based on the alignment of 34 CDS from Cephalochordata, Tunicata and Vertebrata species, considering *Acanthaster planci* (Echinodermata) as outgroup. Again, the ML tree was reconstructed with the ATGC bioinformatics platform, under the GTR +G +I substitution model and bootstrap based on 100 replicates. The resulting phylogenetic trees were edited with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree) and TreeViewer v1.2.2 (https://treeviewer.org).

Evaluation of selection

We evaluated molecular adaptation in *TOP2A* and *TOP2B* protein-coding sequence alignments with the nonsynonymous/synonymous substitution rate ratio (*dN*/*dS*) (Del Amparo et al. 2021; Jeffares et al. 2015). We started by identifying the best-fitting substitution model of DNA evolution with jModelTest2 (Darriba et al. 2012) and reconstructed a ML phylogenetic tree under the selected substitution model. Next, we

estimated *dN*/*dS* under a ML method considering the reconstructed phylogenetic tree with the well-established evolutionary framework Hyphy (Kosakovsky Pond and Frost 2005; Kosakovsky Pond et al. 2020). In particular, we applied the single-likelihood ancestor counting (SLAC) method, which provides *dN*/*dS* estimation with accuracy similar to that obtained with other likelihood-based methods (Kosakovsky Pond and Frost 2005). We identified global (entire sequence) genetic signatures of selection and positively selected sites (PSSs). The difference *dN*-*dS* was also used to evaluate selection at the site codon level.

TOPIIA protein structure

We used the TOP2A domains and other regions previously described (Broeck et al. 2021) and we considered them also in TOP2B by aligning the human reference sequences of both proteins. The TOP2A protein structure with PDB (Protein Data Bank) (Berman et al. 2000) code 6ZY7 (Broeck et al. 2021) and TOP2B structure with code 5ZAD (Sun et al. 2018) were obtained with Mol* (Sehnal et al. 2021) and RCSB PDB. The structures were coloured according to the protein domains and main regions.

4.4. Results and Discussion

TOPIIA proteins are conserved across Metazoa

The phylogenetic tree built with TOPIIA protein sequences placed Cnidaria at the root of Metazoa, which was expected considering that it was the only phylum in our dataset not belonging to Bilateria (Fig. 4.2, Supplementary Figure S4.1). In addition to Cnidaria, our dataset included eight Protostomia and 13 Deuterostomia phyla. Our phylogeny supports the split of Protostomia in Ecdysozoa (those exhibiting moulting) and Spiralia or Lophotrochozoa (those having lophophores and trochophore larvae), although only Spiralia was retrieved as a monophyletic group. Within Ecdysozoa, our results did not support the existence of Cycloneuralia, a clade including Scalidophora (represented here by Priapulida) and Nematoida (represented here by Nematoda) (Dunn et al. 2008). Indeed, our phylogenetic tree placed Nematoida more related to Panarthropoda (represented here by Tardigrada and Arthropoda) than to Scalidophora (Priapulida) (Campbell et al. 2011; Pisani et al. 2013). Spiralia formed a well-supported monophyletic group including Annelida, Mollusca, Brachiopoda and Platyhelminthes. The relationships within Spiralia are poorly resolved and often a matter of debate (Dunn et al. 2014). The clustering of Brachiopoda with Platyhelminthes supports the hypothesis that Lophophorata, organisms with a rake-like feeding structure (represented here by Brachiopoda), forms a separate clade from Trochozoa, a group defined by trochophore larvae, including at least Annelida and Mollusca (Dunn et al. 2014; Nesnidal et al. 2013) (Fig. 4.2).

The Deuterostomia clade includes a well-supported monophyletic group with Hemichordata and Echinodermata, which together form Ambulacraria. Within the Chordata, cephalochordates and tunicates diverged first and vertebrates formed a monophyletic group with the two paralogues *TOP2A* and *TOP2B* in different branches. In order to have a better resolution within Chordata, we built a phylogeny with all available TOPIIA CDS sequences from chordates (Fig. 4.3A). In the CDS tree, cephalochordates diverged first, with urochordates and vertebrates forming a sister group known as Olfactores, an arrangement that has now been widely accepted (Delsuc et al. 2006; Putnam et al. 2008; Satoh et al. 2014). We found particularly long branches in tunicates that suggests a high rate of molecular evolution, as previously noted for other genes (Delsuc et al. 2006; Tsagkogeorga et al. 2010). In particular, the extremely long branch of *Oikopleura dioica* (Fig. 4.3A) supports the claim that it is the fastest evolving metazoan recorded so far (Berna and Alvarez-Valin 2014; Denoeud et al. 2010).

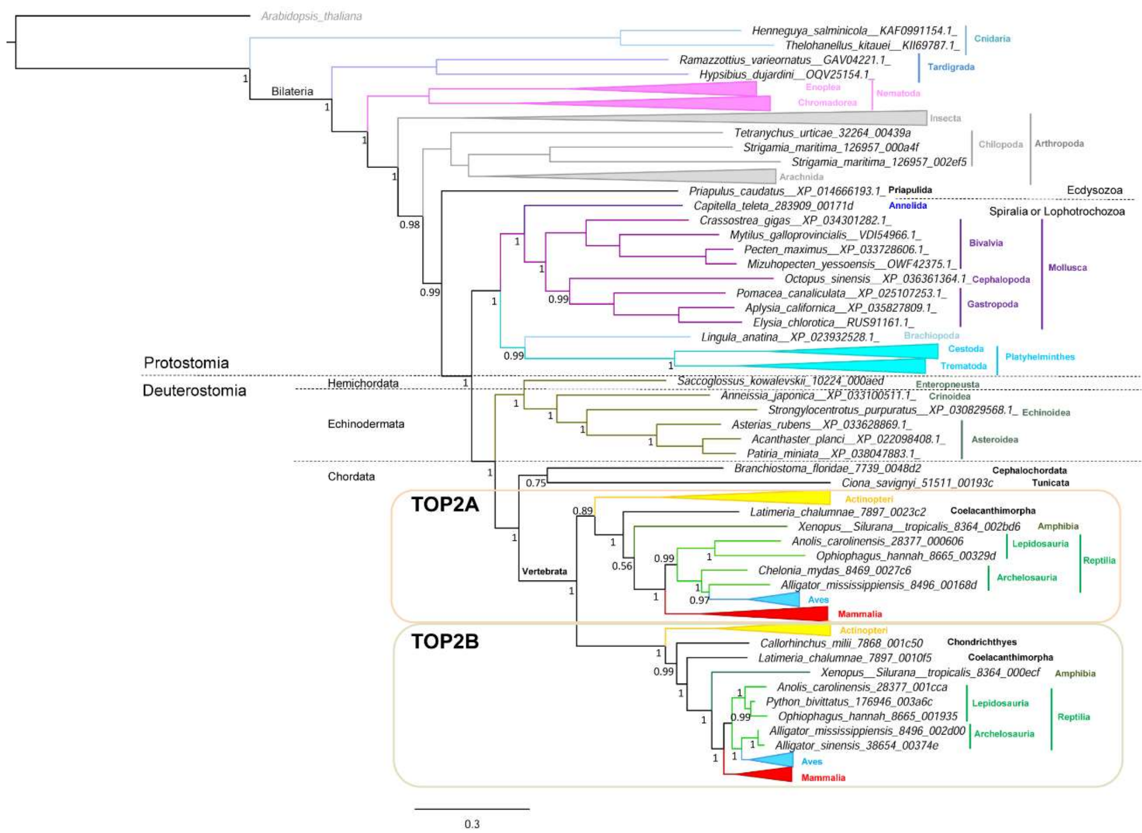


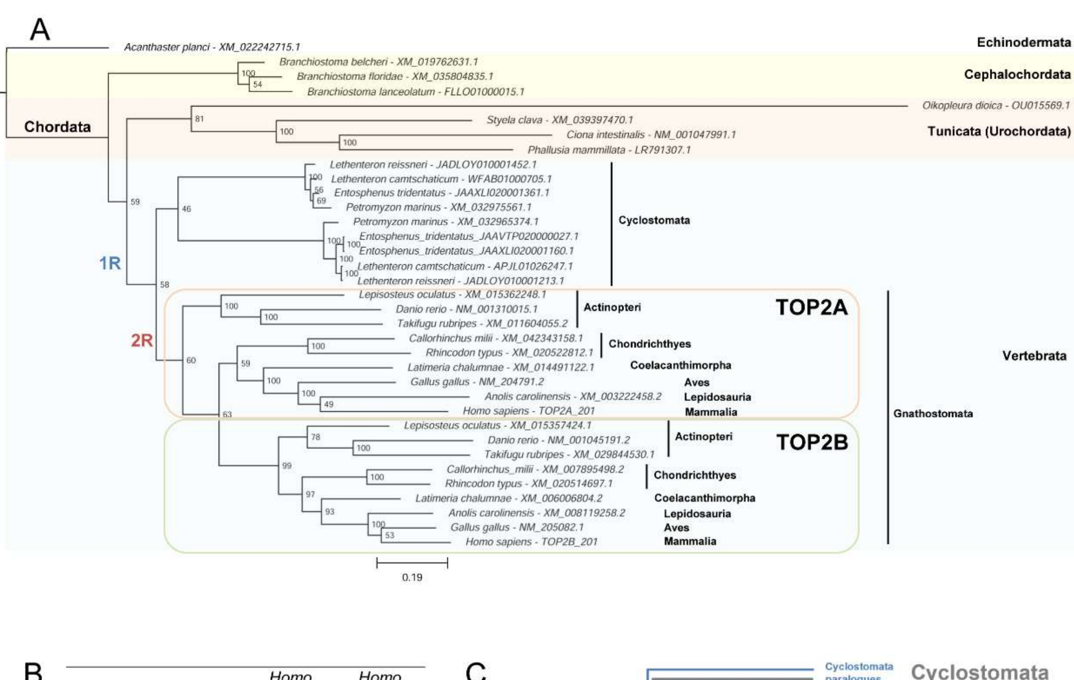
Figure 4.2. Phylogenetic analysis of TOPIIA in Metazoa. Maximum likelihood (ML) phylogenetic tree built with an alignment of 389 TOPIIA protein sequences from metazoans and considering *Arabidopsis thaliana* as outgroup. The branch support was estimated with aBayes, shown on the internal nodes. The scale bar indicates substitutions per site. Major monophyletic clades are collapsed for visualization purposes. The complete tree can be accessed in the supplementary material.

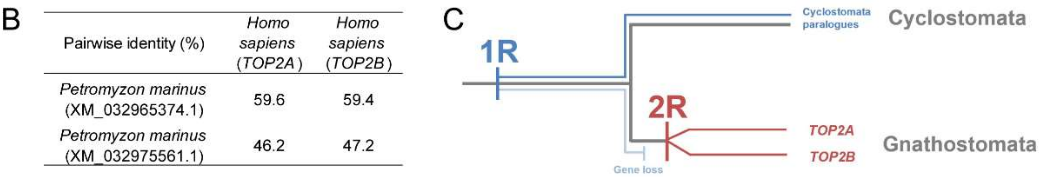
TOPIIA **genes duplicated independently in Cyclostomata and Gnathostomata**

We extensively searched for TOPIIA sequences in all available chordate genomes, and only found cases of TOPIIA paralogue genes within vertebrates. It is therefore evident that the formation of TOPIIA paralogues is related with vertebrate-specific genome duplication events. There is now convincing evidence that early vertebrate evolution is characterized by two rounds of tetraploidization (known as 1R and 2R), whose timing is still a topic of debate (Ohno 2013; Smith and Keinath 2015; Van de Peer et al. 2009). Irrespectively of when tetraploidization events occurred, our analyses suggest that the formation of paralogues in jawless vertebrates (Cyclostomata) and in jawed vertebrates (Gnathostomata) were independent events. Although with weak bootstrap values (~60%), our analyses always placed Cyclostomata paralogues in a different group from Gnathostomata *TOP2A* and *TOP2B* paralogues (Fig. 4.3A). In fact, Cyclostomata TOPIIA paralogues cannot even be classified as *TOP2A* or *TOP2B,* as they equally

diverged from Gnathostomata paralogues, as shown by the similar pairwise sequence identities between *Petromyzon marinus* and *Homo sapiens* paralogues (Fig. 4.3B).

There is a growing consensus that all vertebrates share the first tetraploidization event (1R), but only jawed vertebrates had a second whole genome duplication (2R) (Aase-Remedios and Ferrier 2021; Nakatani et al. 2021; Simakov et al. 2020), following previous studies (Escriva et al. 2002; Stadler et al. 2004). If that is the case, the most parsimonious succession of events to explain TOPIIA paralogues was: a) Cyclostomata paralogues originated during the 1R event; b) Gnathostomata lost one of the duplicated genes from 1R and c) Gnathostomata *TOP2A* and *TOP2B* originated in the 2R event (Fig. 4.3C). This scenario only assumes a single gene loss to explain the observed phylogeny. Other scenarios are less parsimonious by requiring two or more gene loss/duplication events (Supplementary Figure S4.2). For example, if 1R duplicated genes were retained in Gnathostomata at the time of 2R, two of the four resulting copies had to be subsequently lost. Nonetheless, further studies on early vertebrates will elucidate these evolutionary events.




B

Pairwise identity (%)	*Homo sapiens* (*TOP2A*)	*Homo sapiens* (*TOP2B*)
Petromyzon marinus (XM_032965374.1)	59.6	59.4
Petromyzon marinus (XM_032975561.1)	46.2	47.2

Figure 4.3. Evolutionary history of TOPIIA in basal chordates. A) Maximum likelihood (ML) phylogenetic tree built with an alignment of 34 TOPIIA coding domain sequences (CDS) from Cephalochordata, Tunicata and Vertebrata, considering *Acanthaster planci* (Echinodermata) as outgroup. The branch support was estimated with 100 bootstrap cycles, shown on the internal nodes. The scale bar indicates substitutions per site. The putative occurrence of two rounds of tetraploidization (1R and 2R) is indicated. B) Pairwise identity between *Petromyzon marinus* and *Homo sapiens* paralogues. C) Hypothetical succession of events to explain the presence of duplicated genes in Cyclostomata and Gnathostomata. The Cyclostomata paralogues originated

during the 1R event, the Gnathostomata lost one of the duplicated genes from 1R and the *TOP2A* and *TOP2B* paralogues originated in the 2R event.

Non-synonymous changes in TOP2A and TOP2B among modern and archaic humans

Neanderthals and Denisovans are the closest relatives to modern humans, diverging from the modern human lineage early in the Middle Pleistocene (Green et al. 2010; Reich et al. 2010). The recent shared ancestry explains the low genetic divergence among modern and archaic humans. We identified four polymorphisms among human, Denisovan and Neanderthal TOPIIA coding sequences (Table 4.1). The polymorphic positions were observed in the Transducer region of the ATPase domain (n = 1) and CTD (n = 3). The mutation in the *TOP2A* Transducer (position 267) occurred in the Denisovan lineage (A>G) without replacing the amino acid, perhaps the only possible type of mutation considering that this site is invariable in the 389 species analysed here. Two missense mutations were found in the *TOP2A* CTD. A T>C mutation in the human lineage replaced aspartate by glycine at position 1386. A C>A mutation in the Neanderthal lineage changed alanine by serine in position 1515 (Table 4.1). In both cases, the replacement was between amino acids with different charges or polarities. Such replacements may not affect the protein structure as the CTD is believed to be disordered. However, it is possible that such substitutions may impact the CTD interactions with other cellular components, raising the possibility that it may have contributed to the evolution of present-day humans. For example, position 1515 is within the chromatin tether domain (ChT), shown to be essential for TOP2A to interact robustly with chromosomes in mitosis (Lane et al. 2013). The only polymorphism detected in *TOP2B* occurs also in the CTD in the human lineage without replacing the amino acid. We found that all the polymorphic positions in the CTD are not conserved (< 50% pairwise identity), as expected considering the high variability of this protein region.

Very few amino acid changes become fixed in modern humans when comparing with archaic humans. For example, only 78 of those substitutions were described in the original publication of the Neandertal genome (Green et al. 2010). A recent survey, using data from different Neanderthal and Denisovan genomes, identified 571 genes with human-specific amino acid-changes (Kuhlwilm and Boeckx 2019). In fact, human *TOP2A* was identified as the protein with the largest number of interactions with other proteins (n = 53), suggesting that it may operate as an interaction hub in modifications of the cell division complex (Kuhlwilm and Boeckx 2019). Experimental studies exploring the influence of these changes on the cell cycle machinery could evaluate this intriguing hypothesis.

Table 4.1. Polymorphisms identified in Denisovan (Denis.) and *Homo neanderthalensis* (Neand.) *TOP2A* and *TOP2B* coding sequences.

Gene	***Homo sapiens* reference sequences**						**Ancestral State***		**Variant in archaic humans**		**Mutational event**				**Pairwise Identity (%) in Chordata**	
	Sequence	**Genome position**	**nt**	**Protein position**	**aa**	**Protein domain**	**nt**	**aa**	**Species**	**nt**	**Probable event**	**Lineage**	**Type**	**Amino acid replacement**	**TOP2A**	**TOP2B**
TOP2A	NC_000017.11	40,411,807	A	267	Phe	Transducer	A	Phe	Denis.	G	A>G	Denis.	Silent		100	
		40,391,616	C	1386	Gly	CTD	T	Asp	Denis.	T	T>C	*H. sapiens*	Missense	Asp>Gly	36	
		40,389,572	C	1515	Ala	CTD (ChT)	C	Ala	Neand.	A	C>A	Neand.	Missense	Ala>Ser	49.2	
TOP2B	NC_000003.12	25,609,599	A	1300	Gly	CTD	G	Gly	Denis.	G	G>A	*H. sapiens*	Silent			37

* Nucleotide in *Pan troglodytes*, *Pan paniscus* and *Gorilla gorilla*

***TOP2A* and *TOP2B* evolved under strong purifying selection but a few sites were positively selected**

We tested for the strength and mode of selection acting on *TOP2A* and *TOP2B* using *dN*/*dS* in chordates (Table 4.2). We found molecular signatures of purifying selection in both genes (*dN*/*dS* < 0.3), indicative of a strong selective pressure to conserve both TOP2A and TOP2B. It has been suggested that paralogues are subject to weaker purifying selection than single-copy genes (Kondrashov et al. 2002; Scannell and Wolfe 2008). In this concern, *dN/dS* values for *TOP2A* and *TOP2B* are higher than those estimated for Topoisomerase III Beta (*TOP3B*) in chordates (*dN*/*dS* < 0.1) (Moreira et al. 2021), which suggests that functional constraints were relaxed during the functional divergence to *TOP2A* and *TOP2B*. Moreover, the presence of (at least partially) redundant gene copies may have permitted the accumulation of previously forbidden deleterious mutations, which can explain the higher *dN/dS* values.

Paralogues may exhibit asymmetric rates of sequence evolution (Conant and Wagner 2003; Scannell and Wolfe 2008; Van de Peer et al. 2001). The strength of purifying selection was stronger in *TOP2B* (i.e., *dN*/*dS* = 0.156 on Chordata) than in *TOP2A* (e.g., *dN*/*dS* = 0.238 on Chordata), which suggests that *TOP2B* is under stronger functional constraints. Moreover, *TOP2A* displayed a higher nucleotide diversity and substitution rate than *TOP2B* (Table 4.2), which can also be noted in its longer branches in the reconstructed phylogenetic trees (Figs. 4.2, 4.3A). The specific activity of TOP2B in nerve growth and brain development (Lyu and Wang 2003; Yang et al. 2000) could impose relevant constrains on molecular evolution by the need of interacting with different partners and chemical environments. For example, it has been suggested that the role of TOP2B in the organism development involves the activation and repression of specific developmental genes (e.g., *Myt1l*, *Cacna2d1*, *Syt1*, *Kcnd2*) in association with diverse proteins (Lyu et al. 2006; Lyu and Wang 2003; Tiwari et al. 2012).

We detected signatures of positive selection in three *TOP2A* and two *TOP2B* sites in mammals (Table 4.2). Three of these positively selected sites (PSSs) are placed in the CTD, which is believed to conform specificity to the different activities of TOP2A and TOP2B (Kozuki et al. 2017; Linka et al. 2007). The TOP2A position 928 in the Tower domain was also found under positive selection (*dN*-*dS* = 4.832). The highest *dN*-*dS* values were obtained for the TOP2B position 28 in the ATPase domain in both the Chordata and Mammalia datasets (Table 4.2). The ATPase domain binds to ATP for a nucleotide-actuated protein dimerization gate through which DNA duplexes are passed. Considering the domains where the PSSs are located, we believe that these sites accumulated nonsynonymous substitutions over time to improve TOP2A and TOP2B

interaction and cellular functions, although it should be evaluated with experimental studies.

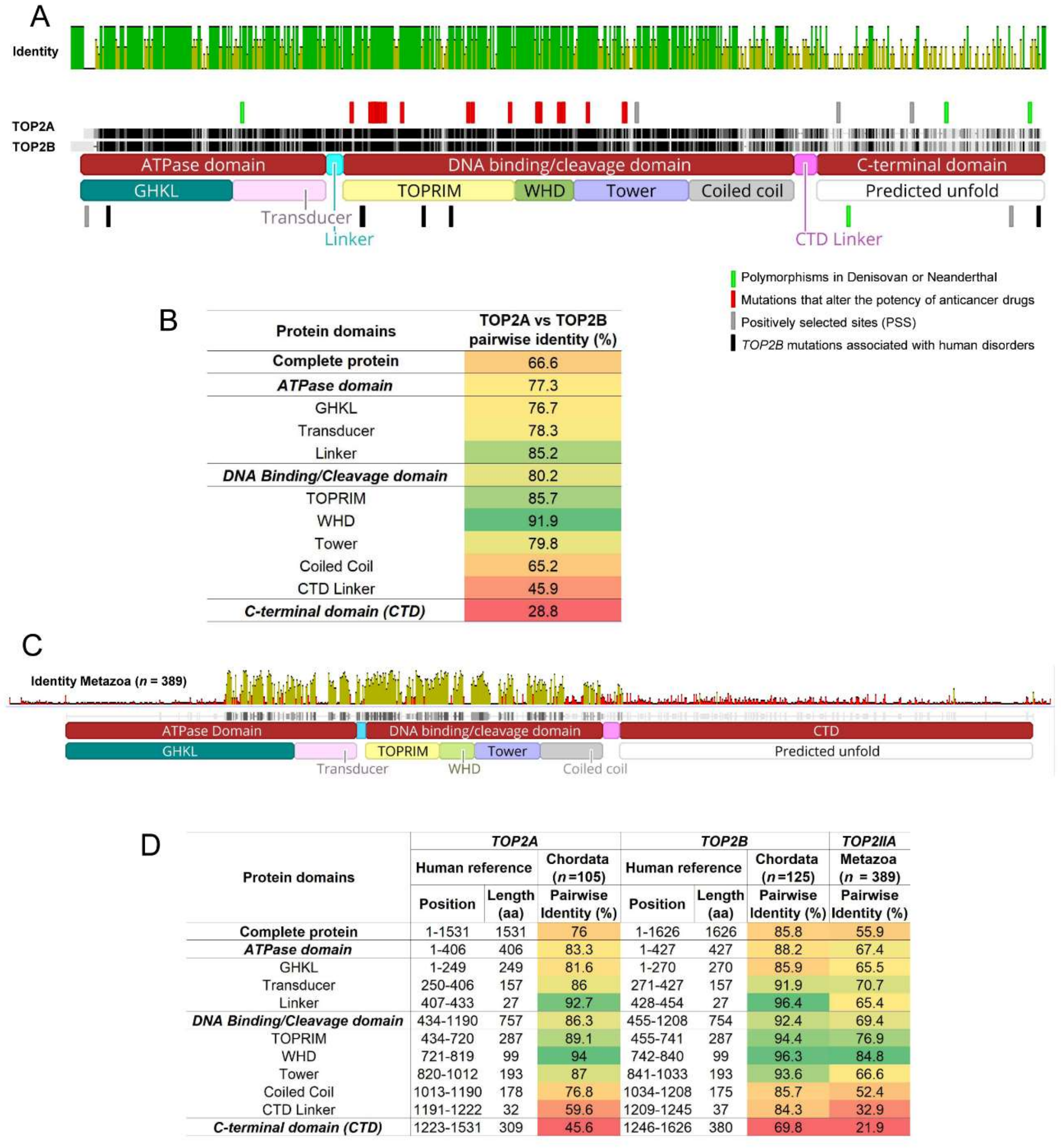
Table 4.2. Nucleotide diversity, substitution rates and selection pressure in *TOP2A* and *TOP2B*.

Gene	Dataset	*n*	Nucleotide diversity (π)	Substitution rate (Θ)*	Best substitution model	Global *dN/dS***	Positively selected sites***		
							Position	Region	*dN-dS* (*p*-value)
TOP2A	Chordata	159	0.16	813.76	GTR+I+G	0.238 [0.235–0.242]	-	-	-
	Actinopteri	51	0.17	780.58	GTR+I+G	0.226 [0.221–0.231]	-	-	-
	Aves	11	0.09	571.53	GTR+G	0.270 [0.255–0.285]	-	-	-
	Mammalia	86	0.07	616.23	GTR+I+G	0.220 [0.213–0.227]	928	Tower	4.832 (0.028)
							1255	CTD	7.009 (0.031)
							1343	CTD	6.911 (0.026)
TOP2B	Chordata	159	0.14	693.78	GTR+I+G	0.156 [0.152–0.158]	28	GHKL	20.444 (0.004)
	Actinopteri	51	0.12	719.68	GTR+G	0.164 [0.160–0.169]	-	-	-
	Aves	11	0.04	326.94	GTR+G	0.102 [0.091–0.115]	-	-	-
	Mammalia	86	0.06	470.58	GTR+I+G	0.119 [0.114–0.124]	28	GHKL	24.670 (4e-6)
							1572	CTD	5.318 (0.043)

* $\Theta = 4N\mu$, where N is the effective population size and μ is the mutation rate per generation, also named as Watterson Θ (Watterson 1975).
** Global (entire sequences) *dN/dS* including the 95% confidence interval.
*** Considering a cut-off of 0.05 for the *p*-value, position number according to the human reference sequence.

Low conservation in putative regulatory regions of the C-terminal domain (CTD)

Human TOP2A and TOP2B proteins had an overall pairwise identity of 66.6% (Figs. 4.4A, 4.4B). The DNA Binding/Cleavage domain was slightly more conserved (80.2%) than the ATPase domain (77.3%), while the paralogues mainly diverged in the CTD (28.8%). In this concern, the alignment of TOPIIA sequences of metazoans clearly showed the contrast between the poorly conservated CTD and the other well-conserved protein domains (Fig. 4.4C). The most conserved regions were the WHD (84.8%) and TOPRIM (76.9%) domains (Fig. 4.4D), where conservation can be explained by their critical role in interacting with DNA (Aravind et al. 1998; Gajiwala and Burley 2000; McKay and Steitz 1981; Roca et al. 1996). The same domains stand out as the most conserved when comparing TOP2A with TOP2B (Fig. 4.4D). All the other TOPIIA domains displayed a pairwise identity of around 50 to 70%, excepting the CTD linker (32.9%) and CTD (21.9%), which are almost impossible to align when using all metazoans. The CTD regulates nuclear localization and protein–protein interactions, which could have differentially evolved among species. Moreover, the CTD is believed to be a disordered region, which is known to evolve faster than well-structured regions (Brown et al. 2011). Overall, we found that TOP2B domains are more conserved than TOP2A domains (Fig. 4.4D), in agreement with the strong purifying selection observed in TOP2B (Table 4.2). We found poor conservation in the ending CTD 30 amino acids (positions 1502–1531) that constitute the ChT domain (Lane et al. 2013). The ChT domain had a 24.8% pairwise identity in Metazoa while a 55.7% in TOP2A Chordata. Similarly, the bipartite nuclear localisation signal (NLS) near the TOP2A CTD end (positions 1454 to 1497) (Mirski et al. 1997) was found poorly conserved in Metazoa (24.9%) and Chordata TOP2A (49.9%).



Protein domains	TOP2A vs TOP2B pairwise identity (%)
Complete protein	66.6
ATPase domain	77.3
GHKL	76.7
Transducer	78.3
Linker	85.2
DNA Binding/Cleavage domain	80.2
TOPRIM	85.7
WHD	91.9
Tower	79.8
Coiled Coil	65.2
CTD Linker	45.9
C-terminal domain (CTD)	28.8

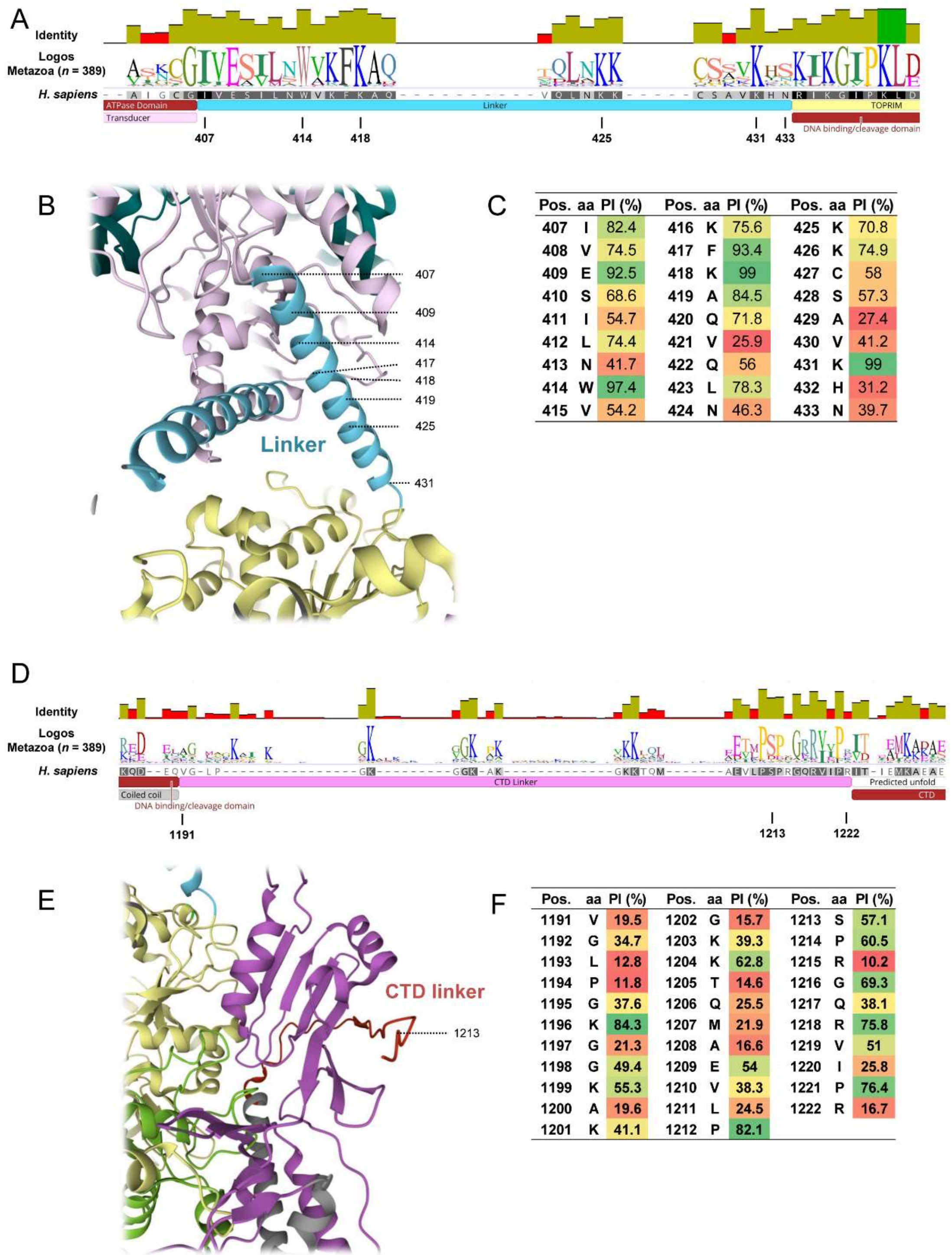
Protein domains	*TOP2A*			*TOP2B*			*TOP2IIA*
	Human reference		Chordata (*n*=105)	Human reference		Chordata (*n*=125)	Metazoa (*n* = 389)
	Position	Length (aa)	Pairwise Identity (%)	Position	Length (aa)	Pairwise Identity (%)	Pairwise Identity (%)
Complete protein	1-1531	1531	76	1-1626	1626	85.8	55.9
ATPase domain	1-406	406	83.3	1-427	427	88.2	67.4
GHKL	1-249	249	81.6	1-270	270	85.9	65.5
Transducer	250-406	157	86	271-427	157	91.9	70.7
Linker	407-433	27	92.7	428-454	27	96.4	65.4
DNA Binding/Cleavage domain	434-1190	757	86.3	455-1208	754	92.4	69.4
TOPRIM	434-720	287	89.1	455-741	287	94.4	76.9
WHD	721-819	99	94	742-840	99	96.3	84.8
Tower	820-1012	193	87	841-1033	193	93.6	66.6
Coiled Coil	1013-1190	178	76.8	1034-1208	175	85.7	52.4
CTD Linker	1191-1222	32	59.6	1209-1245	37	84.3	32.9
C-terminal domain (CTD)	1223-1531	309	45.6	1246-1626	380	69.8	21.9

Figure 4.4. TOPIIA diversity and structural organization. A) Identity plot for the alignments of human TOP2A and TOP2B reference sequences. The identical positions are shown in green bars and the different positions are shown in yellow bars. Highlighted are the positions polymorphic in Denisovan or Neanderthal (green bars), associated with resistance to anticancer drugs (red bars) and positively selected (grey bars). TOP2B mutations resulting in disease are indicated by black bars. B) Percentage of pairwise identity between human TOP2A and TOP2B protein complete sequence and domains. C) Identity plot for the alignment of 389 TOPIIA protein sequences from metazoan species. The most conserved positions are indicated with brown bars, the less conserved with red bars. The main protein domains are included. D) TOPIIA conservation across Metazoa. The percentage of pairwise identity was calculated for the full TOPIIA protein and domains in three different alignments, TOP2A (chordates), TOP2B (chordates) and all metazoans.

High conservation of the linker connecting the ATPase and the TOPRIM domains

We found that the linker connecting the ATPase and the TOPRIM domains (Fig. 4.5A) is well conserved in both TOP2A (92.7%) and TOP2B (96.4%), but more variable (65.4%) when all metazoans are compared (Fig. 4.4D). Two insertions of several amino acids are observed in *Trichuris suis* and *Habropoda laboriosa*. This linker forms an alpha helix with 29 amino acids connecting the N-gate to the DNA-gate (Figs. 4.1, 4.5B). Broeck et al. (2021) identified four highly conserved residues (positions 414, 417, 418 and 425) in this linker and tested different mutants to assess their contribution to the allosteric regulation of the human TOP2A. Our dataset of metazoan TOPIIA sequences confirmed that the positions 414, 417 and 418 were highly conserved (> 93%), but the position 425 showed a moderate level of conservation (70.8%; Fig. 4.5C). Indeed, positions 431 (99%), 409 (92.5%), 419 (84.5%) and 407 (82.4%) were also conserved, suggesting that they may play an important role in connections between the ATPase and TOPRIM domains. These residues could be tested in future experiments on allosteric regulation of TOP2A.

It has been suggested that the CTD linker (Figs. 4.5D, 5E) can structurally favour the curvature of the G-segment, stimulating DNA cleavage and facilitating strand passage (Broeck et al. 2021). Our analyses show that the CTD linker is poorly conserved in Metazoa, with several insertions and deletions (Figs. 4.4C, 4.4D). A better conservation was observed in Chordata TOP2A and TOP2B, but still much lower than other domains. In the Metazoa alignment, only the CTD linker region near the CTD is relatively conserved, with a few sites showing a pairwise identity above 75% (Fig. 4.5F). A few lysines (K) sites stand out as relatively conserved (Fig. 4.5F). The TOP2A position 1213 is phosphorylated during mitosis and contributes to localization of the protein to the centromere (Ishida et al. 2001). This position is moderately conserved in Metazoa (57.1%) and Chordata TOP2A (68.5%), but completely conserved in TOP2B (100%), which suggests that it is fundamental for the localization of TOP2B.



Pos.	aa	PI (%)	Pos.	aa	PI (%)	Pos.	aa	PI (%)
407	I	82.4	416	K	75.6	425	K	70.8
408	V	74.5	417	F	93.4	426	K	74.9
409	E	92.5	418	K	99	427	C	58
410	S	68.6	419	A	84.5	428	S	57.3
411	I	54.7	420	Q	71.8	429	A	27.4
412	L	74.4	421	V	25.9	430	V	41.2
413	N	41.7	422	Q	56	431	K	99
414	W	97.4	423	L	78.3	432	H	31.2
415	V	54.2	424	N	46.3	433	N	39.7

Pos.	aa	PI (%)	Pos.	aa	PI (%)	Pos.	aa	PI (%)
1191	V	19.5	1202	G	15.7	1213	S	57.1
1192	G	34.7	1203	K	39.3	1214	P	60.5
1193	L	12.8	1204	K	62.8	1215	R	10.2
1194	P	11.8	1205	T	14.6	1216	G	69.3
1195	G	37.6	1206	Q	25.5	1217	Q	38.1
1196	K	84.3	1207	M	21.9	1218	R	75.8
1197	G	21.3	1208	A	16.6	1219	V	51
1198	G	49.4	1209	E	54	1220	I	25.8
1199	K	55.3	1210	V	38.3	1221	P	76.4
1200	A	19.6	1211	L	24.5	1222	R	16.7
1201	K	41.1	1212	P	82.1			

Figure 4.5. Structure and diversity of TOPIIA linker regions. A) Identity plot and sequence logo for the linker joining the N-gate to DNA-gate. The most conserved positions are indicated by brown bars and the less conserved by red bars. B) Cartoon representation of the linker region in the human TOP2A structure. C) Percentage of pairwise identity per site for the linker region obtained from the alignment of 389 animal species. From D) to F), we show the same information as from A) to C) but for the C-terminal domain (CTD) linker.

***TOP2B* mutations associated with disease occur in conserved sites and replace amino acids with different physicochemical properties**

Alterations in topoisomerases have been associated with neurodegenerative and immune disorders and cancer (Pommier et al. 2016). TOP2A is essential for life, therefore mutations that significantly affect its activity in relaxing topological stress are expected to be lethal. In fact, human disorders caused by mutations in *TOP2A* are rare. For example, the Human Gene Mutation Database (http://www.hgmd.cf.ac.uk/) only reports a gross deletion associated with congenital heart disease (Glessner et al. 2014). On the other hand, TOP2B is not embryonic lethal and acts particularly in postmitotic cells. Perhaps because of this, a few cases of inherited *TOP2B* mutation associated with diseases have been reported (Broderick et al. 2019; Erdős et al. 2021; Lam et al. 2017; Papapietro et al. 2020; Xia et al. 2019). We identified six TOP2B mutations in the literature related with impaired B-cell development and function, hearing loss or neurodevelopmental disease (Table 4.3). A mutation at position 63 replaced a histidine by tyrosine in the Bergerat fold of the ATP binding domain, in an invariable site across Chordates. The mutation replaces a positively charged (histidine) by a neutral (tyrosine) amino acid. A total of four mutations were observed in the TOPRIM domain, all of them in highly conserved residues (PI > 95.2%). The heterozygous mutations affecting the TOPRIM were shown as partially dominant loss-of-function mutations (Broderick et al. 2019) and affect essential sites for the catalytic activity of the TOP2B. For example, the alanine to proline replacement at position 490 is predicted to destabilize an alfa helix within the TOPRIM domain (Papapietro et al. 2020). The Ser>Leu and Gly>Ser replacements change polar and nonpolar amino acids.

A single mutation (position 1618) was described in the TOP2B CTD (Table 4.3). The residue is poorly conserved (PI of 54.9%). The mutation occurs near the end of the TOP2B sequence in a region that is homologous to the TOP2A ChT domain (Lane et al. 2013), that facilitates stable binding to chromatin. Still, it remains to be determined if TOP2B presents a similar domain and if the identified mutation could affect its activity. Overall, *TOP2B* replacement mutations resulting in disease are found at conserved amino acid positions, a pattern also observed in other genes (Miller and Kumar 2001). The high conservation of several TOP2B residues suggests that other undetected mutations could cause similar diseases.

Table 4.3. *TOP2B* mutations associated with human disorders.

Position (CDS)		**Mutation**	**Position (Protein)**		**Protein region**	**Amino acid replacement**	**Disease**	**Reference**	**Pairwise Identity (%) in Chordata**
TOP2B-204*	**TOP2B-201****		**TOP2B-204***	**TOP2B-201****					
172	187	C>T	58	63	GHKL	His>Tyr	Global developmental delay and intellectual disability	(Lam et al. 2017)	100
1448	1463	C>T	483	488	TOPRIM	Ser>Leu	B cell immunodeficiency	(Broderick et al. 2019)	95.2
1453	1468	G>C	485	490	TOPRIM	Ala>Pro	B cell immunodeficiency	(Papapietro et al. 2020)	95.2
1761-1763	1776-1778	delAGA	587	592	TOPRIM	GluGlu>Glu	B cell immunodeficiency	(Broderick et al. 2019)	95.3
1897	1912	G>A	633	638	TOPRIM	Gly>Ser	B cell immunodeficiency	(Broderick et al. 2019)	98.4
4837	4852	G>C	1613	1618	CTD	Asp>His	Hereditary hearing loss	(Xia et al. 2019)	54.9

* TOP2B-204 (ENST00000435706.6); 5389nt; 1621aa
** TOP2B-201 (ENST00000264331.9); 5814nt; 1626aa

Six residues conferring resistance to TOP2 poisons differ among TOP2A and TOP2B, and can be used to develop paralogue-specific drugs

TOPIIA is a molecular target of several important classes of anticancer drugs, whose efficiency can be affected by mutations in critical protein sites (Delgado et al. 2018; Nitiss 2009). We analysed 27 amino acid replacements previously shown to confer resistance to anticancer drugs (Beck et al. 1993; Gilroy et al. 2006; Leontiou et al. 2006; Leontiou et al. 2007; Vassetzky et al. 1995), all of them located in the DNA Binding/Cleavage domain (Table 4.4). A total of six out of the 27 sites represented different amino acids in TOP2A and TOP2B sequences (positions 450, 480, 762, 763, 908 and 909 in TOP2A). These six residues are among the most variable sites in Chordata TOP2A. However, we did not find the same pattern in TOP2B, with only the position 929 being variable among chordates. TOP2B is believed to be responsible for undesirable side effects of anticancer chemotherapy by leading to therapy-related leukaemia (Azarova et al. 2007; Cowell et al. 2012). Therefore, we believe that future works could explore these six variable sites to design TOP2A-specific anticancer drugs with less undesirable side effects caused by interfering with TOP2B (Wu et al. 2013). With the exceptions mentioned above, positions conferring resistance to chemotherapy were well conserved (Table 4.4). This pattern is expected because target sites of TOP2 poisons should disrupt functionally relevant protein sites, which therefore are under strong negative selection. Nevertheless, such sites can vary in cases of resistance to TOP2 poisons and still allow functional topoisomerases, perhaps only possible in the specific environment of cancer cells under different selective pressures.

Table 4.4. TOPIIA amino acid replacements known to affect the efficiency of anticancer drugs. The six positions (out of the 27) that represent different amino acids in TOP2A and TOP2B are identified by an asterisk.

TOP2A				TOP2B				TOPIIA
Human reference			Chordata (*n* = 105)	Human reference			Chordata (*n* = 125)	Metazoa (*n* = 389)
Position	**Amino acid**	**Protein domain**	**Pairwise Identity (%)**	**Position**	**Amino acid**	**Protein domain**	**Pairwise Identity (%)**	**Pairwise Identity (%)**
449	G	TOPRIM	50.3	470	G	TOPRIM	95.2	38.7
450*	R		51.2	471	K		89.2	46.9
480*	K		61	501	R		95.2	32.6
485	P		98.1	506	P		95.2	98
486	L		98.1	507	L		98.4	99
487	R		98.1	508	R		95.2	86.5
489	K		98.1	510	K		96.8	97.5
491	L		98.1	512	L		96.8	96.4
494	R		98.1	515	R		95.2	95.9
498	H		90.8	519	H		96.8	70.7
506	E		100	527	E		96.8	97.5
534	G		100	555	G		96.8	97.5
645	D		98.1	666	D		98.4	96.9
652	A		100	673	A		100	100
716	P		96.2	737	P		100	98.5
760	G	WHD	98.1	781	G	WHD	100	98.5
761	E		98.1	782	E		100	99
762*	M		41.2	783	Q		100	27.3
763*	S		87.4	784	A		98.4	54.1
766	M		96.2	787	M		100	48.6
798	K		96.2	819	K		100	98
803	P		96.2	824	P		100	75.4
805	Y		96.2	826	Y		100	98
846	P	Tower	100	867	P	Tower	100	99
906	Q		92.5	927	Q		98.4	45
908*	V		59.7	929	A		40	35
909*	I		67.6	930	V		96.8	27.4

4.5. Conclusions

Our results suggest that the long-term evolution of TOPIIA is primarily driven by strong purifying selection, which also explains the high levels of sequence conservation. *TOP2B* is under stronger selective constraints than *TOP2A*, which may be explained by the specialized role of *TOP2B* in the genetic programming of postmitotic cells that impose additional constrains to its evolution. The TOPIIA phylogeny lead us to conclude that Cyclostomata TOPIIA paralogues have evolved independently from jawed vertebrates. Therefore, jawless vertebrates are a good model to uncover the role of additional TOPIIA genes in vertebrate evolution. Our study also identified two missense mutations in the *TOP2A* CTD when comparing modern and archaic humans, that may have contributed to the evolution of human-specific features. We found that almost all mutations related with resistance to chemotherapy or causing diseases occur in conserved sites of the *TOP2B* ATPase and DNA Binding/Cleavage domains, including their linker region. Therefore, we recommend that these domains should be included in the screening of undiagnosed diseases, particularly considering the multiple roles of TOP2B in cells. Similarly, we provide a list of residues that could be a good target to design TOP2A-specific anticancer drugs that would avoid the undesirable side effects caused by interfering with TOP2B. Overall, our study provides important insights into the evolution of TOPIIA in animals and represents a valuable resource for future functional studies of topoisomerases.

Acknowledgements

This work was partially supported by a research grant to FM (SFRH/BD/131584/2017) from Fundação para a Ciência e a Tecnologia. MA was supported by the Spanish Ministerio de Ciencia e Innovación through the Grants [RYC-2015-18241] and [PID2019-107931GA-I00].

4.6. References

Aase-Remedios ME, Ferrier DEK (2021) Improved understanding of the role of gene and genome duplications in chordate evolution with new genome and transcriptome sequences. Frontiers in Ecology and Evolution 9:429.

Akimitsu N et al. (2003) Enforced cytokinesis without complete nuclear division in embryonic cells depleting the activity of DNA topoisomerase IIα. Genes to Cells 8:393-402.

Ali Y, Abd Hamid S (2016) Human topoisomerase II alpha as a prognostic biomarker in cancer chemotherapy. Tumor Biology 37:47-55.

Anisimova M, Gil M, Dufayard J-F, Dessimoz C, Gascuel O (2011) Survey of branch support methods demonstrates accuracy, power, and robustness of fast likelihood-based approximation schemes. Systematic biology 60:685-699.

Aravind L, Leipe DD, Koonin EV (1998) Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Nucleic acids research 26:4205-4213.

Austin CA, Fisher LM (1990) Isolation and characterization of a human cDNA clone encoding a novel DNA topoisomerase II homologue from HeLa cells. FEBS letters 266:115-117.

Austin CA et al. (2018) TOP2B: the first thirty years. International Journal of Molecular Sciences 19:2765.

Austin CA, Sng J-H, Patel S, Fisher LM (1993) Novel HeLa topoisomerase II is the IIβ isoform: complete coding sequence and homology with other type II topoisomerases. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1172:283-291.

Azarova AM, Lyu YL, Lin C-P, Tsai Y-C, Lau JY-N, Wang JC, Liu LF (2007) Roles of DNA topoisomerase II isozymes in chemotherapy and secondary malignancies. Proceedings of the National Academy of Sciences 104:11014-11019.

Beck WT, Danks MK, Wolverton JS, Kim R, Chen M (1993) Drug resistance associated with altered DNA topoisomerase II. Advances in enzyme regulation 33:113-116.

Berger JM, Gamblin SJ, Harrison SC, Wang JC (1996) Structure and mechanism of DNA topoisomerase II. Nature 379:225-232.

Berman HM et al. (2000) The protein data bank. Nucleic acids research 28:235-242.

Berna L, Alvarez-Valin F (2014) Evolutionary genomics of fast evolving tunicates. Genome biology and evolution 6:1724-1738.

Bollimpelli VS, Dholaniya PS, Kondapi AK (2017) Topoisomerase IIβ and its role in different biological contexts. Archives of Biochemistry and Biophysics 633:78-84.

Broderick L et al. (2019) Mutations in topoisomerase IIβ result in a B cell immunodeficiency. Nature communications 10:1-15.

Broeck AV, Lotz C, Drillien R, Haas L, Bedez C, Lamour V (2021) Structural basis for allosteric regulation of Human Topoisomerase IIα. Nature Communications 12:1-13.

Brown CJ, Johnson AK, Dunker AK, Daughdrill GW (2011) Evolution and disorder. Current opinion in structural biology 21:441-446.

Cabral JHM, Jackson AP, Smith CV, Shikotra N, Maxwell A, Liddington RC (1997) Crystal structure of the breakage–reunion domain of DNA gyrase. Nature 388:903-906.

Campbell LI et al. (2011) MicroRNAs and phylogenomics resolve the relationships of Tardigrada and suggest that velvet worms are the sister group of Arthropoda. Proceedings of the National Academy of Sciences 108:15920-15924.

Chakraborty AK, Majumder HK (1987) Decatenation of kinetoplast DNA by an ATP-dependent DNA topoisomerase from the kinetoplast hemoflagellate Leishmania donovani. Molecular and biochemical parasitology 26:215-224.

Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. Annual review of biochemistry 70:369-413.

Cheesman S, McAleese S, Goman M, Johnson D, Horrocks P, Ridley RG, Kilbey BJ (1994) The gene encoding topoisomerase II from Plasmodium falciparum. Nucleic acids research 22:2547-2551.

Chen W, Qiu J, Shen Y (2012) Topoisomerase IIα, rather than IIβ, is a promising target in development of anti-cancer drugs. Drug discoveries & therapeutics 6:230-237.

Classen S, Olland S, Berger JM (2003) Structure of the topoisomerase II ATPase region and its mechanism of inhibition by the chemotherapeutic agent ICRF-187. Proceedings of the National Academy of Sciences 100:10629-10634.

Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. Genome research 13:2052-2058.
Corbett KD, Berger JM (2003) Structure of the topoisomerase VI-B subunit: implications for type II topoisomerase mechanism and evolution. The EMBO Journal 22:151-163.
Corbett KD, Schoeffler AJ, Thomsen ND, Berger JM (2005) The structural basis for substrate specificity in DNA topoisomerase IV. Journal of molecular biology 351:545-561.
Cornarotti M, Tinelli S, Willmore E, Zunino F, Fisher LM, Austin CA, Capranico G (1996) Drug sensitivity and sequence specificity of human recombinant DNA topoisomerases IIalpha (p170) and IIbeta (p180). Molecular Pharmacology 50:1463-1471.
Cowell IG et al. (2012) Model for MLL translocations in therapy-related leukemia involving topoisomerase IIβ-mediated DNA strand breaks and gene proximity. Proceedings of the National Academy of Sciences 109:8989-8994.
Cuvier O, Hirano T (2003) A role of topoisomerase II in linking DNA replication to chromosome condensation. The Journal of cell biology 160:645-655.
Darriba D, Taboada GL, Doallo R, Posada D (2012) jModelTest 2: more models, new heuristics and parallel computing. Nature methods 9:772-772.
De H, Wen JF, Chen WQ, Lu SQ, De Dong X (2005) Identification, characteristic and phylogenetic analysis of type II DNA topoisomerase gene in Giardia lamblia. Cell research 15:474-482.
Del Amparo R, Branco C, Arenas J, Vicens A, Arenas M (2021) Analysis of selection in protein-coding sequences accounting for common biases. Briefings in Bioinformatics.
Delgado JL, Hsieh C-M, Chan N-L, Hiasa H (2018) Topoisomerases as anticancer targets. Biochemical Journal 475:373-398.
Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965-968.
Denoeud F et al. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science 330:1381-1385.
Dong KC, Berger JM (2007) Structural basis for gate-DNA recognition and bending by type IIA topoisomerases. Nature 450:1201-1205.
Drake FH, Hofmann GA, Bartus HF, Mattern MR, Crooke ST, Mirabelli CK (1989) Biochemical and pharmacological properties of p170 and p180 forms of topoisomerase II. Biochemistry 28:8154-8160.
Drake FH et al. (1987) Purification of topoisomerase II from amsacrine-resistant P388 leukemia cells. Evidence for two forms of the enzyme. Journal of Biological Chemistry 262:16739-16747.
Dunn CW, Giribet G, Edgecombe GD, Hejnol A (2014) Animal phylogeny and its evolutionary implications. Annual review of ecology, evolution, and systematics 45:371-395.
Dunn CW et al. (2008) Broad phylogenomic sampling improves resolution of the animal tree of life. Nature 452:745-749.
Dutta R, Inouye M (2000) GHKL, an emergent ATPase/kinase superfamily. Trends in biochemical sciences 25:24-28.
Erdős M, Lányi Á, Balázs G, Casanova J-L, Boisson B, Maródi L (2021) Inherited TOP2B Mutation: Possible Confirmation of Mutational Hotspots in the TOPRIM Domain. Journal of clinical immunology 41:817-819.
Escriva H, Manzon L, Youson J, Laudet V (2002) Analysis of lamprey and hagfish genes reveals a complex history of gene duplications during early vertebrate evolution. Molecular biology and evolution 19:1440-1450.
Fass D, Bogden CE, Berger JM (1999) Quaternary changes in topoisomerase II may direct orthogonal movement of two DNA strands. Nature structural biology 6:322-326.
Forterre P, Gribaldo S, Gadelle D, Serre M-C (2007) Origin and evolution of DNA topoisomerases. Biochimie 89:427-446.
Gadelle D, Filée J, Buhler C, Forterre P (2003) Phylogenomics of type II DNA topoisomerases. Bioessays 25:232-242.
Gajiwala KS, Burley SK (2000) Winged helix proteins. Current opinion in structural biology 10:110-116.
Gilroy KL, Leontiou C, Padget K, Lakey JH, Austin CA (2006) mAMSA resistant human topoisomerase IIβ mutation G465D has reduced ATP hydrolysis activity. Nucleic acids research 34:1597-1607.
Glessner JT et al. (2014) Increased frequency of de novo copy number variants in congenital heart disease by integrative analysis of single nucleotide polymorphism array and exome sequence data. Circulation research 115:884-896.

Green RE et al. (2010) A draft sequence of the Neandertal genome. science 328:710-722.
Grue P et al. (1998) Essential mitotic functions of DNA topoisomerase IIα are not adopted by topoisomerase IIβ in human H69 cells. Journal of Biological Chemistry 273:33660-33666.
Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O (2010) New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of PhyML 3.0. Systematic Biology 59:307-321.
Haffner MC et al. (2010) Androgen-induced TOP2B-mediated double-strand breaks and prostate cancer gene rearrangements. Nature Genetics 42:668-675.
Hunt SE et al. (2018) Ensembl variation resources. Database 2018:bay 119.
Ishida R et al. (2001) Mitotic specific phosphorylation of serine-1212 in human DNA topoisomerase IIα. Cell structure and function 26:215-226.
Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In: Parasite Genomics Protocols. Springer, pp 65-90
Jenkins JR et al. (1992) Isolation of cDNA clones encoding the β isozyme of human DNA topoisomerase II and localisation of the gene to chromosome 3p24. Nucleic acids research 20:5587-5592.
Ju B-G, Lunyak VV, Perissi V, Garcia-Bassets I, Rose DW, Glass CK, Rosenfeld MG (2006) A topoisomerase IIß-mediated dsDNA break required for regulated transcription. science 312:1798-1802.
Katoh K, Rozewicki J, Yamada KD (2019) MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Briefings in bioinformatics 20:1160-1166.
Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome research 12:996-1006.
Kingma P, Osheroff N, Bjergbaek L, Nielsen IS, Wang Y, Westergaard O, Andersen AH (2000) Communication between the ATPase and cleavage/religation domains of human topoisomerase IIα. Journal of Biological Chemistry 275:13041-13048.
Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome biology 3:1-9.
Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular biology and evolution 22:1208-1222.
Kosakovsky Pond SL et al. (2020) HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. Molecular biology and evolution 37:295-299.
Kozuki T et al. (2017) Roles of the C-terminal domains of topoisomerase IIα and topoisomerase IIβ in regulation of the decatenation checkpoint. Nucleic acids research 45:5995-6010.
Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM (2019) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic acids research 47:D807-D811.
Kuhlwilm M, Boeckx C (2019) A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. Scientific reports 9:1-14.
Lam C-W, Yeung W-L, Law C-Y (2017) Global developmental delay and intellectual disability associated with a de novo TOP2B mutation. Clinica Chimica Acta 469:63-68.
Lane AB, Giménez-Abián JF, Clarke DJ (2013) A novel chromatin tether domain controls topoisomerase IIα dynamics and mitotic chromosome formation. Journal of Cell Biology 203:471-486.
Laponogov I et al. (2007) Breakage-reunion domain of Streptococcus pneumoniae topoisomerase IV: crystal structure of a gram-positive quinolone target. PLoS One 2:e301.
Lefort V, Longueville J-E, Gascuel O (2017) SMS: smart model selection in PhyML. Molecular biology and evolution 34:2422-2424.
Leontiou C, Lakey JH, Lightowlers R, Turnbull RM, Austin CA (2006) Mutation P732L in human DNA topoisomerase IIβ abolishes DNA cleavage in the presence of calcium and confers drug resistance. Molecular pharmacology 69:130-139.
Leontiou C, Lightowlers R, Lakey JH, Austin CA (2003) Kinetic analysis of human topoisomerase IIα and β DNA binding by surface plasmon resonance. FEBS letters 554:206-210.
Leontiou C et al. (2007) Differential selection of acridine resistance mutations in human DNA topoisomerase IIβ is dependent on the acridine structure. Molecular pharmacology 71:1006-1014.

Linka RM, Porter AC, Volkov A, Mielke C, Boege F, Christensen MO (2007) C-terminal regions of topoisomerase II α and II β determine isoform-specific functioning of the enzymes in vivo. Nucleic acids research 35:3810-3822.

Liu LF, Rowe TC, Yang L, Tewey KM, Chen GL (1983) Cleavage of DNA by mammalian DNA topoisomerase II. Journal of Biological Chemistry 258:15365-15370.

Lotz C, Lamour V (2020) The interplay between DNA topoisomerase 2α post-translational modifications and drug resistance. Cancer Drug Resistance:149-160.

Lyu YL, Lin C-P, Azarova AM, Cai L, Wang JC, Liu LF (2006) Role of Topoisomerase IIβ in the Expression of Developmentally RegulatedGenes. Molecular and cellular biology 26:7929-7941.

Lyu YL, Wang JC (2003) Aberrant lamination in the cerebral cortex of mouse embryos lacking DNA topoisomerase IIβ. Proceedings of the National Academy of Sciences 100:7123-7128.

Marsh KL et al. (1996) Amsacrine-promoted DNA cleavage site determinants for the two human DNA topoisomerase II isoforms α and β. Biochemical pharmacology 52:1675-1685.

McClendon AK, Rodriguez AC, Osheroff N (2005) Human topoisomerase IIα rapidly relaxes positively supercoiled DNA: implications for enzyme action ahead of replication forks. Journal of Biological Chemistry 280:39337-39345.

McKay DB, Steitz TA (1981) Structure of catabolite gene activator protein at 2.9 Å resolution suggests binding to left-handed B-DNA. Nature 290:744-749.

McNamara S, Wang H, Hanna N, Miller Jr WH (2008) Topoisomerase IIβ negatively modulates retinoic acid receptor α function: a novel mechanism of retinoic acid resistance. Molecular and cellular biology 28:2066-2077.

Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Human molecular genetics 10:2319-2328.

Mirski SE, Gerlach JH, Cummings HJ, Zirngibl R, Greer PA, Cole SP (1997) Bipartite nuclear localization signals in the C terminus of human topoisomerase IIα. Experimental cell research 237:452-455.

Moreira F, Arenas M, Videira A, Pereira F (2021) Molecular Evolution of DNA Topoisomerase III Beta (TOP3B) in Metazoa. Journal of molecular evolution 89:384-395.

Morrison A, Cozzarelli NR (1979) Site-specific cleavage of DNA by E. coli DNA gyrase. Cell 17:175-184.

Nakatani Y, Shingate P, Ravi V, Pillai NE, Prasad A, McLysaght A, Venkatesh B (2021) Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. Nature communications 12:1-14.

Nesnidal MP et al. (2013) New phylogenomic data support the monophyly of Lophophorata and an Ectoproct-Phoronid clade and indicate that Polyzoa and Kryptrochozoa are caused by systematic bias. BMC evolutionary biology 13:1-13.

Niimi A, Suka N, Harata M, Kikuchi A, Mizuno S (2001) Co-localization of chicken DNA topoisomerase IIα, but not β, with sites of DNA replication and possible involvement of a C-terminal region of α through its binding to PCNA. Chromosoma 110:102-114.

Nitiss J, Beck W (1996) Antitopoisomerase drug action and resistance. European Journal of Cancer 32:958-966.

Nitiss JL (2009) Targeting DNA topoisomerase II in cancer chemotherapy. Nature Reviews Cancer 9:338-350.

Ohno S (2013) Evolution by gene duplication. Springer Science & Business Media,

Papapietro O et al. (2020) Topoisomerase 2β mutation impairs early B-cell development. Blood, The Journal of the American Society of Hematology 135:1497-1501.

Pisani D, Carton R, Campbell LI, Akanni WA, Mulville E, Rota-Stabelli O (2013) An overview of arthropod genomics, mitogenomics, and the evolutionary origins of the arthropod proteome. Arthropod biology and evolution:41-61.

Pommier Y, Sun Y, Shar-yin NH, Nitiss JL (2016) Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nature reviews Molecular cell biology 17:703-721.

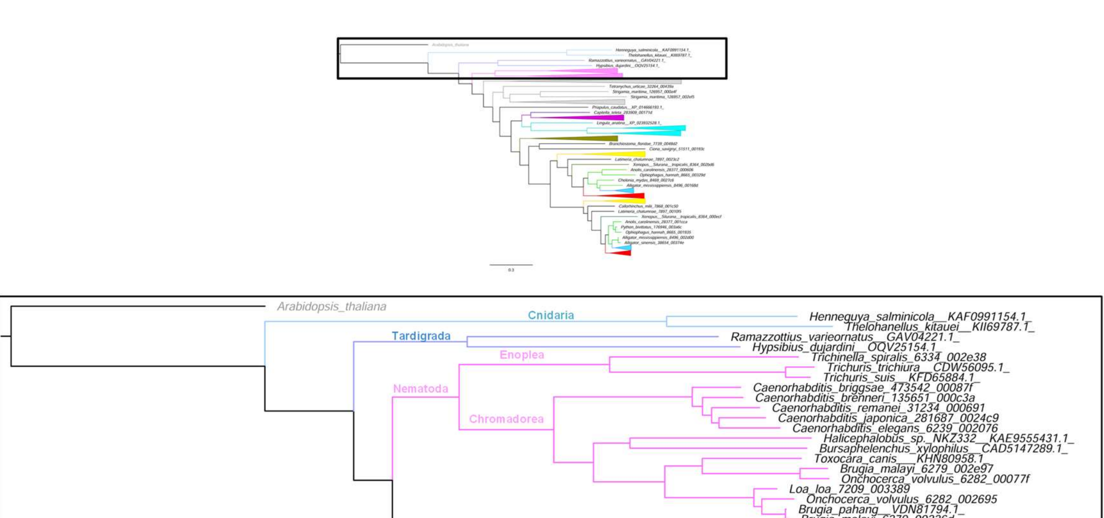
Putnam NH et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064-1071.

Reich D et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053-1060.

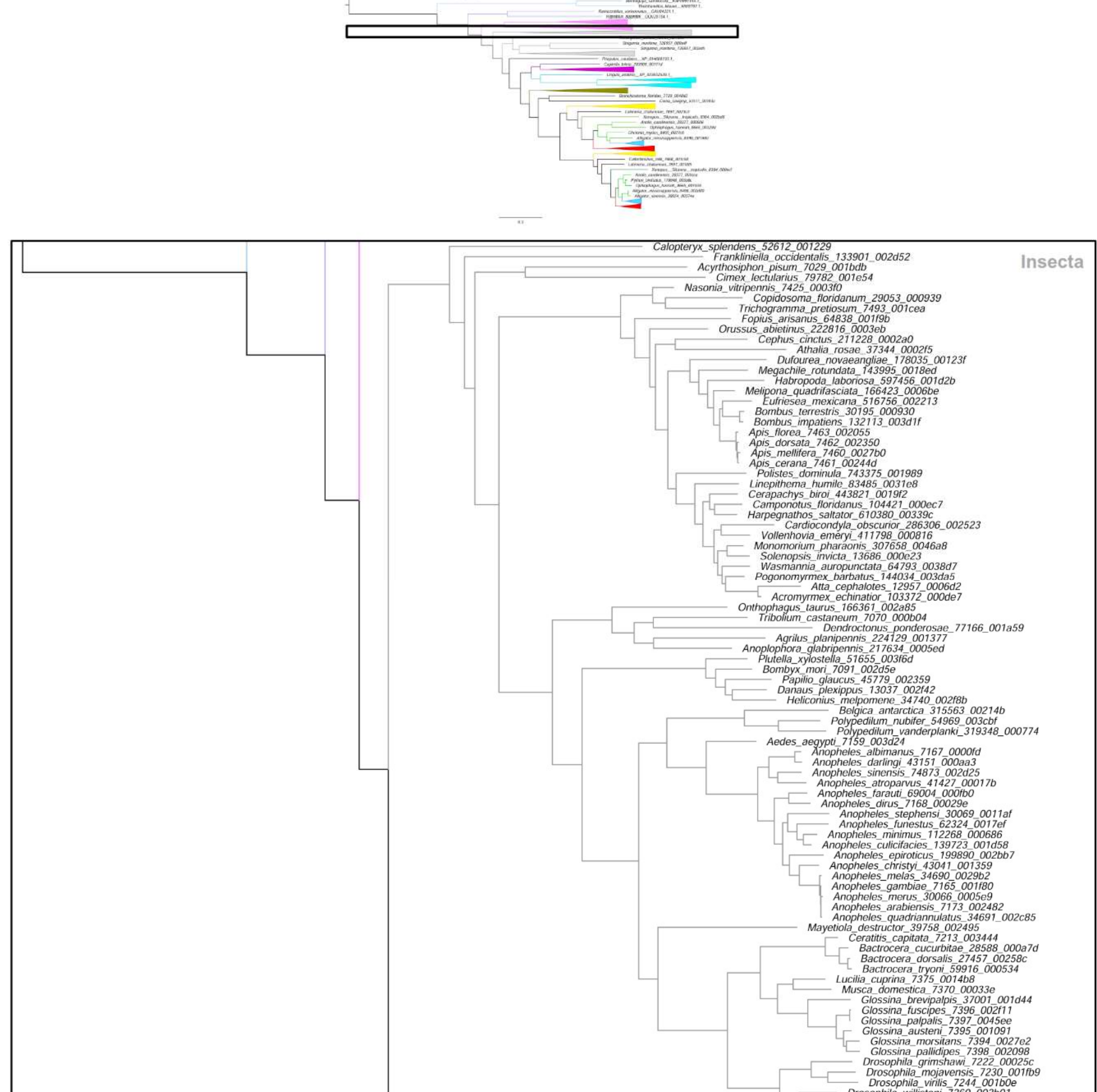
Roca J, Berger JM, Harrison SC, Wang JC (1996) DNA transport by a type II topoisomerase: direct evidence for a two-gate mechanism. Proceedings of the National Academy of Sciences 93:4057-4062.
Roca J, Wang JC (1992) The capture of a DNA double helix by an ATP-dependent protein clamp: a key step in DNA transport by type II DNA topoisomerases. Cell 71:833-840.
Roca J, Wang JC (1994) DNA transport by a type II DNA topoisomerase: evidence in favor of a two-gate mechanism. Cell 77:609-616.
Sander M, Hsieh T-S (1983) Double strand DNA cleavage by type II DNA topoisomerase from Drosophila melanogaster. Journal of Biological Chemistry 258:8421-8428.
Satoh N, Rokhsar D, Nishikawa T (2014) Chordate evolution and the three-phylum system. Proceedings of the Royal Society B: Biological Sciences 281:20141729.
Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome research 18:137-147.
Schoeffler AJ, Berger JM (2008) DNA topoisomerases: harnessing and constraining energy to govern chromosome topology. Quarterly reviews of biophysics 41:41-101.
Sehnal D et al. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Research.
Simakov O et al. (2020) Deeply conserved synteny resolves early events in vertebrate evolution. Nature ecology & evolution 4:820-830.
Smith JJ, Keinath MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. Genome research 25:1081-1090.
Stadler PF, Fried C, Prohaska SJ, Bailey WJ, Misof BY, Ruddle FH, Wagner GP (2004) Evidence for independent Hox gene duplications in the hagfish lineage: a PCR-based gene inventory of *Eptatretus stoutii*. Molecular phylogenetics and evolution 32:686-694.
Strauss PR, Wang JC (1990) The TOP2 gene of Trypanosoma brucei: a single-copy gene that shares extensive homology with other TOP2 genes encoding eukaryotic DNA topoisomerase II. Molecular and biochemical parasitology 38:141-150.
Sun L-Y, Zhu L-W, Tang Y-J (2018) Increasing the distance between two monomers of topoisomerase IIβ under the action of antitumor agent 4 β-sulfur-(benzimidazole) 4′-demethylepipodophyllotoxin. Scientific reports 8:1-9.
Tan K, Dorman TE, Falls KM, Chung TD, Mirabelli CK, Crooke ST, Mao J-i (1992) Topoisomerase IIα and topoisomerase IIβ genes: characterization and mapping to human chromosomes 17 and 3, respectively. Cancer Research 52:231-234.
Tiwari VK et al. (2012) Target genes of Topoisomerase IIβ regulate neuronal survival and are defined by their chromatin state. Proceedings of the National Academy of Sciences 109:E934-E943.
Tsagkogeorga G, Turon X, Galtier N, Douzery EJ, Delsuc F (2010) Accelerated evolutionary rate of housekeeping genes in tunicates. Journal of molecular evolution 71:153-167.
Tsai-Pflugfelder M et al. (1988) Cloning and sequencing of cDNA encoding human DNA topoisomerase II and localization of the gene to chromosome region 17q21-22. Proceedings of the National Academy of Sciences 85:7177-7181.
Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nature Reviews Genetics 10:725-732.
Van de Peer Y, Taylor JS, Braasch I, Meyer A (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. Journal of molecular evolution 53:436-446.
Vassetzky YS, Alghisi GC, Gasser SM (1995) DNA topoisomerase II mutations and resistance to anti-tumor drugs. Bioessays 17:767-774.
Wang JC (1996) DNA topoisomerases. Annual review of biochemistry 65:635-692.
Wang JC (2002) Cellular roles of DNA topoisomerases: a molecular perspective. Nature reviews Molecular cell biology 3:430-440.
Watterson G (1975) On the number of segregating sites in genetical models without recombination. Theoretical population biology 7:256-276.
Wigley DB, Davies GJ, Dodson EJ, Maxwell A, Dodson G (1991) Crystal structure of an N-terminal fragment of the DNA gyrase B protein. Nature 351:624-629.
Wu C-C et al. (2011) Structural basis of type II topoisomerase inhibition by the anticancer drug etoposide. Science 333:459-462.
Wu C-C, Li Y-C, Wang Y-R, Li T-K, Chan N-L (2013) On the structural basis and design guidelines for type II topoisomerase-targeting anticancer drugs. Nucleic acids research 41:10630-10640.

Xia W et al. (2019) Mutations in TOP 2B cause autosomal-dominant hereditary hearing loss via inhibition of the PI 3K-Akt signalling pathway. FEBS letters 593:2008-2018.
Yang X, Li W, Prescott ED, Burden SJ, Wang JC (2000) DNA topoisomerase IIβ and neural development. Science 287:131-134.
Ye J et al. (2010) TRF2 and Apollo cooperate with topoisomerase 2α to protect human telomeres from replicative damage. Cell 142:230-242.

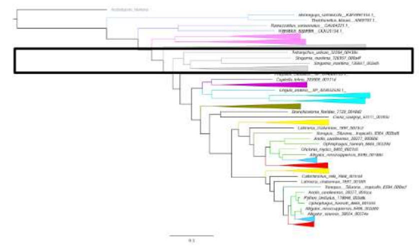
4.7. Supplementary Material

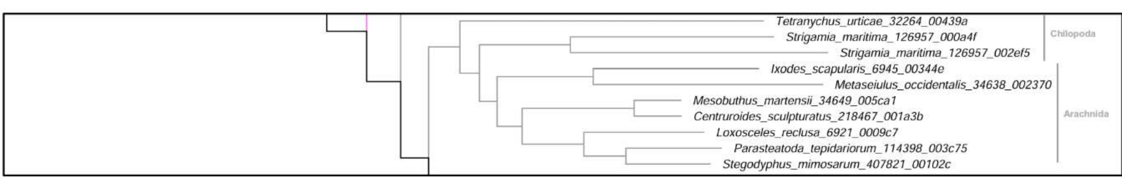


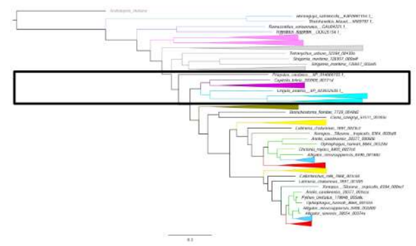
Supplementary Figure S4.1. Detailed view of the maximum likelihood (ML) phylogenetic tree built with an alignment of 389 TOPIIA protein sequences from metazoans, considering Arabidopsis thaliana as outgroup.

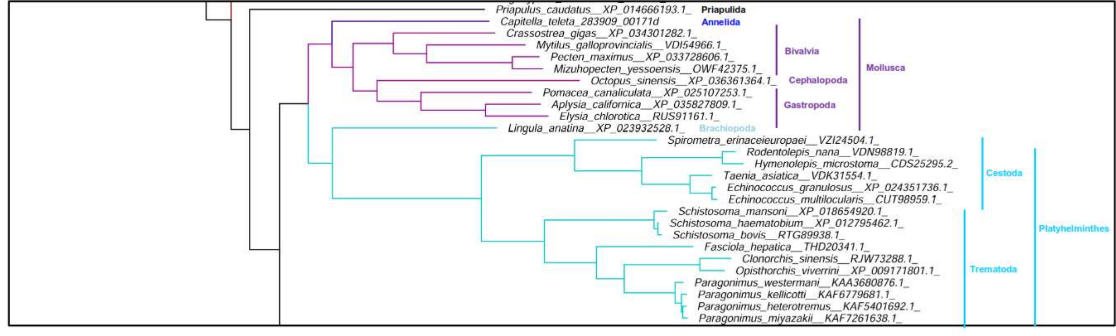


Supplementary Figure S4.1 (cont.)

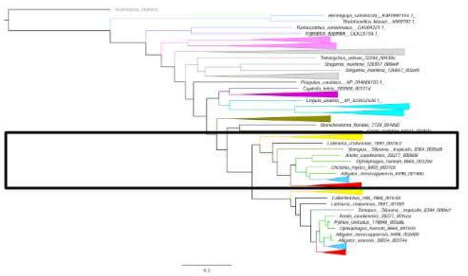




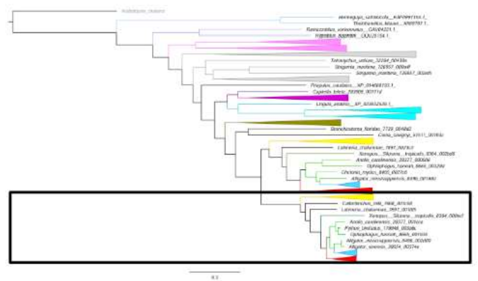




Supplementary Figure S4.1 (cont.)



Supplementary Figure S4.1 (cont.)



TOP2B

Lepisosteus_oculatus_7918_001af2
Danio_rerio_7955_00163c
Esox_lucius_8010_0017d6
Cynoglossus_semilaevis_244447_004025
Takifugu_rubripes_31033_001b48
Tetraodon_nigroviridis_99883_00460b
Stegastes_partitus_144197_002158
Larimichthys_crocea_215358_0048ee
Gasterosteus_aculeatus_69293_001efd
Oreochromis_niloticus_8128_002849
Pundamilia_nyererei_303518_0038e8
Maylandia_zebra_106582_000ca3
Haplochromis_burtoni_8153_003bcd
Oryzias_latipes_8090_002021
Fundulus_heteroclitus_8078_005015
Poecilia_formosa_48698_0019d1
Xiphophorus_maculatus_8083_00420e

Actinopteri

Callorhinchus_milii_7868_001c50
Chondrichthyes

Latimeria_chalumnae_7897_0010f5
Coelacanthimorpha

Xenopus__Silurana__tropicalis_8364_000ecf
Anolis_carolinensis_28377_001cca
Python_bivittatus_176946_003a6c
Ophiophagus_hannah_8665_001935

Lepidosauria

Alligator_mississippiensis_8496_002d00
Alligator_sinensis_38654_00374e

Archelosauria

Reptilia

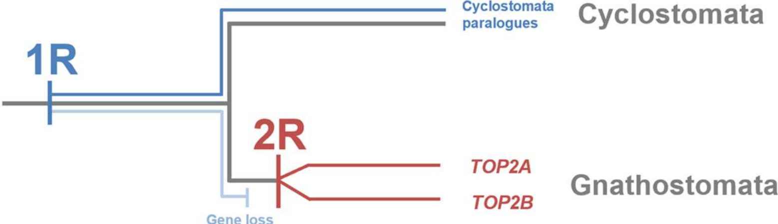
Tinamus_guttatus_94827_000891
Struthio_camelus_8801_00359b
Anas_platyrhynchos_8839_00182d
Meleagris_gallopavo_9103_0022fb
Gallus_gallus_9031_001e09
Buceros_rhinoceros_175835_002780
Charadrius_vociferus_50402_001433
Ficedula_albicollis_59894_0023bf
Pseudopodoces_humilis_181119_002c2a
Manacus_vitellinus_328815_001775
Corvus_cornix_181096_001528
Taeniopygia_guttata_59729_000d06
Serinus_canaria_9135_00029f
Geospiza_fortis_48883_001707
Zonotrichia_albicollis_44394_001790
Haliaeetus_leucocephalus_52644_0036e5
Aquila_chrysaetos_8962_0027ad
Chaetura_pelagica_8897_0006df
Picoides_pubescens_118200_0027a0
Gavia_stellata_37040_001361
Nipponia_nippon_128390_001d52
Egretta_garzetta_188379_001030
Phaethon_lepturus_97097_002998
Eurypyga_helias_54383_0027b5
Cariama_cristata_54380_001536
Falco_cherrug_345164_0035dd
Falco_peregrinus_8954_000a8c
Aptenodytes_forsteri_9233_0034ff
Pygoscelis_adeliae_9238_0019ca
Nestor_notabilis_176057_001e94
Opisthocomus_hoazin_30419_0023f7
Columba_livia_8932_001947

Aves

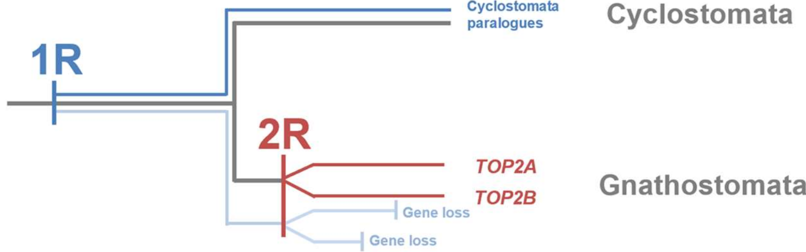
Ornithorhynchus_anatinus_9258_002d24
Monodelphis_domestica_13616_002bbf
Sarcophilus_harrisii_9305_0029cc
Sus_scrofa_9823_001f66
Balaenoptera_acutorostrata_9767_000e32
Physeter_catodon_9755_0024bc
Tursiops_truncatus_9739_00285b
Lipotes_vexillifer_118797_002760
Camelus_dromedarius_9838_00254c
Camelus_ferus_419612_003a5d
Camelus_bactrianus_9837_0005a8
Ovis_aries_9940_00105b
Capra_hircus_9925_002738
Pantholops_hodgsonii_59538_003d97
Bubalus_bubalis_89462_00186c
Bison_bison_9901_002a88
Bos_taurus_9913_000bad
Pteropus_alecto_9402_0045da
Eptesicus_fuscus_29078_002c49
Myotis_lucifugus_59463_00056e
Myotis_brandtii_109478_000116
Myotis_davidii_225400_003643
Condylura_cristata_143302_0023b2
Equus_przewalskii_9798_001c13
Equus_caballus_9796_000f21
Panthera_tigris_9694_0023a5
Felis_catus_9685_0049d1
Canis_lupus_familiaris_9615_001088
Ailuropoda_melanoleuca_9646_000251
Ursus_maritimus_29073_0049bc
Mustela_putorius_furo_9669_002e89
Leptonychotes_weddellii_9713_002fe1
Odobenus_rosmarus_9707_002a58
Dasypus_novemcinctus_9361_00111b
Trichechus_manatus_9778_003b2d
Loxodonta_africana_9785_002aa1
Elephantulus_edwardii_28737_000a8b
Chrysochloris_asiatica_185453_0031ac
Oryctolagus_cuniculus_9986_00021f
Tupaia_chinensis_246437_00290b
Propithecus_coquereli_379532_001289
Otolemur_garnettii_30611_000731
Octodon_degus_10160_003087
Fukomys_damarensis_885580_002108
Cavia_porcellus_10141_003237
Chinchilla_lanigera_34839_001ba4
Jaculus_jaculus_51337_0043f6
Nannospalax_galili_1026970_001c1c
Mus_musculus_10090_00070a
Peromyscus_maniculatus_10042_0042ba
Mesocricetus_auratus_10036_003556
Galeopterus_variegatus_482537_002e82
Aotus_nancymaae_37293_0002bf
Callithrix_jacchus_9483_000cc2
Pongo_abelii_9601_0033ce
Nomascus_leucogenys_61853_000d67
Gorilla_gorilla_gorilla_2B__XP_030864923.1_
Homo_sapiens_9606_000552
Pan_paniscus_9597_001682
Pan_troglodytes_2B__XP_009443307.2_
Rhinopithecus_roxellana_61622_000c4c
Colobus_angolensis_54131_003d9b
Chlorocebus_sabaeus_60711_0024dd
Cercocebus_atys_9531_00122c
Mandrillus_leucophaeus_9568_00161a
Papio_anubis_9555_001827
Macaca_mulatta_9544_0016e9
Macaca_nemestrina_9545_00388b

Mammalia

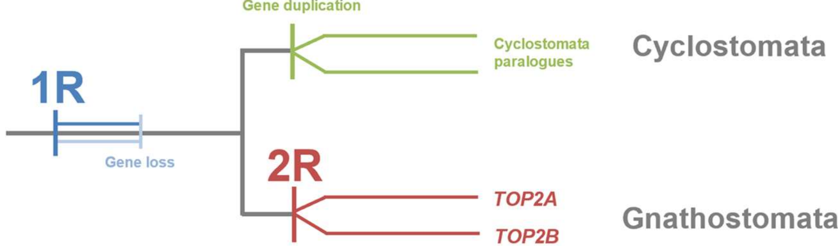
Supplementary Figure S4.1 (cont.)



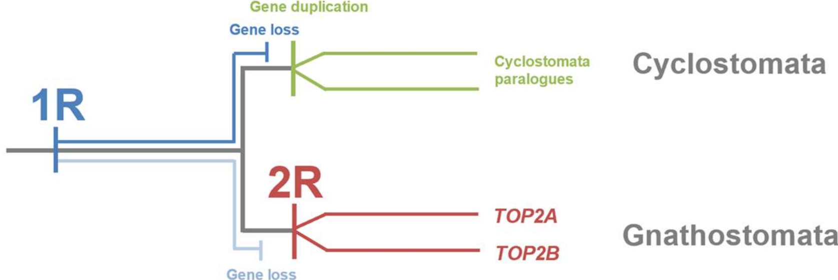
One event: a single gene loss in Gnathostomata lineage before 2R



Two events: Two gene losses in Gnathostomata after 2R



Two events: a gene loss before the Cyclostomata/Gnathostomata divergence and a gene duplication in the Cyclostomata lineage



Three events: two gene losses after the Cyclostomata/Gnathostomata divergence and a gene duplication in the Cyclostomata lineage

Supplementary Figure S4.2. Different scenarios to explain the observed placement of Cyclostomata and Gnathostomata TOPIIA paralogues in the metazoan phylogeny. The models assume the occurrence of two rounds of tetraploidization, 1R before the Cyclostomata / Gnathostomata divergence and 2R in the Gnathostomata lineage, that duplicated all TOPIIA genes found in the genome at that time.

Chapter 5
Publication nº3

Journal of Molecular Evolution
https://doi.org/10.1007/s00239-022-10091-z

ORIGINAL ARTICLE

Evolution of *TOP1* and *TOP1MT* Topoisomerases in Chordata

Filipa Moreira[1,2] · Miguel Arenas[3,4,5] · Arnaldo Videira[2,6,7] · Filipe Pereira[8,9]

Received: 10 September 2022 / Accepted: 30 December 2022

5. Evolution of *TOP1* and *TOP1MT* topoisomerases in Chordata (Scientific Publication nº 3)

Filipa Moreira[1,2], Miguel Arenas[3,4,5], Arnaldo Videira[2,6,7] and Filipe Pereira[8,9*]

[1] Interdisciplinary Centre of Marine and Environmental Research (CIIMAR), University of Porto, Terminal de Cruzeiros do Porto de Leixões, Av. General Norton de Matos s/n 4450-208, Matosinhos, Portugal
[2] ICBAS - Instituto de Ciências Biomédicas de Abel Salazar, Universidade do Porto, Rua Jorge de Viterbo Ferreira 228, 4050-313, Porto, Portugal
[3] Department of Biochemistry, Genetics and Immunology, University of Vigo, 36310 Vigo, Spain
[4] CINBIO, Universidade de Vigo, 36310 Vigo, Spain
[5] Galicia Sur Health Research Institute (IIS Galicia Sur), 36310 Vigo, Spain
[6] IBMC-Instituto de Biologia Molecular e Celular, Universidade do Porto, Porto, Portugal
[7] i3S-Instituto de Investigação e Inovação em Saúde, Universidade do Porto, Porto, Portugal
[8] IDENTIFICA genetic testing, Rua Simão Bolívar 259 3º Dir Tras. 4470-214 Maia, Portugal
[9] Centre for Functional Ecology, Department of Life Sciences, University of Coimbra, Calçada Martim de Freitas, 3000-456 Coimbra, Portugal

* Corresponding author. E-mail address: fpereirapt@gmail.com (F. Pereira)

5.1. Abstract

Type IB topoisomerases relax the torsional stress associated with DNA metabolism in the nucleus and mitochondria and constitute important molecular targets of anticancer drugs. Vertebrates stand out among eukaryotes by having two Type IB topoisomerases acting specifically in the nucleus (TOP1) and mitochondria (TOP1MT). Despite their major importance, the origin and evolution of these paralogues remain unknown. Here, we examine the molecular evolutionary processes acting on both TOP1 and TOP1MT in Chordata, taking advantage of the increasing number of available genome sequences. We found that both TOP1 and TOP1MT evolved under strong purifying selection, as expected considering their essential biological functions. Critical active sites, including those associated with resistance to anticancer agents, were found particularly conserved. However, TOP1MT presented a higher rate of molecular evolution than TOP1, possibly related with its specialized activity on the mitochondrial genome and a less critical role in cells. We could place the duplication event that originated the *TOP1* and *TOP1MT* paralogues early in the radiation of vertebrates, most likely associated with the first round of vertebrate tetraploidization (1R). Moreover, our data suggest that cyclostomes present a specialized mitochondrial Type IB topoisomerase. Interestingly, we identified two missense mutations replacing amino acids in the Linker region of TOP1MT in Neanderthals, which appears as a rare event when comparing the genome of both species. In conclusion, TOP1 and TOP1MT differ in their rates of evolution, and their evolutionary histories allowed us to better understand the evolution of chordates.

Keywords: Type IB topoisomerases, molecular phylogeny, purifying selection, archaic humans, functional divergence

5.2. Introduction

DNA topoisomerases introduce reversible breaks in the DNA phosphodiester backbone allowing for modifications in DNA topology during DNA replication, recombination, transcription and chromosome condensation (Pommier et al. 2022; Pommier et al. 2016). Concerning Type I topoisomerases, they are monomeric and cleave one DNA strand at a time without requiring an energy cofactor. These topoisomerases are traditionally classified in two groups (Type IA and Type IB) without sequence and structural similarity. Indeed, while Type IA breaks the DNA by forming a covalent bond to the 5′ end, Type IB bind covalently to the 3′ end of the break (Capranico et al. 2017; Cheng et al. 1998; Redinbo et al. 1998).

Type IB topoisomerases were found in some bacteria and Poxviruses and in eukaryotes (Champoux 2001; Forterre et al. 2007). All eukaryotes have at least one topoisomerase I (TOP1) for relaxing both negative and positive supercoils in front of moving polymerases during replication and transcription. Studies in yeast suggest that a single TOP1 may act in both the nuclear and mitochondrial genomes (de la Loza and Wellinger 2009; Wang et al. 1995). However, a second Type IB topoisomerase (TOP1MT) was identified in vertebrates, encoded in the nuclear genome. The TOP1MT exclusively localizes to mitochondria via a mitochondrial targeting sequence (MTS) at its N-terminal domain (Zhang et al. 2001). Among model organisms, TOP1 is essential for mouse and fruit fly development (Lee et al. 1993; Morham et al. 1996). TOP1MT seems to be dispensable for mouse development, but its absence causes increased negative supercoiling of mitochondrial DNA (mtDNA) and affects cellular energy metabolism (Douarre et al. 2012; Zhang et al. 2014) by interfering with biological processes such as liver regeneration (Khiati et al. 2015). Despite the biological relevance of both genes, their origin and molecular evolutionary patterns are still unknown.

In humans, the *TOP1* gene is located in the chromosome region 20q12 (Juan et al. 1988) and encodes a 91 kDa protein with 765 amino acids. Two *TOP1* pseudogenes have been identified on chromosomes 1 (ψ1-h*TOP1*) and 22 (ψ2-h*TOP1*) resulting from truncated mRNA transcripts of the active gene (Fig. 1A) (Yang et al. 1990). The *TOP1MT* gene maps to chromosome region 8q24 resulting in a 70 kDa protein with 601 amino acids (Zhang et al. 2001). Although *TOP1* has 21 exons and *TOP1MT* has 14 exons, the terminal 13 exons are conserved between both genes (Zhang et al. 2004).

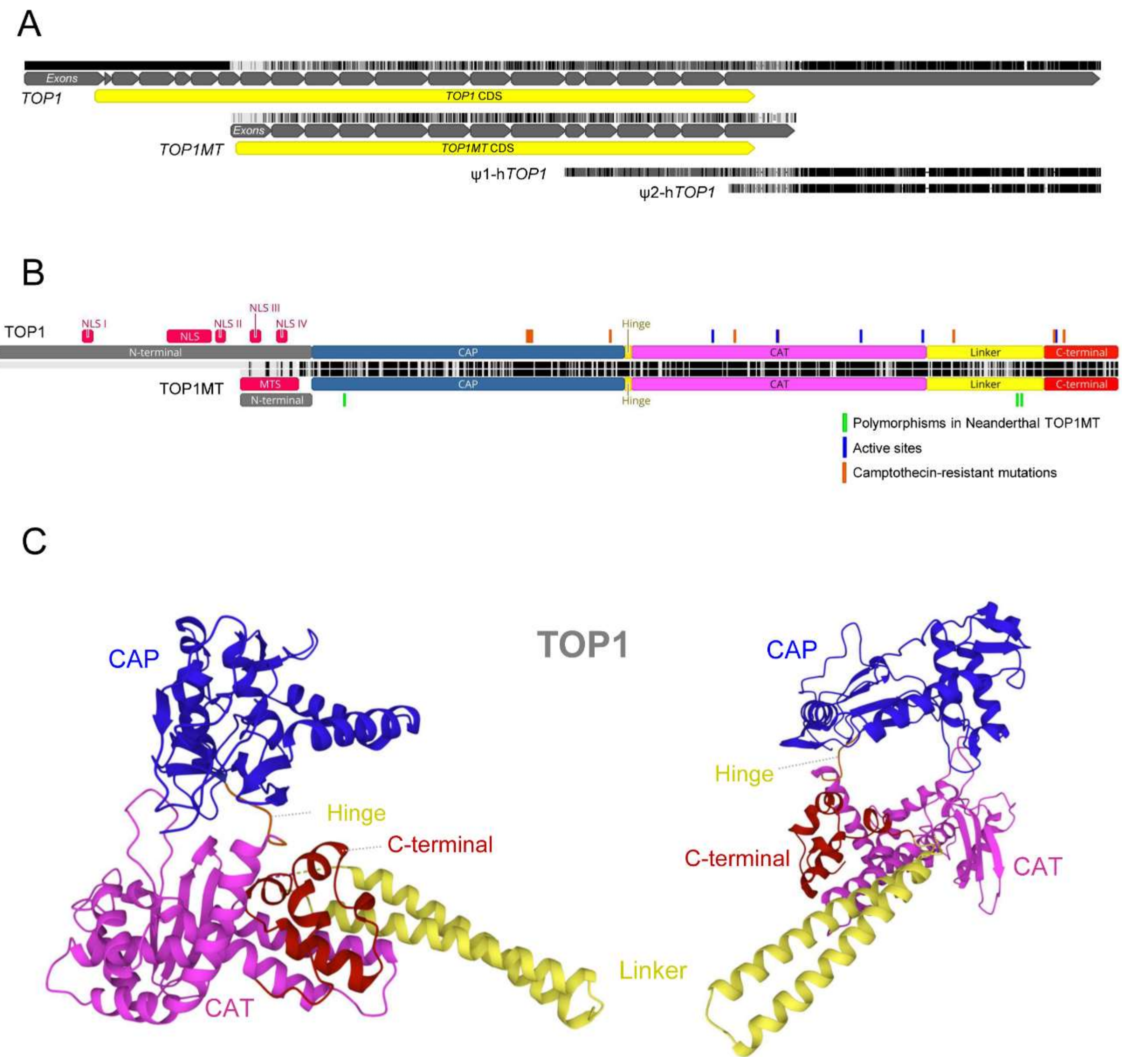


Figure 5.1. Organization of human nuclear (TOP1) and mitochondrial (TOP1MT) DNA topoisomerases I. A) Multiple sequence alignment of human *TOP1* and *TOP1MT* mRNA sequences and the two *TOP1* pseudogenes identified in chromosomes 1 (ψ1-h*TOP1*) and 22 (ψ2-h*TOP1*). B) Pairwise alignment of TOP1 and TOP1MT protein sequences, annotated with the most relevant protein domains and sites. C) Illustrative representation of the human TOP1 protein structure with major domains highlighted.

Considering the molecular structure and sequence conservation, TOP1 and TOP1MT proteins are organized into four distinct domains: N-terminal, Core, Linker and C-terminal domains (Fig. 5.1B, C). The N-terminal domain is poorly conserved across species and varies considerably when comparing both proteins. In particular, the TOP1 N-terminal is highly charged and relatively unstructured, being dispensable for the enzyme activity, mediates protein-protein interactions and includes nuclear localization signals (NLSs) (Alsner et al. 1992; Mo et al. 2000; Palle et al. 2008). The TOP1MT N-terminal is much shorter than that from TOP1 and includes a MTS. The core domain is highly conserved and contains essential catalytic residues, being connected to the C-terminal domain by

a poorly conserved Linker region formed by an extended pair of α-helices. TOP1 forms a toroidal fold with two modules entrapping the DNA molecule, a capping module matching the first half of the core domain (CAP domain or core sub-domains I and II) and a catalytic module comprising the second half of the core domain (CAT domain or core sub-domain III) the Linker and the C-terminal domain (Redinbo et al. 1998; Stewart et al. 1998; Takahashi et al. 2022). The catalytic module includes several active sites relevant for the protein activity (Champoux 2001). The Hinge is a five-residue loop connecting the capping and catalytic modules whose flexibility permits the opening/closing of the enzyme and the entry of DNA (Takahashi et al. 2022). The C-terminal domain is highly conserved and includes the Tyr723 active site which forms a transient phosphotyrosyl linkage to one DNA strand, catalysing changes in DNA topology (Stewart et al. 1996).

Importantly, TOP1 is the target of the camptothecin family of anticancer agents that binds to and reversibly stabilizes the covalent TOP1-DNA complex, resulting in double stranded DNA breaks and apoptosis, preferentially in cancer cells that often overexpress TOP1 (Pommier 2006; Pommier et al. 2010). TOP1MT is also sensitive to camptothecin agents, but it is not an *in vivo* target due to the alkaline mitochondria matrix that inactivates the drug (Tua et al. 1997; Zhang et al. 2001; Zhang and Pommier 2008). However, several mutations in TOP1 are known to impact the efficacy of camptothecin (Chrencik et al. 2004; Cretaio et al. 2007; Saleem et al. 2000).

Previous works have compared Type IB topoisomerases from different species, but often focused on a specific section of the protein or explored only a few animal species [e.g., (Champoux 1998; Takahashi et al. 2022; Zhang et al. 2004)]. Here we present a detailed examination of the evolutionary history of Type IB topoisomerases using a variety of animals that represent the main taxonomic groups of Metazoa. In particular, we evaluated the molecular evolution and adaptation processes and the origin of the TOP1 and TOP1MT paralogues in vertebrates.

5.3. Material and Methods

TOPIB sequences

TOPIB protein sequences from the main Metazoa phyla were retrieved from the NCBI non-redundant protein sequences (nr) database via the protein-protein BLAST (blastp) suite, using as query sequences from species close to the target taxonomic group (Supplementary Fig. S5.1). Short sequences with less than half of the average of TOPIB length were ignored since they often represent partial protein sequences derived from gaps in assembled genomes in which the contigs do not cover the complete genomic region. Possibly by the same reason, we fail to detect one or both the paralogues in the sequenced genome of some species.

Denisovan and Neanderthal *TOP1* and *TOP1MT* sequences were downloaded from the UCSC Genome Browser (http://genome.ucsc.edu/) (Kent et al. 2002). All BAM reads for tracks *Denisova* and *Neanderthal Cntgs* matching the Human Mar. 2006 (NCBI36/hg18) chr20:39,090,876-39,186,540 (*TOP1*) and chr8:144,462,903-144,488,425 (*TOP1MT*) were downloaded. The BAM reads from each track were then reassembled against the human *TOP1* (NC_000020.11) and *TOP1MT* (NC_000008.11) reference sequences using Geneious v2022.1.1 (http://www.geneious.com). We only considered a variable position in Denisovan and Neanderthal genomes when: 1) at least two reads overlap in that position; 2) the variant represents more than 75% of all the reads and 3) the difference is not at the end of a read. The variations between modern humans and Neanderthals were also confirmed in the assembly available at The Neandertal Genome Project (http://neandertal.ensemblgenomes.org).

TOPIB sequence alignments

The TOPIB protein sequences were aligned with the Geneious alignment in three datasets: Metazoa (n = 161), Chordata TOP1 (n = 48) and Chordata TOP1MT (n = 48). The conservation across the alignments was measured with the percentage of pairwise identity (PI) that compares base pairs at every site. The same species were used in the Chordata alignments to avoid biases and facilitate the comparison of results. The coding domain sequences (CDS) of the orthologues of human TOP1 (ENSG00000198900) and TOP1MT (ENSG00000184428) were obtained from the Ensembl Genome Server (Hunt et al. 2018).

Phylogenetic analyses

We analysed the TOPIB duplication events in chordates with a phylogenetic tree built with 37 protein sequences from Cephalochordata, Tunicata and Vertebrata species, and considering *Acanthaster planci* and *Strongylocentrotus purpuratus* (Echinodermata) as outgroups. We used Gblocks 0.91b server, running on Phylogeny.fr (Dereeper et al. 2008), to remove poorly aligned positions under the settings for a less stringent selection (Castresana 2000; Talavera and Castresana 2007). The best-fitting amino acid substitution model of evolution (LG+I+G4+F) was determined with ModelTest-NG (Darriba et al. 2020; Flouri et al. 2015). Next, we build a Bayesian phylogenetic tree with MrBayes v3.2.7a (Huelsenbeck and Ronquist 2001; Ronquist and Huelsenbeck 2003) running on the CIPRES Science Gateway v3.3 (Miller et al. 2010). The Metropolis-coupled Markov chain Monte Carlo (MCMC) process was set with two independent runs, each with four independent chains that ran simultaneously during 4,000,000 iterations. The average standard deviation of split frequencies of the final tree was 0.002339, indicating convergence among the independent runs. A burn-in value of 0.25 was applied following the program recommendation. The resulting phylogenetic tree was edited with FigTree v1.4.3 (http://tree.bio.ed.ac.uk/software/figtree).

Evaluation of selection

Molecular adaptation signatures in TOP1 and TOP1MT protein-coding sequence alignments were evaluated with the nonsynonymous/synonymous substitution rates ratio (*dN*/*dS*) (Del Amparo et al. 2021; Jeffares et al. 2015). First, we selected the best-fitting substitution model of DNA evolution and reconstructed a maximum likelihood (ML) phylogenetic tree. Next, we estimated *dN*/*dS* under a ML method, considering the reconstructed phylogenetic tree, implemented in the evolutionary framework Hyphy (Kosakovsky Pond and Frost 2005; Kosakovsky Pond et al. 2020). In particular, we applied the single-likelihood ancestor counting (SLAC) method for the *dN*/*dS* estimation, which has an accuracy similar to that from other likelihood-based methods and includes statistical evaluations (Kosakovsky Pond and Frost 2005).

Template of TOPIB protein structure

We considered as an illustrative template of the human TOPIB protein structure, the protein structure of the Protein Data Bank (PDB) (Berman et al. 2000) with code 1A36 (Stewart et al. 1998). The structure was analysed with Mol* (Sehnal et al. 2021) and RCSB PDB.

5.4. Results and Discussion

***TOP1* and *TOP1MT* paralogues originated in the first round of vertebrate tetraploidization (1R)**

Previous works have shown that TOPIB topoisomerases are ubiquitous in eukaryotes, and that only vertebrates have two *TOPIB* paralogues, named *TOP1* and *TOP1MT* (Forterre et al. 2007; Zhang et al. 2004). Our extensive search for *TOPIB* genes in the genome of all available chordates only retrieved paralogues in the cyclostomes (jawless vertebrates) and gnathostomes (jawed vertebrates), confirming the previous claiming that *TOPIB* paralogues only occur in vertebrates (Zhang et al. 2004). Our phylogeny placed cephalochordates at the root of Chordata (Fig. 5.2). The Tunicata (Urochordata) and Vertebrata form a clade known as Olfactores (Delsuc et al. 2006; Putnam et al. 2008; Satoh et al. 2014). The fast-evolving *Oikopleura dioica* forms a particularly long branch, as we previously found for TOPIIA (Moreira et al. 2022).

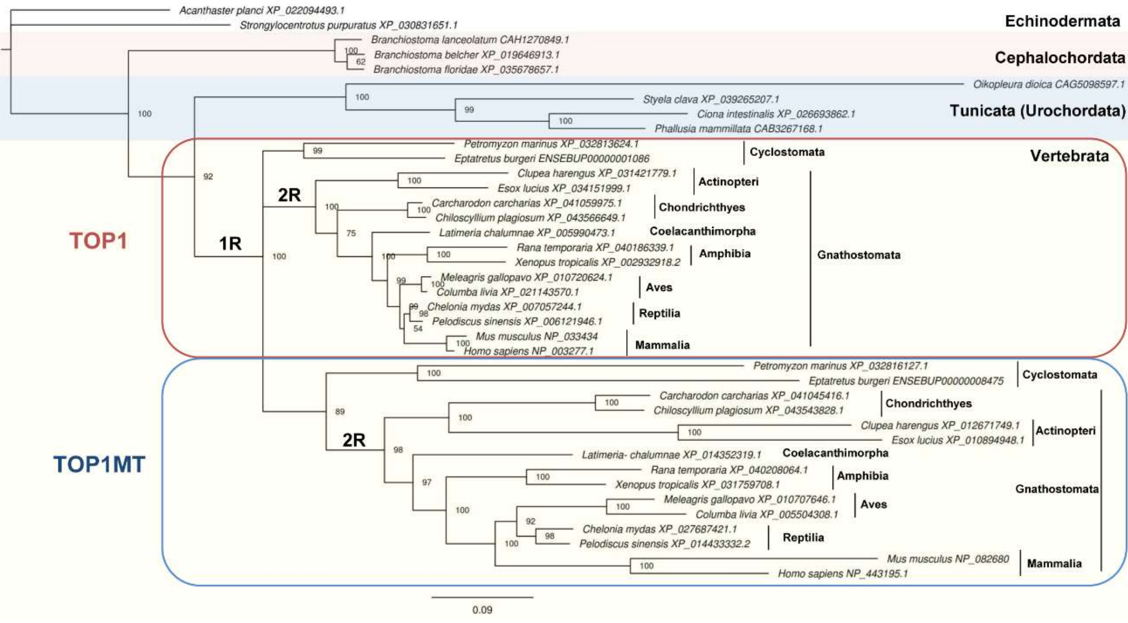


Figure 5.2. Phylogenetic analysis of TOPIB in chordates. Bayesian phylogenetic tree built with an alignment of 37 TOPIB protein sequences from chordates and considering two Echinodermata species as outgroup. The Bayesian posterior probabilities are shown on the internal nodes. The scale bar indicates substitutions per site. The putative occurrence of two rounds of tetraploidization (1R and 2R) is indicated.

The timing of the duplication event that gave rise to both paralogues remains unclear, particularly considering that the origin of vertebrates is associated with several gene and genome duplication events. Two rounds of tetraploidization, known as 1R and 2R, are believed to have occurred early in vertebrate evolution (Ohno 2013; Smith and Keinath 2015; Van de Peer et al. 2009). The timing of the tetraploidization events is still a matter of debate, but it was recently proposed that 1R preceded the divergence between cyclostomes and gnathostomes and 2R only occurred in gnathostomes (Aase-Remedios and Ferrier 2021; Nakatani et al. 2021; Simakov et al. 2020). Previous works observed that vertebrata *TOP1* and *TOP1MT* form two separate clusters (Forterre et al. 2007; Wang et al. 2009; Zhang et al. 2007), but were performed without sequences from cyclostomes. Our search for *TOPIB* genes in cyclostomes allowed us to retrieve two complete *TOPIB* sequences in two species, *Petromyzon marinus* and *Eptatretus burgeri.* We also noticed the presence of at least two paralogues in other cyclostomes (e.g., *Lethenteron camtschaticum*, *Entosphenus tridentatus*), but the genomic sequences were incomplete and thus were not used in the phylogenies. Therefore, it is likely that cyclostomes have at least two TOPIB paralogues, as observed in other vertebrates. In this concern, the *TOPIB* paralogues from *P. marinus* and *E. burgeri* did not cluster together in our phylogeny (Fig. 5.2). Instead, one pair clusters with TOP1MT sequences. Indeed, these two paralogues also display long branches, which are typical for the fast-evolving TOP1MT. Therefore, our analyses suggest that cyclostomes have a mitochondrial Type IB topoisomerase. The other pair of TOPIB paralogues from *P. marinus* and *E. burgeri* split from Gnathostomata TOP1 and TOP1MT at similar times. Our analysis is compatible with the idea that the duplication event that originated *TOP1* and *TOP1MT* is related with the first round of tetraploidization (1R). In this situation, *TOP1* and *TOP1MT* originated during the whole genome duplication in the early vertebrate evolution. The paralogues then diverged independently during the evolution of vertebrates, clustering in two separate branches (Fig. 5.2). The main difference between the phylogeny of the two genes is the placement of TOP1 from cyclostomes, which does not cluster with TOP1 from gnathostomes, as in the TOP1MT clade. Further analyses with additional sequences from Cyclostomata are necessary to better define the evolutionary history of these genes.

The specialization for acting on mtDNA may have occurred early in the radiation of vertebrates. In this concern, we previously identified that TOPIIA paralogues (*TOP2A* and *TOP2B*) present a different origin within chordates (Moreira et al. 2022). Here we found that *TOP2A* and *TOP2B* paralogues from Cyclostomata cluster together in a separate branch from all Gnathostomata paralogues. Altogether, our findings suggest

that the different classes of topoisomerases present different evolutionary histories in chordates.

Strong purifying selection acting on *TOP1* and *TOP1MT*

We estimated the *dN*/*dS* ratio to evaluate selection acting on TOP1 and TOP1MT paralogues of chordates (Table 5.1). We found that both genes present genetic signatures of negative (purifying) selection (*dN*/*dS* < 1), as noticed before in other topoisomerases (TOP3B, TOP2A, TOP2B) (Moreira et al. 2021; Moreira et al. 2022). The paralogue pairs TOP1/TOP1MT and TOP2B (*dN*/*dS* = 0.156) / TOP2A (*dN*/*dS* = 0.238) (Moreira et al. 2022) presented higher *dN*/*dS* ratios than TOP3B (*dN*/*dS* = 0.076) (Moreira et al. 2022), which has no paralogue.

Paralogues can exhibit asymmetric rates of sequence evolution (Conant and Wagner 2003; Scannell and Wolfe 2008; Van de Peer et al. 2001). The strength of negative selection was higher in TOP1 (*dN*/*dS* = 0.154) than in TOP1MT (*dN*/*dS* = 0.307). Indeed, TOP1 also exhibits a lower diversity compared with TOP1MT (Table 5.1). The essential activity of TOP1 across species (Lee et al. 1993; Morham et al. 1996) in different biological processes can explain its relatively high conservation. On the other hand, TOP1MT presents the highest *dN*/*dS* ratio among all the topoisomerases studied by us (Moreira et al. 2021; Moreira et al. 2022). Although it still evolved under negative selection, TOP1MT seems more permissive to accept amino acids changes than other topoisomerases. The higher diversity estimated in TOP1MT (in comparison to TOP1) can also be observed in the Chordata phylogeny, where TOP1MT branches are considerably longer than those for TOP1 (Fig. 5.2). The fast rate of change in *TOP1MT* can explain why finding orthologues for this gene is difficult. For example, the Ensembl genome browser only recognizes 77 orthologues for *TOP1MT*, in comparison with the 272 orthologues identified for TOP1 (accessed in April 2022). TOP1MT was also recognized as the only topoisomerase with highly frequent single nucleotide variants (SNVs) in the human population (Zhang et al. 2017). It was speculated that TOP1MT varies more than other topoisomerases due to several factors: *i*) it is a nonessential gene under less constraints to mutate; *ii*) it is in a subtelomeric end of a chromosome and/or *iii*) it is a relatively recent gene under adaptation to its activity in mitochondria (Zhang et al. 2017). Thus, the observed pattern can be the result from a combination of those factors. Comparing with our previous results, *TOP2B* and *TOP2A* are more conserved than *TOP1MT* despite being also paralogues that originated early in vertebrate evolution

(Moreira et al. 2022). Thus, we believe that these paralogues could be a good comparative model to study *TOP1MT* in future investigations.

Table 5.1. Selection pressure in *TOP1* and *TOP1MT*.

Gene	Dataset	*n*	Best substitution model	Global *dN/dS*[*]	Pairwise Identity (%)
TOP1	Chordata	74	SYM + G	0.154 [0.147–0.162]	82.0
TOP1MT	Chordata	74	SYM + I +G	0.307 [0.300–0.315]	61.3

[*] Global (entire sequences) *dN/dS* including the 95% confidence interval. Positively selected sites (PSS) were not detected.

Two missense mutations identified in the Neandertals TOP1MT Linker region

Neanderthals and Denisovans are extinct groups of hominins that inhabited Eurasia until around 40,000 years ago (Green et al. 2010; Reich et al. 2010). Previous works identified a few amino acid changes among modern humans and other hominins, some of which may have contributed to unique human traits (Green et al. 2010; Kuhlwilm and Boeckx 2019). Here, we searched for sequence differences in coding regions among modern human, Denisovan and Neanderthal *TOP1* and *TOP1MT* genes. However, we did not identify polymorphic positions in *TOP1* coding regions covered by Neanderthals or Denisovans sequence reads. On the contrary, we identified three nucleotide differences in the coding regions of *TOP1MT* (Table 5.2). A silent mutation in the CAP TOP1MT domain occurred in the human lineage. Next, two missense mutations were identified in the Neanderthal lineage. In particular, the mutations involved changes in two close amino acid positions (533 and 536) that belong to the Linker region (Fig. 5.3). Notice that the occurrence of missense mutations between modern humans and Neanderthals is rare (Green et al. 2010; Kuhlwilm and Boeckx 2019). When comparing present-day human and Neanderthals, Kuhlwilm and Boeckx (2019) identified 647 amino acid-changes in 571 genes. Among those genes, only 68 had two or more amino acid changes. Assuming that humans have 19,969 genes (Nurk et al. 2022), only 0.34% of those genes have more than one amino acid change, making it a rare event.

Two mutations occurring in the same sequence read seems particularly improbable. However, we identified the mutations in several reads, including both our assembly and the assembly available at the Neandertal Genome Project (Supplementary Fig. S5.2). Moreover, we fail to align the Neandertal reads with any other available sequence in GenBank, including *TOP1* gene and pseudogenes, which excludes a possible misplacement of reads from those regions in *TOP1MT*. The two mutations involved

amino acids with different physicochemical properties. In particular, two glutamines (polar uncharged side chain) were replaced by an arginine and a lysine (positively charged, basic, side chain). These different properties could affect the protein function, but further experimental analyses are required to corroborate this possibility. We previously identified two missense mutations in TOP2A when comparing present-day humans and Neanderthals (Moreira et al. 2022). It is interesting to note that missense mutations were only identified in the two topoisomerases (TOP1MT and TOP2A) that are less conserved in chordates, which supports the credibility of the identified sequence differences. The sequencing of additional Neanderthal and Denisovan samples will allow us to confirm if these sequence variations were fixed among these species.

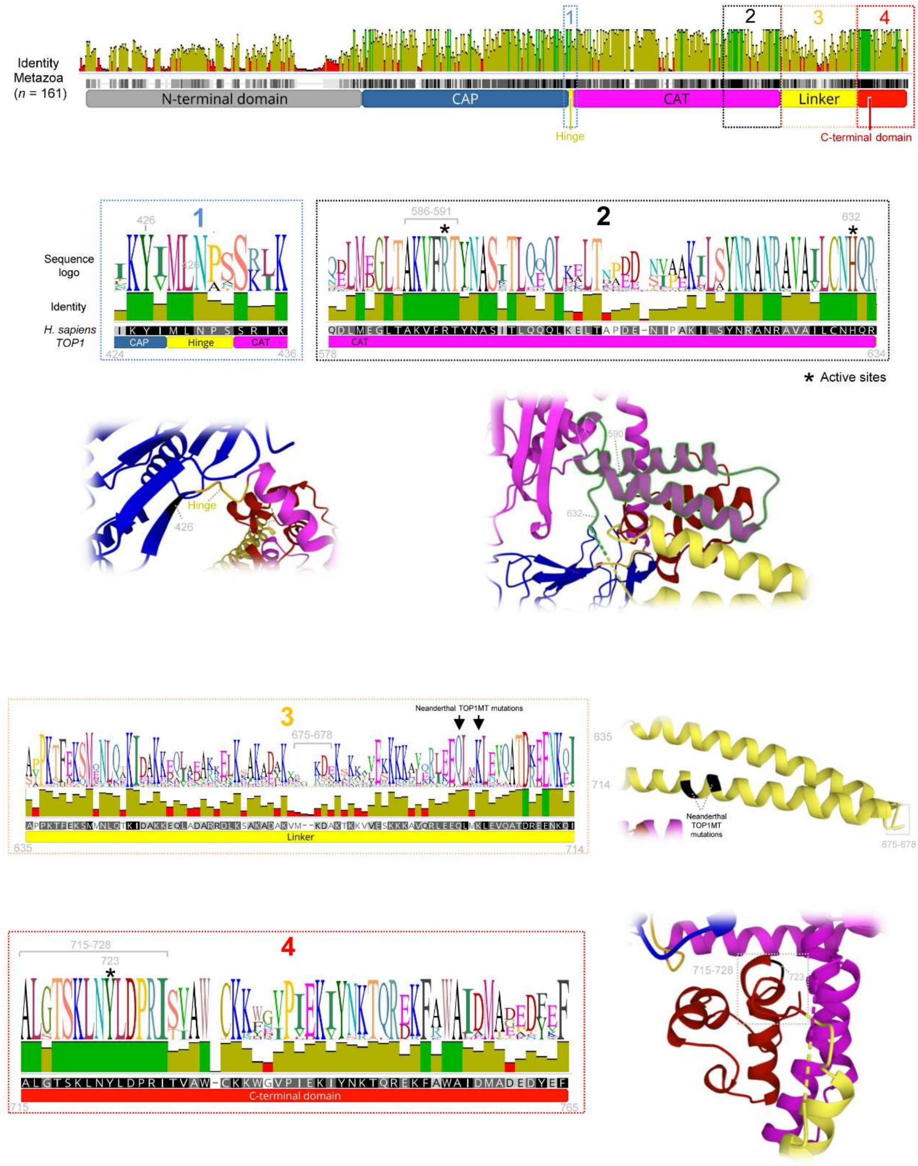


Figure 5.3. Structural conservation of TOPIB. The sequence identity plot was estimated from the 161 TOPIB protein sequences of the major chordate groups. The most conserved positions are indicated with brown bars, while the less conserved positions are shown using red bars. The sequence logo and an illustration of the protein structure for the highlighted regions are included.

Table 5.2. Sequence variants identified in *TOP1MT* coding sequences among modern humans (*H. sapiens*), Denisovan (*Denis*) and *Homo neanderthalensis* (*Neand*).

Gene	***Homo sapiens* reference sequences**						**Ancestral State***		**Variant in Denisovan and Neanderthal**		**Mutational event**				**Pairwise Identity (%) in Chordata**
	Sequence	**Genome position**	**nt**	**Protein position**	**aa**	**Protein domain**	**nt**	**aa**	**Species**	**nt**	**Probable event**	**Lineage**	**Type**	**Amino acid replaceme nt**	
TOP1MT	NC_000008.11	143,331,246	G	72	Asp	CAP	A	Asp	Denis	A	A>G	*H. sapiens*	Silent	-	26.4
		143,310,173	T	533	Gln	Linker	T	Gln	Neand	C	T>C	Neand	Missense	Gln-Arg	92.7
		143,310,165	G	536	Gln	Linker	G	Gln	Neand	T	G>T	Neand	Missense	Gln-Lys	84.7

* Nucleotide in *Pan paniscus* and *Gorilla gorilla*

Relevant TOP1 and TOP1MT sites for catalytic activities tend to be conserved across animals

The alignment of TOP1 and TOP1MT protein sequences from 48 representative chordate species confirms that TOP1MT (77% of pairwise sequence identity) is less conserved than TOP1 (sequence identity of 83.6%) (Fig. 5.3, Table 5.3). This result agrees with the long branches of TOP1MT in the Chordata phylogeny (Fig. 5.2) and its higher genetic diversity (Table 5.2). The N-terminal domain is the less conserved region in both proteins (sequence identities of 64.1% in TOP1 and 53.3% in TOP1MT), as noticed since the first studies on TOP1 (Champoux 1998; Champoux 2001; Stewart et al. 1996). The function of the N-terminal domain remains poorly understood partially due to a lack of structural information. However, it is dispensable for the catalytic activity of the enzyme (Alsner et al. 1992), suggesting that it could accept mutations without compromising the protein activity. Moreover, the N-terminal domain mediates TOP1 interactions with other proteins (Czubaty et al. 2005). These protein-protein interactions might experience different co-evolution processes among species that could explain the poor sequence conservation of the domain. The protein-protein binding regions identified in N-terminal domains of TOP1 (NLSs) and TOP1MT (MTS) are also poorly conserved, possibly due to evolution driven by different species requirements. Only TOP1 NLS-II and NLS-IV are relatively conserved in chordates (Table 5.3).

Table 5.3. Organization and conservation of TOP1 and TOP1MT protein domains and relevant sites. The percentage of pairwise identity was obtained in an alignment with all metazoans and with chordates alone.

Protein domains		TOP1			TOP1MT			TOPIB
		Human reference		Chordata (*n* = 48)	Human reference		Chordata (*n* = 48)	Metazoa (*n* = 161)
		Position	Length (aa)	Pairwise Identity (%)	Position	Length (aa)	Pairwise Identity (%)	Pairwise Identity (%)
Complete protein		1-765	765	83.6	1-601	601	77	70.5
N-terminal domain		1-213	213	64.1	1-49	49	53.3	41.9
Core domain	***CAP***	214-427	214	89	50-263	214	77.3	69.8
	Hinge	428-432	5	98.3	264-268	5	84.1	84.8
	CAT	433-635	202	94.5	269-470	202	81.4	78.4
Linker		636-714	80	87	471-550	80	63.1	57.3
C-terminal domain		715-765	51	93	551-601	51	85.4	82.5
N-terminal localization signals	**TOP1 NLS**	117-146	30	56.8	-	-	-	-
	TOP1 NLS-I	59-65	7	62.9	-	-	-	-
	TOP1 NLS-II	150-156	7	86.9	-	-	-	-
	TOP1 NLS-III	174-180	7	35.4	-	-	-	-
	TOP1 NLS-IV	192-198	7	79.2	-	-	-	-
	TOP1MT MTS	-	-	-	1-40	40	34.8	-
Active sites	**CAT**	488	R	100	324	R	91.8	97.5
	CAT	532	K	100	368	K	95.8	100
	CAT	590	R	100	426	R	100	100
	CAT	632	H	100	468	H	100	100
	C-terminal	723	Y	100	559	Y	100	100
Camptothecin-resistant mutations	**CAP**	361	F	100	197	F	100	98.8
	CAP	363	G	100	199	G	100	98.8
	CAP	364	R	100	200	R	100	100
	CAP	418	E	100	254	E	100	97.5
	CAT	503	G	100	339	G	91.8	97.5
	CAT	533	D	100	369	D	95.8	98.8
	Linker	653	A	77.2	489	A	62.8	51.6
	C-terminal	722	N	100	558	N	100	100
	C-terminal	729	T	66.3	565	S	71.6	65.8

The core domain (CAP, Hinge and CAT) is highly conserved due to its fundamental function on DNA binding during catalysis. We also found a high conservation in the DNA-binding regions in other topoisomerases (Moreira et al. 2021; Moreira et al. 2022), suggesting that these regions cannot accommodate changes due to maintaining the topoisomerase activity through a proper interaction with DNA. The CAT region is slightly more conserved than the CAP region, which agrees with the observation that only the CAT region is conserved in bacterial, viral and eukaryotic topoisomerases (Patel et al. 2010; Perry et al. 2006). The five-residue loop Hinge is conserved across metazoan (84.8% sequence identity), specifically the first two residues (TOP1 positions 428-429) that present the same amino acids in all the analysed species (Fig. 5.3). In addition, the tyrosine upstream of the Hinge (position 426) was also found conserved, in agreement

with a previous work suggesting that this position interacts with the DNA duplex and guides the motion of the CAP domain upon DNA binding to enable the enzyme closing (Takahashi et al. 2022). Within the CAT region, we noticed that near the Linker there are two conserved stretches of around 20 amino acids that flank a poorly conserved region (Fig. 5.3). In particular, we identified a region with 6 amino acids AKVFRT (TOP1 reference positions 586-591) that is 100% conserved across all the 161 metazoan analysed species. This region included several active sites. The CAP and CAT regions include sites conferring resistance to camptothecin and all of them are 100% conserved. The only variable sites conferring resistance to camptothecin were observed in the Linker (site 653, sequence identity of 77.2%) and C-terminal (site 729, sequence identity of 66.3%) regions.

We found that the Linker region is more variable than the surrounding core and C-terminal domains (Fig. 5.3, Table 5.3). The Linker consists of two long alpha helices connected by a short turn, forming an antiparallel coiled-coil configuration that protrudes away from the remainder of the enzyme (Stewart et al. 1998). We found that its conservation decreases with the increasing distance to the flanking domains and to the catalytic region of the enzyme (Fig. 5.3). The short turn at the end of the Linker (TOP1 positions 675-678) is extremely variable across species (21.4% of sequence identity), including some variation in length, suggesting that it can vary without affecting the protein function. The increase in conservation of the Linker in regions closer to the core of the enzyme indicates that amino acid replacements are less tolerated if they occur close to the catalytic region, possibly due to affecting the protein activity or the Linker connections to the DNA strand.

The C-terminal domain folds into a globular structure (Figs. 5.1C, 5.3) that includes the active-site nucleophile Tyr723 (Redinbo et al. 1998). This region also includes 8 residues near the Linker (718-722) with significant structural similarity with the bacteriophage family of DNA integrases (Redinbo et al. 1998). Our results confirmed previous observations about the high conservation of the C-terminal domain (Champoux 2001). In particular, we found that the 14 amino acids closer to the Linker (human TOP1 positions 715-728) are almost 100% conserved in all the analysed animal species (Fig. 5.3).

5.5. Conclusions

Type IB topoisomerases are widespread in the animal kingdom. Indeed, vertebrates present specialized topoisomerases to operate with the nuclear and mitochondrial genomes. However, little is known about its evolution and its genetic similarities among species. Here we analysed the molecular evolution of topoisomerases among a variety of animal species. Our phylogenetic investigation placed the event that originated the specialized TOP1 and TOP1MT proteins in the early evolution of vertebrates, possibly associated with whole-genome duplications. After the duplication event, the long-term evolution of both paralogues was primarily driven by strong purifying selection probably to maintain the protein function. However, we found that TOP1MT evolved much faster than TOP1 and other topoisomerases, perhaps related with its specific role within the mitochondria. The fast evolution of TOP1MT was also evident in the missense mutations detected in the Neanderthals, displaying a rare case of protein differences among hominids. Finally, comparison of topoisomerases among species showed that the relevant protein sites for catalytic activities are mainly conserved across animals, again probably caused by their relevant biological roles.

Acknowledgements

This work was partially supported by a research grant to FM (SFRH/BD/131584/2017) from Fundação para a Ciência e a Tecnologia. MA was supported by the Spanish Ministerio de Ciencia e Innovación through the Grants [RYC-2015-18241] and [PID2019-107931GA-I00].

5.6 References

Aase-Remedios ME, Ferrier DEK (2021) Improved understanding of the role of gene and genome duplications in chordate evolution with new genome and transcriptome sequences. Frontiers in Ecology and Evolution 9:429.

Alsner J, Svejstrup J, Kjeldsen E, Sørensen B, Westergaard O (1992) Identification of an N-terminal domain of eukaryotic DNA topoisomerase I dispensable for catalytic activity but essential for in vivo function. Journal of Biological Chemistry 267:12408-12411.

Berman HM et al. (2000) The protein data bank. Nucleic acids research 28:235-242.

Capranico G, Marinello J, Chillemi G (2017) Type i DNA topoisomerases. Journal of medicinal chemistry 60:2169-2192.

Castresana J (2000) Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Molecular biology and evolution 17:540-552.

Champoux JJ (1998) Domains of human topoisomerase I and associated functions. Progress in nucleic acid research and molecular biology 60:111-132.

Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. Annual review of biochemistry 70:369-413.

Cheng C, Kussie P, Pavletich N, Shuman S (1998) Conservation of structure and mechanism between eukaryotic topoisomerase I and site-specific recombinases. Cell 92:841-850.

Chrencik JE, Staker BL, Burgin AB, Pourquier P, Pommier Y, Stewart L, Redinbo MR (2004) Mechanisms of camptothecin resistance by human topoisomerase I mutations. Journal of molecular biology 339:773-784.

Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. Genome Research 13:2052-2058.

Cretaio E, Pattarello L, Fontebasso Y, Benedetti P, Losasso C (2007) Human DNA topoisomerase IB: structure and functions. The Italian journal of biochemistry 56:91-102.

Czubaty A et al. (2005) Proteomic analysis of complexes formed by human topoisomerase I. Biochimica et Biophysica Acta (BBA)-Proteins and Proteomics 1749:133-141.

Darriba D, Posada D, Kozlov AM, Stamatakis A, Morel B, Flouri T (2020) ModelTest-NG: a new and scalable tool for the selection of DNA and protein evolutionary models. Molecular Biology and Evolution 37:291-294.

de la Loza MD, Wellinger RE (2009) A novel approach for organelle-specific DNA damage targeting reveals different susceptibility of mitochondrial DNA to the anticancer drugs camptothecin and topotecan. Nucleic acids research 37:e26-e26.

Del Amparo R, Branco C, Arenas J, Vicens A, Arenas M (2021) Analysis of selection in protein-coding sequences accounting for common biases. Briefings in Bioinformatics 22:bbaa431.

Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965-968.

Dereeper A et al. (2008) Phylogeny. fr: robust phylogenetic analysis for the non-specialist. Nucleic acids research 36:W465-W469.

Douarre C et al. (2012) Mitochondrial topoisomerase I is critical for mitochondrial integrity and cellular energy metabolism. PloS one 7:e41094.

Flouri T et al. (2015) The phylogenetic likelihood library. Systematic biology 64:356-362.

Forterre P, Gribaldo S, Gadelle D, Serre M-C (2007) Origin and evolution of DNA topoisomerases. Biochimie 89:427-446.

Green RE et al. (2010) A draft sequence of the Neandertal genome. science 328:710-722.

Huelsenbeck JP, Ronquist F (2001) MRBAYES: Bayesian inference of phylogenetic trees. Bioinformatics 17:754-755.

Hunt SE et al. (2018) Ensembl variation resources. Database 2018:bay 119.

Jeffares DC, Tomiczek B, Sojo V, dos Reis M (2015) A beginners guide to estimating the non-synonymous to synonymous rate ratio of all protein-coding genes in a genome. In: Parasite Genomics Protocols. Springer, pp 65-90

Juan C-C et al. (1988) Human DNA topoisomerase I is encoded by a single-copy gene that maps to chromosome region 20q12-13.2. Proceedings of the National Academy of Sciences 85:8910-8913.

Kent WJ, Sugnet CW, Furey TS, Roskin KM, Pringle TH, Zahler AM, Haussler D (2002) The human genome browser at UCSC. Genome research 12:996-1006.

Khiati S et al. (2015) Lack of mitochondrial topoisomerase I (TOP1mt) impairs liver regeneration. Proceedings of the National Academy of Sciences 112:11282-11287.

Kosakovsky Pond SL, Frost SD (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. Molecular biology and evolution 22:1208-1222.

Kosakovsky Pond SL et al. (2020) HyPhy 2.5—a customizable platform for evolutionary hypothesis testing using phylogenies. Molecular biology and evolution 37:295-299.

Kuhlwilm M, Boeckx C (2019) A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. Scientific reports 9:1-14.

Lee MP, Brown SD, Chen A, Hsieh T-s (1993) DNA topoisomerase I is essential in Drosophila melanogaster. Proceedings of the National Academy of Sciences 90:6656-6660.

Miller MA, Pfeiffer W, Schwartz T Creating the CIPRES Science Gateway for inference of large phylogenetic trees. In: 2010 gateway computing environments workshop (GCE), 2010. Ieee, pp 1-8

Mo Y-Y, Wang C, Beck WT (2000) A novel nuclear localization signal in human DNA topoisomerase I. Journal of Biological Chemistry 275:41107-41113.

Moreira F, Arenas M, Videira A, Pereira F (2021) Molecular Evolution of DNA Topoisomerase III Beta (TOP3B) in Metazoa. Journal of molecular evolution 89:384-395.

Moreira F, Arenas M, Videira A, Pereira F (2022) Evolutionary History of TOPIIA Topoisomerases in Animals. Journal of Molecular Evolution 90:149-165.

Morham SG, Kluckman KD, Voulomanos N, Smithies O (1996) Targeted disruption of the mouse topoisomerase I gene by camptothecin selection. Molecular and cellular biology 16:6804-6809.

Nakatani Y, Shingate P, Ravi V, Pillai NE, Prasad A, McLysaght A, Venkatesh B (2021) Reconstruction of proto-vertebrate, proto-cyclostome and proto-gnathostome genomes provides new insights into early vertebrate evolution. Nature communications 12:1-14.

Nurk S et al. (2022) The complete sequence of a human genome. Science 376:44-53.

Ohno S (2013) Evolution by gene duplication. Springer Science & Business Media,

Palle K, Pattarello L, van der Merwe M, Losasso C, Benedetti P, Bjornsti M-A (2008) Disulfide cross-links reveal conserved features of DNA topoisomerase I architecture and a role for the N terminus in clamp closure. Journal of Biological Chemistry 283:27767-27775.

Patel A, Yakovleva L, Shuman S, Mondragón A (2010) Crystal structure of a bacterial topoisomerase IB in complex with DNA reveals a secondary DNA binding site. Structure 18:725-733.

Perry K, Hwang Y, Bushman FD, Van Duyne GD (2006) Structural basis for specificity in the poxvirus topoisomerase. Molecular cell 23:343-354.

Pommier Y (2006) Topoisomerase I inhibitors: camptothecins and beyond. Nature Reviews Cancer 6:789-802.

Pommier Y, Leo E, Zhang H, Marchand C (2010) DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. Chemistry & biology 17:421-433.

Pommier Y, Nussenzweig A, Takeda S, Austin C (2022) Human topoisomerases and their roles in genome stability and organization. Nature Reviews Molecular Cell Biology:1-21.

Pommier Y, Sun Y, Shar-yin NH, Nitiss JL (2016) Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nature reviews Molecular cell biology 17:703-721.

Putnam NH et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064-1071.

Redinbo MR, Stewart L, Kuhn P, Champoux JJ, Hol WG (1998) Crystal structures of human topoisomerase I in covalent and noncovalent complexes with DNA. Science 279:1504-1513.

Reich D et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053-1060.

Ronquist F, Huelsenbeck JP (2003) MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19:1572-1574.

Saleem A, Edwards TK, Rasheed Z, Rubin EH (2000) Mechanisms of resistance to camptothecins. Annals of the New York Academy of Sciences 922:46-55.

Satoh N, Rokhsar D, Nishikawa T (2014) Chordate evolution and the three-phylum system. Proceedings of the Royal Society B: Biological Sciences 281:20141729.

Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome research 18:137-147.

Sehnal D et al. (2021) Mol* Viewer: modern web app for 3D visualization and analysis of large biomolecular structures. Nucleic Acids Research.

Simakov O et al. (2020) Deeply conserved synteny resolves early events in vertebrate evolution. Nature ecology & evolution 4:820-830.

Smith JJ, Keinath MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. Genome research 25:1081-1090.

Stewart L, Ireton GC, Parker LH, Madden KR, Champoux JJ (1996) Biochemical and Biophysical Analyses of Recombinant Forms of Human Topoisomerase I (∗). Journal of Biological Chemistry 271:7593-7601.

Stewart L, Redinbo MR, Qiu X, Hol WG, Champoux JJ (1998) A model for the mechanism of human topoisomerase I. Science 279:1534-1541.

Takahashi DT et al. (2022) Topoisomerase I (TOP1) dynamics: conformational transition from open to closed states. Nature communications 13:1-11.

Talavera G, Castresana J (2007) Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Systematic biology 56:564-577.

Tua A, Wang J, Kulpa V, Wernette C (1997) Mitochondrial DNA topoisomerase I of Saccharomyces cerevisiae. Biochimie 79:341-350.

Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nature Reviews Genetics 10:725-732.

Van de Peer Y, Taylor JS, Braasch I, Meyer A (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. Journal of molecular evolution 53:436-446.

Wang J, Kearney K, Derby M, Wernette CM (1995) On the Relationship of the ATP-Independent, Mitochondrial-Associated DNA Topoisomerase of Saccharomyces cerevisiae to the Nuclear Topoisomerase I. Biochemical and biophysical research communications 214:723-729.

Wang X, Huang Y, Lavrov DV, Gu X (2009) Comparative study of human mitochondrial proteome reveals extensive protein subcellular relocalization after gene duplications. BMC Evolutionary Biology 9:1-11.

Yang G, Kunze N, Baumgärtner B, Jiang Z, Sapp M, Knippers R, Richter A (1990) Molecular structures of two human DNA topoisomerase I retrosequences. Gene 91:247-253.

Zhang H, Barceló JM, Lee B, Kohlhagen G, Zimonjic DB, Popescu NC, Pommier Y (2001) Human mitochondrial topoisomerase I. Proceedings of the National Academy of Sciences 98:10608-10613.

Zhang H, Meng L-H, Pommier Y (2007) Mitochondrial topoisomerases and alternative splicing of the human TOP1mt gene. Biochimie 89:474-481.

Zhang H, Meng LH, Zimonjic DB, Popescu NC, Pommier Y (2004) Thirteen-exon-motif signature for vertebrate nuclear and mitochondrial type IB topoisomerases. Nucleic acids research 32:2087-2092.

Zhang H, Pommier Y (2008) Mitochondrial topoisomerase I sites in the regulatory D-loop region of mitochondrial DNA. Biochemistry 47:11196-11203.

Zhang H, Seol Y, Agama K, Neuman KC, Pommier Y (2017) Distribution bias and biochemical characterization of TOP1MT single nucleotide variants. Scientific reports 7:1-11.

Zhang H, Zhang Y-W, Yasukawa T, Dalla Rosa I, Khiati S, Pommier Y (2014) Increased negative supercoiling of mtDNA in TOP1mt knockout mice and presence of topoisomerases IIα and IIβ in vertebrate mitochondria. Nucleic acids research 42:7259-7267.

5.7. Supplementary Material

Supplementary Fig. S5.1. TOP1 and TOP1MT protein sequences used in this study.

Taxonomy	***Species - Accession number***
Annelida; Clitellata	*Helobdella robusta - XP_009017579.1*
Annelida; Polychaeta	*Capitella teleta - ELU02442.1*
	Owenia fusiformis - CAC9596629.1
	Dimorphilus gyrociliatus - CAD5111131.1
Arthropoda; Arachnida	*Stegodyphus dumicola - XP_035223075.1*
	Parasteatoda tepidariorum - XP_015914810.1
	Araneus ventricosus - GBM09797.1
	Varroa jacobsoni - XP_022703641.1
	Varroa destructor - XP_022671948.1
	Rhipicephalus sanguineus - XP_037529830.1
	Dermacentor silvarum - XP_037558923.1
	Ixodes scapularis - XP_029851295.2
Arthropoda; Insecta	*Coptotermes formosanus - GFG35199.1*
	Cephus cinctus - XP_015608843.1
	Polistes dominula - XP_015173950.1
	Harpegnathos saltator - XP_011140118.2
	Pogonomyrmex barbatus - XP_011636012.1
	Formica exsecta - XP_029668836.1
	Bombus impatiens - XP_024221338.1
	Bombus pyrosoma - XP_043594904.1
	Bombus terrestris - XP_012169008.1
	Tribolium madens - XP_044267953.1
Chordata; Actinopterygii	*Esox lucius - XP_034151999.1_TOP1*
	Esox lucius - XP_010894948.1_MT
	Danio rerio - XP_002666448.2_TOP1
	Danio rerio - XP_002665372.2_MT
	Astyanax mexicanus - XP_022535346.1_MT
	Astyanax mexicanus - KAG9269888.1_TOP1
	Pundamilia nyererei - XP_005734805.1_MT
	Maylandia zebra - XP_004571958.1_TOP1
	Maylandia zebra - XP_004568268.1_MT
	Xiphophorus maculatus - XP_005801712.1_TOP1
	Xiphophorus maculatus - XP_005795820.2_MT
	Poecilia formosa - XP_007565717.1_TOP1
	Poecilia formosa - XP_007553331.1_MT
	Oryzias latipes - XP_023811830.1_MT
	Oryzias latipes - XP_023810923.1_TOP1
	Clupea harengus - XP_031421779.1_TOP1
	Clupea harengus - XP_012671749.1_MT
	Lepisosteus oculatus - XP_015209608.1_MT
	Lepisosteus oculatus - XP_006639584.2_TOP1
	Pundamilia nyererei - XP_005740471.1_TOP1

Supplementary Fig. S5.1. *cont*.

Chordata; Amphibia	*Microcaecilia unicolor* - XP_030075669.1_MT
	Microcaecilia unicolor - XP_030069429.1_TOP1
	Rhinatrema bivittatum - XP_029468354.1_TOP1
	Rhinatrema bivittatum - XP_029448031.1_MT
	Xenopus laevis - XP_018123795.1_MT
	Xenopus laevis - NP_001084031.1_TOP1
	Xenopus tropicalis - XP_031759708.1_MT
	Xenopus tropicalis - XP_002932918.2_TOP1
	Rana temporaria - XP_040208064.1_MT
	Rana temporaria - XP_040186339.1_TOP1
Chordata; Aves	*Dromaius novaehollandiae* - XP_025969476.1_MT
	Dromaius novaehollandiae - XP_025948785.1_TOP1
	Melopsittacus undulatus - XP_033922770.1_TOP1
	Melopsittacus undulatus - XP_033918251.1_MT
	Fulmarus glacialis - XP_009585324.1_TOP1
	Fulmarus glacialis - XP_009575693.1_MT
	Chiroxiphia lanceolata - XP_032569132.1_MT
	Chiroxiphia lanceolata - XP_032560816.1_TOP1
	Ficedula albicollis - XP_005057170.1_TOP1
	Ficedula albicollis - XP_005042662.1_MT
	Meleagris gallopavo - XP_010720624.1_TOP1
	Meleagris gallopavo - XP_010707646.1_MT
	Cygnus olor - XP_040431480.1_TOP1
	Cygnus olor - XP_040403374.1_MT
	Cygnus atratus - XP_035422285.1_MT
	Cygnus atratus - XP_035403913.1_TOP1
	Oxyura jamaicensis - XP_035174776.1_MT
	Columba livia - XP_021143570.1_TOP1
	Columba livia - XP_005504308.1_MT
	Calidris pugnax - XP_014821120.1_MT
	Calidris pugnax - XP_014796749.1_TOP1
	Oxyura jamaicensis - XP_035199725.1_TOP1
Chordata; Cephalochordata	*Branchiostoma lanceolatum* - CAH1270849.1
	Branchiostoma floridae - XP_035678657.1
	Branchiostoma belcheri - XP_019646913.1
Chordata; Chondrichthyes	*Callorhinchus milii* - XP_042197656.1_TOP1
	Chiloscyllium plagiosum - XP_043566649.1_TOP1
	Chiloscyllium plagiosum - XP_043543828.1_MT
	Carcharodon carcharias - XP_041059975.1_TOP1
	Carcharodon carcharias - XP_041045416.1_MT
Chordata; Coelacanthimorpha	*Latimeria chalumnae* - XP_014352319.1
	Latimeria chalumnae - XP_005990473.1
Chordata; Cyclostomata	*Petromyzon marinus* - XP_032816127.1_MT
	Petromyzon marinus - XP_032813624.1_TOP1

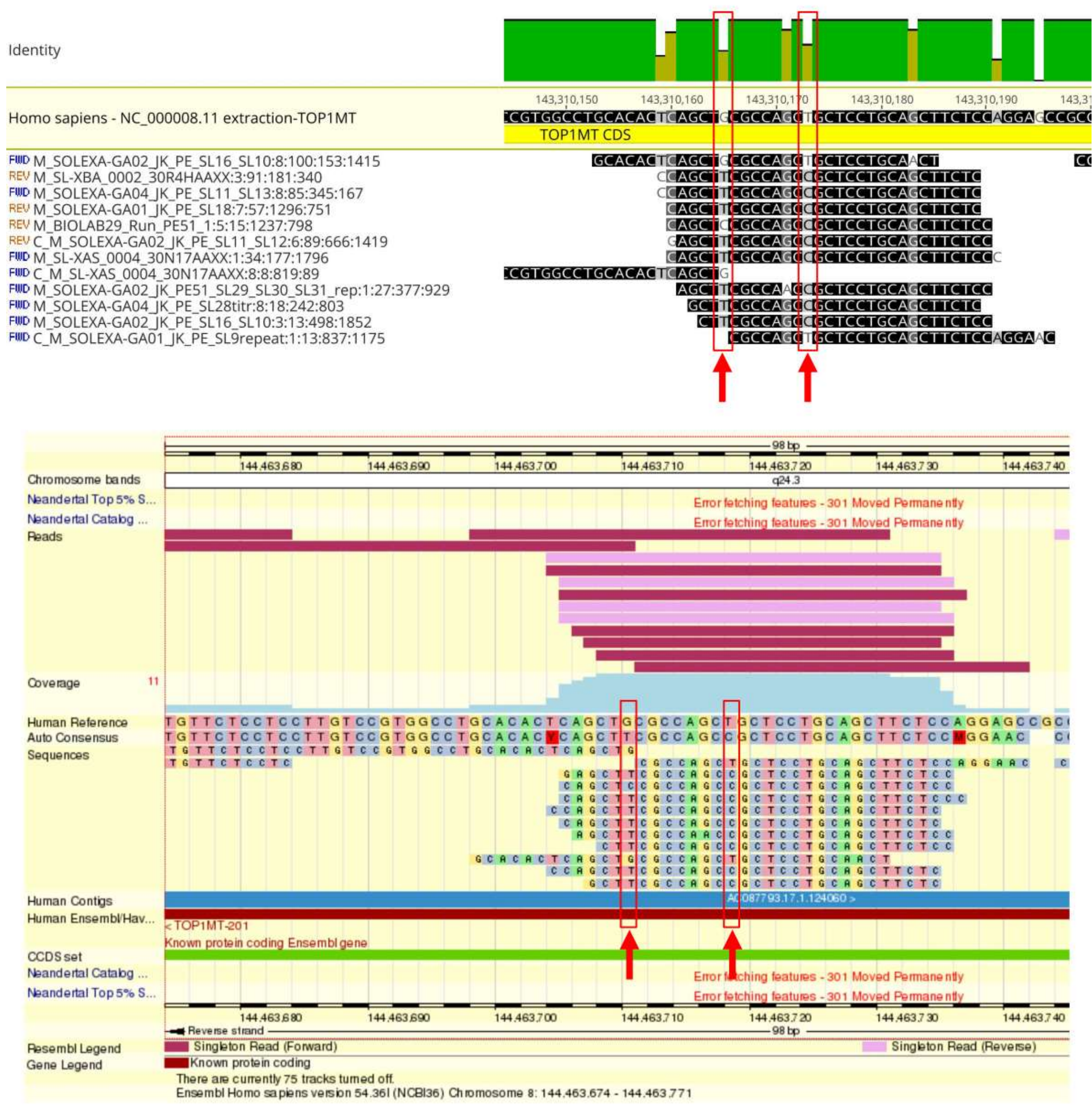
Supplementary Fig. S51. *cont*.

Chordata; Mammalia	*Carlito syrichta* - XP_008058764.1_TOP1
	Carlito syrichta - XP_008049557.1_MT
	Nomascus leucogenys - XP_003280801.2_MT
	Nomascus leucogenys - XP_003253628.1_TOP1
	Pongo abelii - XP_024106654.1_MT
	Pongo abelii - XP_024094656.1_TOP1
	Pan paniscus - XP_008971078.2_MT
	Homo sapiens - NP_443195.1_MT
	Homo sapiens - NP_003277.1_TOP1
	Gorilla gorilla gorilla - XP_004047664.2_MT
	Rhinopithecus roxellana - XP_030777108.1_MT
	Rhinopithecus roxellana - XP_010374253.1_TOP1
	Papio anubis - XP_031525665.1_MT
	Papio anubis - XP_021776748.1_TOP1
	Macaca mulatta - XP_015001609.2_MT
	Macaca mulatta - NP_001253441.1_TOP1
	Cercocebus atys - XP_011921036.1_TOP1
	Cercocebus atys - XP_011905464.1_MT
	Pan paniscus - XP_003825942.1_TOP1
	Gorilla gorilla - XP_004062202.1_TOP1
Chordata; Reptilia	*Pelodiscus sinensis* - XP_014433332.2_MT
	Pelodiscus sinensis - XP_006121946.1_TOP1
	Gopherus evgoodei - XP_030409772.1_MT
	Gopherus evgoodei - XP_030389046.1_TOP1
	Chelonoidis abingdonii - XP_032623226.1_TOP1
	Chelonoidis abingdonii - XP_032619134.1_MT
	Mauremys mutica - XP_044859482.1_MT
	Mauremys mutica - XP_044841689.1_TOP1
	Trachemys scripta elegans - XP_034643158.1_TOP1
	Trachemys scripta elegans - XP_034615486.1_MT
	Chrysemys picta bellii - XP_023958405.2_MT
	Chrysemys picta bellii - XP_008166335.2_TOP1
	Dermochelys coriacea - XP_038249347.1_MT
	Dermochelys coriacea - XP_038225734.1_TOP1
	Chelonia mydas - XP_027687421.1_MT
	Chelonia mydas - XP_007057244.1_TOP1
Chordata; Tunicata	*Styela clava* - XP_039265207.1
	Ciona intestinalis - XP_026693862.1
	Phallusia mammillata - CAB3267168.1
	Oikopleura dioica - CAG5098597.1
Cnidaria; Anthozoa	*Pocillopora damicornis* - XP_027046231.1
	Acropora millepora - XP_029192757.2
	Nematostella vectensis - XP_032231206.1
	Actinia tenebrosa - XP_031560302.1

Supplementary Fig. S5.1. *cont.*

Echinodermata; Asteroidea	*Patiria miniata* - *XP_038077725.1*
	Acanthaster planci - *XP_022094493.1*
	Asterias rubens - *XP_033625993.1*
Echinodermata; Crinoidea	*Anneissia japonica* - *XP_033125260.1*
Echinodermata; Echinoidea	*Lytechinus variegatus* - *XP_041454051.1*
	Strongylocentrotus purpuratus - *XP_030831651.1*
Echinodermata; Holothuroidea	*Apostichopus japonicus* - *PIK62638.1*
Hemichordata; Enteropneusta	*Saccoglossus kowalevskii* - *XP_006814310.1*
Mollusca; Bivalvia	*Pecten maximus* - *XP_033754799.1*
	Mytilus edulis - *CAG2232811.1*
	Dreissena polymorpha - *KAH3774199.1*
Mollusca; Cephalopoda	*Octopus bimaculoides* - *XP_014781973.1*
Mollusca; Gastropoda	*Haliotis rubra* - *XP_046570009.1*
	Aplysia californica - *XP_005091292.1*
	Elysia chlorotica - *RUS80862.1*
	Biomphalaria glabrata - *XP_013086265.1*
	Batillaria attramentaria - *KAG5707923.1*
	Pomacea canaliculata - *XP_025076151.1*
Nematoda; Enoplea	*Trichinella pseudospiralis* - *KRY68761.1*
	Trichinella papuae - *KRZ70158.1*
	Trichinella britovi - *KRY50853.1*
Platyhelminthes; Cestoda	*Echinococcus granulosus* - *KAH9281024.1*
	Echinococcus multilocularis - *CDS39478.1*
	Taenia asiatica - *VDK20333.1*
Platyhelminthes; Trematoda	*Paragonimus heterotremus* - *KAF5397961.1*
	Clonorchis sinensis - *KAG5443823.1*
	Fasciolopsis buski - *KAA0200075.1*
	Fasciola gigantica - *TPP66131.1*
Porifera; Demospongiae	*Amphimedon queenslandica* - *XP_011404857.2*
Tardigrada; Eutardigrada	*Ramazzottius varieornatus* - *GAU96790.1*
	Hypsibius dujardini - *OQV12457.1*

Supplementary Fig. S5.2. Alignment of Neanderthal sequence reads against the human TOP1MT reference sequence. The two missense mutations identified in the Neanderthal lineage and indicated. The top alignment was done by us using the sequences downloaded from the UCSC Genome Browser. The bottom image is a snapshot of the assembly available at The Neandertal Genome Project (http://neandertal.ensemblgenomes.org).



Chapter 6
General Discussion

6. General Discussion

6.1. The use of topoisomerases for phylogenetic inferences

DNA topoisomerases are key DNA metabolic enzymes that change the topology of DNA. By this reason, they are found across different taxonomic groups where topological problems in DNA must be solved (Didier et al. 2021; Forterre et al. 2007). The ubiquitous nature of DNA topoisomerases makes them informative genetic markers to infer deep phylogenies, like those obtained with other important cellular components, such as 16S ribosomal RNA (rRNA), ribosomal proteins, ATP synthases, etc (Forterre and Gadelle 2009). Overall, our phylogenetic inferences using animal topoisomerases are in agreement with the view that these enzymes are conserved and can be used to support prevailing taxonomic and phylogenetic hypotheses (Chapters 3, 4 and 5).

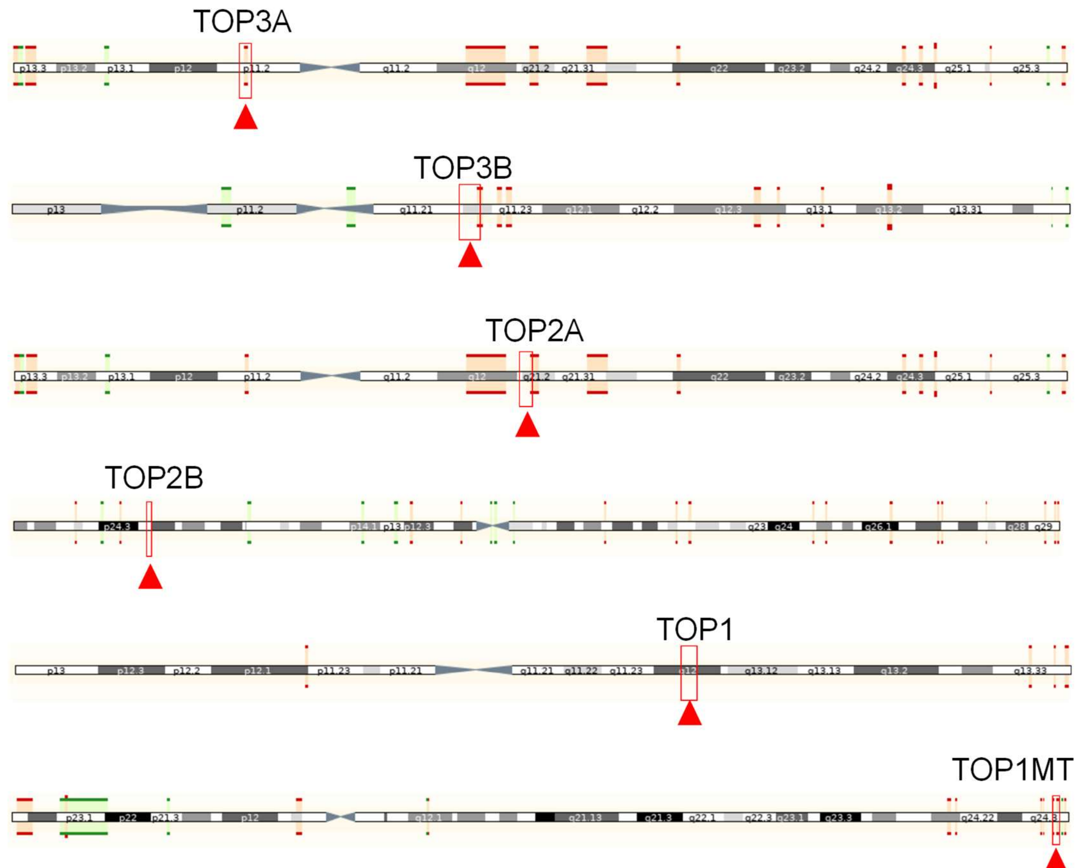
The distribution of the different types of topoisomerases across animals is relatively conserved, with four topoisomerases found in most (if not all) animals: two Type IA (TOP3A and TOP3B), one Type IIA (TOP2) and one Type IB (TOP1). Vertebrates have two additional topoisomerases, one extra Type IIA and one extra Type IB. Despite the apparent similar distribution of topoisomerases across animals, our search for orthologues resulted in very different numbers of identified cases for each protein. It was easier to retrieve orthologues for some genes than for others. For example, Type IA (*TOP3A* and *TOP3B*) had less orthologues than the other types when using the OrthoDB database (Kriventseva et al. 2018), and almost all identified orthologues were for *TOP3B*. By this reason, we only used TOP3B protein sequences in our study of Type IA topoisomerase (Chapter 3). We also noted that *TOP1MT* stood out by having less identified orthologues than the other topoisomerases in the Ensembl databases (Table 6.1 and Chapter 5).

Table 6.1. Number of orthologues reported in the Ensembl genome browser (http://www.ensembl.org) for all human topoisomerase genes (accessed in April 2022).

Gene	Number of sequences
TOP3A	204
TOP3B	205
TOP2A	206
TOP2B	211
TOP1	272
TOP1MT	77

The reasons for the discrepancy in the number of retrieved orthologues could be related with the different rate of evolution observed in topoisomerase genes. For example, *TOP1MT* had the highest nucleotide diversity and substitution rate among topoisomerases (Chapter 5), which makes it difficult to find sufficient homology in genome searches. Moreover, identification of orthologues also depends on the availability of complete genomic sequences. It is well known that some genomic regions are more difficult to sequence and assembly than others (Alkan et al. 2011; Treangen and Salzberg 2012), which may be the case for the location of some of the topoisomerase genes. *TOP1MT* is located near the telomeric region of chromosome 8 in humans (Figure 6.1). It would be interesting to see if the *TOP1MT* is also located near the end of a chromosome in other vertebrates, but the number of genomes with telomere-to-telomere chromosome assemblies is still very small (Logsdon et al. 2020). If *TOP1MT* is indeed located near the end of chromosomes across vertebrates, that would explain the low number of identified orthologues, since peripheral chromosomal regions are often rich in repeats and difficult to sequence.

Figure 6.1. Location of all topoisomerase genes in the human chromosomes. Images retrieved from the Ensembl genome browser.



Despite the differences in the number of retrieved sequences, the phylogenies built for Type IA (Chapter 3) and Type IIA (Chapter 4) showed similar pattern of animal divergence and evolution. Our preliminary trees built for Type IB (TOP1 and TOP1MT) including basal animal phyla yielded low statistical support and were not included in the work of Chapter 5, where we decided to focus on the phylogeny within Chordates. As expected, groups such as Cnidaria were at the base of the TOP3B and Type IIA (TOP2A and TOP2B) trees, supporting their ancient origin within Metazoa. The trees only disagree in the radiation of Mollusca and Annelida, which occurs after Arthropoda and Nematoda in TOP3B and before those groups in Type IIA. In any case, the trees placed Nematoda close to Arthropoda, supporting the Ecdysozoan hypothesis (Aguinaldo et al. 1997), and grouped Annelida and Mollusca supporting the existence of the Spiralia or Lophotrochozoa group, defined as including those having lophophores and trochophore larvae (Nosenko et al. 2013). The sequencing of additional topoisomerases from representative species of these groups will help to elucidate the phylogenetic relationships among these two major taxonomic groups.

The evolution of topoisomerases is particularly interesting to study in chordates due to the existence of paralogues in Type IIA (*TOP2A* and *TOP2B*) and Type IB (*TOP1* and *TOP1MT*) (Forterre et al. 2007; Pommier et al. 2022). Previous works suggested that vertebrates have paralogues for these two topoisomerases [e.g., (Forterre et al. 2007; Zhang et al. 2004)], but did not include analyses of basal chordates (Cephalochordata and Tunicata or Urochordata), remaining some doubts about the possible existence of paralogues in these groups. Our extensive searches fail to identify paralogues in basal chordates, therefore proving that the formation of paralogues in Type IIA and Type IB topoisomerase is indeed a feature exclusive to vertebrates.

The phylogenies built for the three topoisomerase types all agree in placing the Cephalochordata (lancelets, such as *Branchiostoma* sp.) at the base of chordates (Chapters 3, 4 and 5). The phylogenies for Type IIA (TOP2A and TOP2B) and Type IB (TOP1 and TOP1MT) included sequences from Tunicata (Urochordata), and agree in placing them together with Vertebrata to form a sister group known as Olfactores (Delsuc et al. 2006; Putnam et al. 2008; Satoh et al. 2014). Tunicata showed long branches in both phylogenies explained by their high rate of molecular evolution (Delsuc et al. 2006; Tsagkogeorga et al. 2010). Whitin Tunicata, *Oikopleura dioica* stood out by its extremely long branches in the phylogenies (Chapters 4 and 5). Therefore, topoisomerases also support the idea that *O. dioica* is the fastest evolving metazoan recorded so far (Berna and Alvarez-Valin 2014; Denoeud et al. 2010).

The Type IIA (*TOP2A* and *TOP2B*) and Type IB (*TOP1* and *TOP1MT*) paralogues were identified in cyclostomes (jawless vertebrates) and gnathostomes (jawed vertebrates). However, the two families of topoisomerases revealed a different evolutionary history (Chapters 4 and 5). We found that *TOP2A* and *TOP2B* paralogues from Cyclostomata cluster together in a separate branch from all Gnathostomata paralogues. A different phylogeny was obtained for *TOP1* and *TOP1MT*, in which one of the Cyclostomata paralogues cluster with Gnathostomata *TOP1MT* and the other radiated at the same time than Gnathostomata *TOP1*. In both cases, the duplication event that generate the paralogues is most likely related with the different rounds of tetraploidization (1R and 2R) that occurred early in vertebrate evolution (Ohno 2013; Smith and Keinath 2015; Van de Peer et al. 2009). However, the formation of *TOP2A* and *TOP2B* paralogues in Cyclostomata was independent from the formation of paralogues in Gnathostomata (Chapter 4), while *TOP1* and *TOP1MT* paralogues seem to have the same origin in both groups (Chapter 5). Type IIA and Type IB topoisomerases are functionally different and had possibly different origins (Forterre et al. 2007), therefore there is no reason to expect them to have similar evolutionary patterns. It is well known that different genes may have distinct genealogies due to the intricate dynamics of the evolutionary processes. What seems to be concordant is the fact that these duplications occur in the evolutionary transition to vertebrates, where many of these duplicated genes played a role in the increasing biological complexity and in the development of new features (Van de Peer et al. 2009). The topoisomerase paralogues may provide an advantage for specific activities in different tissues, as the case of TOP2A and TOP2B (Capranico et al. 1992), or different genomes as for TOP1 and TOP1MT (Zhang et al. 2007), possibly during different developmental stages.

Within Gnathostomata, we found well-supported monophyletic groups for the main classes: Chondrichthyes, Actinopteri, Amphibia, Aves, Reptilia and Mammalia. However, we found some disagreements within the well-established Chordate phylogeny [e.g. (Amemiya et al. 2013)]. The main disagreement was the radiation of Chondrichthyes (cartilaginous fish) after Actinopteri (ray-finned fish) in TOP2A, TOP2B and TOP1, although with low branch supports in ML and Bayesian trees. The coelacanth (*Latimeria chalumnae*) was at the root of the tetrapods in all trees, with exception of TOP3B. Aves and Reptilia not always formed a monophyletic group. These discrepancies may result from the high conservation of topoisomerases which may provide a limited amount of phylogenetic information for resolution in close related groups, as those within Gnathostomata.

6.2. Topoisomerases evolve under strong purifying selection

Proteins evolve by the interaction between mutational processes and selective forces acting at the molecular level. Among selective forces, purifying or negative selection works to eradicate mutations that are harmful for the organism, while positive selection accelerates the rate of amino acid replacements that may confer an advantage for the organism (Pál et al. 2006). We found strong signs of purifying selection acting in all analysed topoisomerases (Chapters 3, 4 and 5). The nonsynonymous/synonymous substitution rate ratio (*dN/dS*) was less than 1 for all topoisomerases (Table 6.2), meaning that these genes accumulate less non-synonymous mutations than expected (Kryazhimskiy and Plotkin 2008). For example, Chordates showed *dN/dS* values from 0.076 (TOP3B) to 0.307 (TOP1MT). The strong role of purifying selection shaping topoisomerases was expected considering that they act under rigorous functional or structural requirements in their interactions with the DNA molecule and other proteins. Moreover, the activity of most topoisomerases is crucial to all mitotic tissues (Wang 2002), and any disruption on their precise activities has drastic consequences for the organism. It is therefore expectable that most amino acid changes in critical topoisomerase domains will not be accepted, as they will affect the proper protein activity.

Table 6.2. Nucleotide diversity and selective pressure acting in topoisomerases from chordates. The details of the analyses are described in chapters 3, 4 and 5.

Type	Gene	Number of sequences	Nucleotide diversity	Global *dN/dS*	Number of positively selected sites
IA	*TOP3B*	235	0.21	0.076 [0.073–0.078]	1
IIA	*TOP2A*	159	0.16	0.238 [0.235–0.242]	0
	TOP2B	159	0.14	0.156 [0.152–0.158]	1
IB	*TOP1*	74	0.06	0.154 [0.147–0.162]	0
	TOP1MT	74	0.09	0.307 [0.300–0.315]	0

We found that the two pairs of paralogues (*TOP2A*/*TOP2B* and *TOP1*/*TOP1MT*) had higher *dN/dS* values than *TOP3B,* the only gene in our dataset with no paralogue. The estimates are consistent with the idea that paralogues are subject to weaker purifying selection than single-copy genes (Kondrashov et al. 2002; Scannell and Wolfe 2008). Paralogues may provide some redundancy that allow for more amino acid replacements, but under the restricted limits of variability observed in these proteins. Different topoisomerases may be functionally compatible and have some redundant activity

(Sakaguchi and Kikuchi 2004; Zhang et al. 2014). This hypothesis can be supported in the future with studies of other topoisomerase paralogues, such as the bacterial type IIA topoisomerase paralogs (*gyrase* and *topo IV*) (Gellert et al. 1976; Kato et al. 1990).

It has been suggested that paralogues may exhibit asymmetric rates of sequence evolution [e.g., (Conant and Wagner 2003; Steinke et al. 2006; Van de Peer et al. 2001)]. We found that topoisomerase paralogues in vertebrates are examples of duplicated genes that do not diverge from the ancestral state at a similar rate (Chapters 4 and 5). Although all of them evolved under strong negative selection, one of the paralogues has a slightly elevated rate of molecular evolution (Table 6.2). This pattern was also clear in the different length of the paralogue branches in the phylogenetic trees, explained by the different rates of sequence evolution (Chapters 4 and 5). These findings are compatible with the idea that one of the paralogues was recruited to perform a new biological role. In fact, gene duplications contribute to the evolution of genomic novelty by providing raw materials for developing new functions (Chen et al. 2013; Conant and Wolfe 2008). For example, TOP1MT has adapted to resolve topological problems in the mitochondria, contributing to mtDNA integrity and mitochondrial transcription (Baechler et al. 2019; Zhang et al. 2001).

The topoisomerase paralogues may be regarded as cases of neofunctionalization (one copy retains the original function whereas the other copy develops a novel function), subfunctionalization (the two copies develop different functions from each other and compensate for the entire function of the ancestral gene) or specialization (the two copies evolve different functions, and their overall function is also different from the ancestral gene) (Conant and Wolfe 2008; Ohno 2013). The conserved activities and structures of topoisomerase across the animal kingdom suggest that the functions of both paralogues do not significantly differ from the ancestral genes (i.e., they continue to solve topological problems), possibly excluding the specialization hypothesis. The case of TOP1 and TOP1MT seems to be more related with the subfunctionalization model. While vertebrates have specialized paralogues for nuclear and mitochondrial activities, invertebrates have a single Type IB topoisomerase. In yeast, silencing the single Type IB topoisomerase gene suppresses their mitochondrial activity, suggesting that the same gene is functional in nuclei and mitochondria (Tua et al. 1997; Wang et al. 1995). In vertebrates, TOP1 does not seem to work in mitochondria, therefore suggesting that none of the original duplicated genes retained the exact original function (i.e., working on both nuclear and mitochondrial genomes), therefore excluding the neofunctionalization model.

Regarding TOP2A and TOP2B, we found that TOP2B is under stronger selective constraints than TOP2A (Table 6.2). This result could be explained by the specialized role of TOP2B in the genetic programming of postmitotic cells, in particular nerve growth and brain development (Bollimpelli et al. 2017; Lyu et al. 2006; Lyu and Wang 2003), which could impose additional constrains to its evolution. The active role of TOP2B in brain development may suggest it acquired a new role, possibly related with the more complex nervous systems of vertebrates. Nevertheless, further studies are necessary to determine the precise functions of topoisomerases in invertebrates. Such information will help to understand the difference in the activities of the paralogues regarding the ancestral state.

We detected positive selection in two sites of TOP3B and TOP2B when comparing several chordates (Table 6.2 and Chapters 3 and 4). In both cases, the sites under positive selection are located in peripheral regions of the protein, far away from the active core domains interacting with DNA (Chapters 3 and 4). In mammals, an additional positively selected site was also detected in the C-terminal domain of TOP2B (Chapter 4). All these cases seem to suggest that topoisomerase sites under positive selection may be relevant for interaction with other proteins. It is possible that diversification of these proteins interacting with topoisomerase in different lineages may impose different selective pressures on the interacting sites. In other words, mutations in topoisomerase sites that will result in better interactions with other cellular partners (that may have mutated meanwhile) would be positively selected. Several proteins are known to interact with topoisomerases [e.g., (Goto-Ito et al. 2017; Mankouri and Hickson 2007; Shykind et al. 1997; Uusküla-Reimand et al. 2016)], and certainly many more are still to be discovered. A better understanding of these protein-protein interactions will elucidate the possible role of the detected sites under positive selection.

6.3. Variation in topoisomerases amongst modern and archaic humans

The genomes of Neanderthals and Denisovans have been sequenced and compared with that of modern humans, revealing that they contributed genetically to some of the present-day human populations (Green et al. 2010; Reich et al. 2010). Comparative genomic analyses have also revealed that only a small number of proteins have missense changes among modern and archaic humans (Kuhlwilm and Boeckx 2019), some of which are believed to underlie phenotypic differences between the groups. To the best of our knowledge, the topoisomerases of modern and archaic humans have never been compared before. Using the Neanderthals and Denisovans sequence data

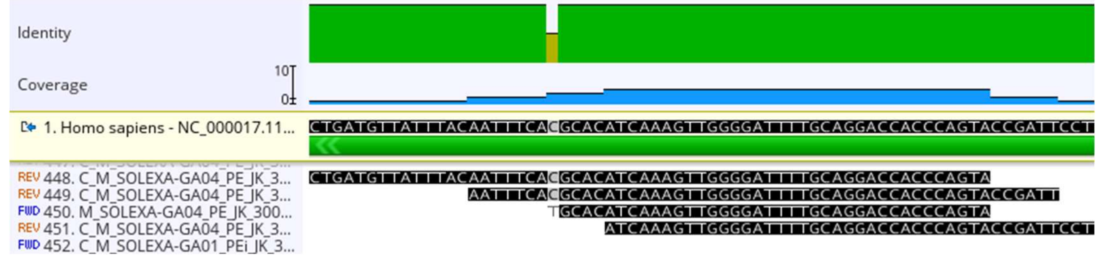
available in the UCSC Genome Browser, we were able to identify eight polymorphisms in the coding regions of the studied topoisomerases (Chapters 3, 4 and 5; Table 6.3). Three of these polymorphisms led to the replacement of an amino acid in the Neanderthals and one amino acid in modern humans.

Table 6.3. Mutational events responsible for missense mutations in topoisomerase genes among archaic and modern humans.

Gene	Mutational event	Amino acid replacement	Species	Pairwise Identity (%) in Chordata
TOP2A	T>C	Asp>Gly	*H. sapiens*	36
	C>A	Ala>Ser	*Neanderthal*	49.2
TOP1MT	T>C	Gln-Arg	*Neanderthal*	92.7
	G>T	Gln-Lys	*Neanderthal*	84.7

It is well known that sequencing errors or nucleotide misincorporations caused by DNA damage occur when analysing ancient DNA samples (Logsdon et al. 2020; Overballe-Petersen et al. 2012). In particular, C to T and G to A substitutions represent the majority of misincorporations, particularly at the ends of the sequence reads (Briggs et al. 2007). Such pattern was clear when inspecting the alignment of sequence reads against the human reference sequence, as shown in Figure 6.2.

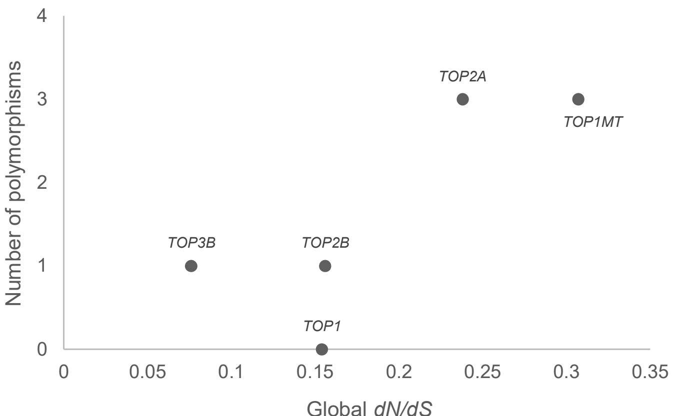
Figure 6.2. Example of a mutation in a sequence read not considered as a polymorphism in our studies. The alignment represents a section of the human *TOP2A* reference sequence with aligned sequence reads from Denisovan.



By this reason, we were particularly cautious when assuming that a polymorphic position really occurred amongst modern and archaic humans (Chapters 3, 4 and 5). Only mutations occurring in most aligned reads and not including those at the ends of the reads were considered as polymorphisms. Moreover, the cases we identified do not belong to the typical C>T or G>A type, suggesting they are not sequencing errors or the result of damage in DNA. In any case, the sequencing of additional Neanderthal and Denisovan samples will allow us to confirm if these polymorphisms were fixed among archaic and modern humans.

The topoisomerases with more polymorphisms were those that showed a higher *dN/dS* ratio (*TOP2A* and *TOP1MT*), *i.e.*, those under a weaker purifying selection (Figure 6.3). More polymorphisms were detected in those genes (*TOP2A* and *TOP1MT*) with a higher diversity among chordates. The observed pattern also supports the credibility of the identified polymorphisms in archaic humans by showing an expected pattern of variability.

Figure 6.3. Relationship between the number of polymorphisms in archaic humans and the strength of selection in topoisomerase coding sequences.



Despite the propensity of these polymorphisms to occur in genes with higher *dN/dS* ratios, the evolutionary conservation of the site where the mutation was identified varied considerably (Table 6.3). The two mutations in *TOP1MT* were found in two conserved amino acids positions (pairwise identity of 92.7% and 84.7% among chordates) close to each other in the linker region (Chapter 5). The *TOP2A* mutations occurred in variable sites (36% and 49.2% pairwise identity) in the C-terminal domain. In both cases, the amino acid replacements were not in active sites of the proteins responsible for the topoisomerase interactions with the DNA molecule. Instead, the polymorphisms occurred in regions that may interact with other cellular components. It was recently noted that TOP2A was the protein with human-lineage high-frequency missense changes that had more connections in a measure of protein-protein interactions (Kuhlwilm and Boeckx 2019). The authors speculate that changes in TOP2A might function as an interaction hub in modifications of the cell division complex. It remains to be tested if such changes result in any alteration of the cell cycle of other cellular processes where TOP2A participates.

6.4. Evolution of topoisomerase domains

All topoisomerases analysed by us have relatively similar levels of conservation, despite their different origins and varied structural features (Chapters 3, 4 and 5; Table 6.4). The percentage of pairwise identity for the complete protein varied from 55.9% in Type IIA to 70.5% in Type IB when considering the multiple sequence alignments with Metazoan species. When only considering chordates, the values only vary from 76% in TOP2A to 87.1% in TOP3B. It should be taken into account that a different number of sequences were used in the alignments of different topoisomerases, which may affect the comparison of results. In any case, the values are relatively similar and demonstrate the high conservation of topoisomerases across different animal phyla (Forterre et al. 2007).

Table 6.4. Degree of conservation in topoisomerases. The percentage of pairwise identity was obtained in an alignment with metazoans or with chordates alone.

			Metazoa		Chordata					
Type	**Protein**	**Human reference length (aa)**	**Pairwise Identity (%)**	***n***	**Pairwise Identity (%)**	***n***	**Most conserved domain**	**Pairwise Identity (%)**	**Most variable domain**	**Pairwise Identity (%)**
IA	**TOP3B**	862	67.6	265	87.1	144	IV	97.1	II	78.1
IIA	**TOP2A**	1531	55.9	389	76	105	WHD	94	C-terminal domain	45.6
	TOP2B	1626			85.8	125	Linker	96.4	C-terminal domain	69.8
IB	**TOP1**	765	70.5	161	83.6	48	Hinge	98.3	N-terminal domain	64.1
	TOP1MT	601			77	48	C-terminal domain	85.4	N-terminal domain	53.3

Although with an overall high conservation when considering the complete protein, topoisomerase domains can vary considerably within the same protein. In general, the most conserved domains are those that bind and cleave the DNA molecule, the catalytic region of the protein. These regions are also conserved when eukaryotic and bacterial topoisomerases are compared, sometimes they are the only domains that are shared between the different kingdoms of life (Champoux 2001). Among the different families, Type IA and Type IB are unrelated in terms of protein sequence, structure and DNA cleavage mechanism. However, Type IA and Type IIA topoisomerases have some mechanistic similarities using an “enzyme-bridging” process to interact with the DNA (Berger et al. 1998; Pommier et al. 2022). Both types share the presence of a Topoisomerase-Primase subdomain (TOPRIM), a conserved catalytic domain also found in DnaG-type primases, OLD family nucleases and RecR proteins (Aravind et al. 1998). Our analyses demonstrate that the TOPRIM is among the most conserved

regions of Type IA and Type IIA topoisomerases in chordates: 86.3% of identity in TOP3B (where is designated subdomain I), 89.1% in TOP2A and 94.4% in TOP2B (Chapters 3 and 4). The conservation decreases in when all Metazoan species are compared (68.2% in Type IA and 76.9% in Type IIA), but are still among the most conserved regions. It would be interesting to verify if the observed variability in the TOPRIM domain results in structural differences. The determination of new topoisomerase 3D structures in other animal species or the use of AI system for predictions of a structure from amino acid sequences, such as the AlphaFold (Jumper et al. 2021), could provide the necessary information. In fact, the AlphaFold database already includes topoisomerase structures for several animal species (e.g., *Danio rerio*, *Gallus gallus*, *Mus musculus*, *Rattus norvegicus*, *Sus scrofa*) that will be of great relevance for future studies on the evolution and function of these proteins.

We also found that the short stretches of amino acids at the junction between major domains can also be highly conserved. For instance, the most conserved domains of TOP2B and TOP1 are regions connecting major domains (Table 6.4.). In the case of TOP2B, we identified the Linker connecting the ATPase and the TOPRIM domains as the most conserved region (Chapter 4). In the case of TOP1, the Hinge connecting the capping and catalytic modules was the most conserved region (Chapter 5). The short length of these regions (27 amino acids in the Linker and five in the Hinge) may facilitate high pairwise identity values, but are undoubtedly highly conserved, as noted before (Takahashi et al. 2022). The conservation at interdomain linkers has been found in other proteins [e.g., (Banjade et al. 2015; Tauber and Fischle 2015)]. These observations suggest that structural flexibility does not always tolerate sequence variability. Nonetheless, we also identified short linker regions that are variable, such as the CTD Linker in TOP2A (Chapter 4), indicating that even within the same protein linker regions can have different variability patterns.

The most notorious differences in the conservation of topoisomerase domains occur in their terminal regions. The concept of N- or C-terminal domain has nothing to do with the structural role of those domains in a protein, as their name only depends on their location in the primary sequence. Still, the terminal regions often contain signal peptide sequences to guide the movement and interaction of the protein with other cellular components [e.g.,(Enz 2012; Hansen et al. 2006; Kumar and Thompson 2003)]. We found that topoisomerase terminal domains can have significantly different levels of conservation. Both terminal regions of TOP3B are poorly conserved, while Type IIA (TOP2A and TOP2B) has a poorly conserved C-terminal domain (Chapters 3 and 4; Table 6.4).

In some cases, these poorly conserved terminal regions include localization signals or other motifs that are relatively conserved despite being within a variable region (Chapters 3, 4 and 5). For example, we found that TOP3B C-terminal CXXC zinc finger motifs, known to participate in protein-DNA and protein–protein interactions (Wilson et al. 2000), are highly conserved in metazoans (Chapters 3). However, we observed a different scenario in the putative regulatory regions of the C-terminal domain of TOP2A and TOP2B. In this case, the ChT domain (Lane et al. 2013) and a bipartite nuclear localization signal (Mirski et al. 1997) were found poorly conserved. This observation suggest that different species may use different localizations signals or other regulatory motifs. Further experimental studies are necessary to prove the use of these regions *in vivo*. Our multiple sequence alignments also identified several well conserved regions in the N- or C-terminal domains of topoisomerases that may indicate some structural or regulatory relevance, perhaps for the interactions with other proteins or the DNA (Chapters 3, 4 and 5). These regions are good candidates for future experimental studies.

Type IB (TOP1 and TOP1MT) are characterized by having a variable N-terminal domain and conserved C-terminus, which is the opposite of the other topoisomerases studied here. The variable N-terminal domain of Type IB not even exists in the homologous archaeal, bacterial, and viral topoisomerases (Takahashi et al. 2022). On the contrary, the C-terminal domain is highly conserved participating in the active core of the enzyme activity. The C-terminal domain folds together into a globular structure that includes the active-site Tyr723 that bounds to the 3'-phosphate of DNA at the site of nicking and forms the “cleavage complex” (Takahashi et al. 2022). Overall, our analyses suggest that unknown interactions or structural dynamics are still to be discovered that could explain the observed patterns of conservation found across animals. The study of additional topoisomerases from diverse animal groups will help to elucidate such features.

6.5. Mutations in topoisomerases causing disease

The essential role of topoisomerases in cleaving the DNA strands to resolve topological problems makes them prone to cause severe cellular damages when not working properly. The formation of irreversible topoisomerase cleavage complexes (TOPccs) could be a relevant cause of harmful genomic lesions (Pommier et al. 2022; Pommier et al. 2016). On a first glance, one should expect a large number of diseases associated with mutations in topoisomerase genes. However, the number of clinical manifestations associated with topoisomerase malfunctions is small, being restricted to a maximum of

four described symptoms for each gene, as recently reviewed (Pommier et al. 2022). We searched for all topoisomerase mutations associated with human inherited disease in *The Human Gene Mutation Database* and also found a small number of reported cases, with most genes only having a single case (Table 6.5). It could be precisely due to the critical role of topoisomerases in so many cellular processes that diseases associated with these enzymes are so rare. It has been suggested that most disease genes are non-essential (Goh et al. 2007). Most topoisomerases could be considered essential, as mouse knockouts are lethal or cause severe disease (Pommier et al. 2022). Therefore, topoisomerases corroborate the idea that essential genes tend to cause developmental abnormalities rather than adult disease.

Table 6.5. List of published topoisomerase mutations responsible for human inherited disease available at The Human Gene Mutation Database (HGMD; http://www.hgmd.cf.ac.uk).

Type	Protein	Missense/ nonsense	Gross deletions	Gross insertions/ duplications	Phenotype
IA	TOP3B	3	4	2	Intellectual disability Schizophrenia Autism spectrum disorder Cognitive impairment and facial dysmorphism
IIA	TOP2A	-	1	-	Congenital heart disease
	TOP2B	1	-	-	Global development delay and intellectual disability
IB	TOP1	1	-	-	Autism
	TOP1MT	2	-	-	Autism spectrum disorder Leukaemia

Despite their rarity, we analysed the few cases with reported point mutations in these genes (Chapters 3, 4 and 5). The *TOP3B* was the gene with more single point mutations or rearrangements associated with diseases (9 cases). We have found that one of these *TOP3B* mutations (R472Q) previously associated with schizophrenia (Xu et al. 2012) is variable in animals (Chapter 3). A different pattern was observed in *TOP2B* mutations associated with human disorders (impaired B-cell development, hearing loss or neurodevelopmental disease), which were almost all found in conserved sites (Chapter 4). It is often assumed that amino acid changes causing disease occur in evolutionary conserved sites (Miller and Kumar 2001), a rule that seems to be followed by *TOP2B.* The reason for such different levels of conservation remains unknown, but may be related with the way the alteration causes the disease. The functional assessment of the mutation has not been determined yet. Our data also suggest that using animal models to study some of the topoisomerase mutations causing disease is not feasible, as the sites are not conserved.

It is interesting to notice that some of the clinical manifestations associated with topoisomerase malfunctions are related with neurodevelopmental and cognitive processes (Pommier et al. 2022) (Table 6.5). There is a high diversity of nervous systems across animals or even across chordates. Therefore, the observation that sites associated with mutations causing disorders typical of complex nervous systems are not conserved is not unexpected. The cellular processes that evolved in certain species are probably unique and therefore require proteins activities and interactions that are also unique.

The potential of topoisomerases to cause cellular damage has been elegantly used to develop anticancer and antibacterial therapeutic approaches (Pommier et al. 2010). If properly induced by drugs, irreversible topoisomerase cleavage complexes (TOPccs) can be used in cancer treatment. The cells accumulate double stranded DNA breaks and undergo apoptosis, which is beneficial when occurring in cancer cells, which often overexpress topoisomerase genes and are highly dependent on topoisomerase activities. Type IA topoisomerases (TOP3A and TOP3B) are not yet clinical therapeutic targets, but Type IB topoisomerases (e.g., camptothecins) and Type IIA topoisomerases (e.g., etoposide) are important targets for anticancer and antibacterial drugs (Bailly 2000; Pommier 2006).

In addition to the problem of the side effects of chemotherapy, the use of topoisomerase poisons is limited by the emergence of drug-resistant mutations (Chrencik et al. 2004; Cretaio et al. 2007; Saleem et al. 2000). The selective pressure of treatment favours mutations that allow cancer cells to avoid inhibitory drugs. We analysed the evolutionary conservation of topoisomerase sites known to be mutated and confer resistance to anticancer drugs (Chapters 4 and 5). Overall, amino acid replacements known to affect the efficiency of anticancer drugs are evolutionarily conserved (Table 6.6).

Table 6.6. Degree of conservation of topoisomerase sites with amino acid replacements known to affect the efficiency of anticancer drugs. The values were obtained in multiple sequence alignments with chordates.

Type	Protein	Number of sites		Pairwise Identity in Chordata (%)			
		Total	Differing in paralogues	Mean	Median	Min	Max
IIA	TOP2A	27	6	87.6	96.2	41.2	100.0
	TOP2B			95.5	98.4	40.0	100.0
IB	TOP1	9	1	93.7	100.0	66.3	100.0
	TOP1MT			91.3	100.0	62.8	100.0

The mean and median values of pairwise identity are near or above 90% in all topoisomerases with known drug-resistant mutations. It is clear that the amino acid sites that change to 'resist' the action of the topoisomerase poisons are functionally relevant

and therefore conserved. In order to properly act, the drugs have to interfere with regions of the protein that are relevant to the formation of the TOPccs. Therefore, the resistance to the drugs occurs in those sites that are affected by it, so there is a good correspondence between the regions of the active sites and those conferring resistance to anticancer drugs.

Our comparative analyses also revealed a few drug-resistance sites that differ between the paralogues (Chapters 4 and 5; Table 6.6). The side effect of chemotherapy regimens is sometimes the result of the inhibition of one of the paralogues, such as leukaemias caused by TOP2B interference (Azarova et al. 2007; Cowell et al. 2012). There is therefore interest in developing targeting drugs specific for the TOP2A paralogue, avoiding the side effect of blocking TOP2B. The differences between the paralogues identified in our work can provide a basis for future improvements in anticancer drugs.

6.6. References

Aguinaldo AMA, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA (1997) Evidence for a clade of nematodes, arthropods and other moulting animals. Nature 387:489-493.

Alkan C, Sajjadian S, Eichler EE (2011) Limitations of next-generation genome sequence assembly. Nature methods 8:61-65.

Amemiya CT et al. (2013) The African coelacanth genome provides insights into tetrapod evolution. Nature 496:311-316.

Aravind L, Leipe DD, Koonin EV (1998) Toprim—a conserved catalytic domain in type IA and II topoisomerases, DnaG-type primases, OLD family nucleases and RecR proteins. Nucleic acids research 26:4205-4213.

Azarova AM, Lyu YL, Lin C-P, Tsai Y-C, Lau JY-N, Wang JC, Liu LF (2007) Roles of DNA topoisomerase II isozymes in chemotherapy and secondary malignancies. Proceedings of the National Academy of Sciences 104:11014-11019.

Baechler SA, Dalla Rosa I, Spinazzola A, Pommier Y (2019) Beyond the unwinding: role of TOP1MT in mitochondrial translation. Cell Cycle 18:2377-2384.

Bailly C (2000) Topoisomerase I poisons and suppressors as anticancer drugs. Current medicinal chemistry 7:39-58.

Banjade S, Wu Q, Mittal A, Peeples WB, Pappu RV, Rosen MK (2015) Conserved interdomain linker promotes phase separation of the multivalent adaptor protein Nck. Proceedings of the National Academy of Sciences 112:E6426-E6435.

Berger JM, Fass D, Wang JC, Harrison SC (1998) Structural similarities between topoisomerases that cleave one or both DNA strands. Proceedings of the National Academy of Sciences 95:7876-7881.

Berna L, Alvarez-Valin F (2014) Evolutionary genomics of fast evolving tunicates. Genome biology and evolution 6:1724-1738.

Bollimpelli VS, Dholaniya PS, Kondapi AK (2017) Topoisomerase IIβ and its role in different biological contexts. Archives of Biochemistry and Biophysics 633:78-84.

Briggs AW et al. (2007) Patterns of damage in genomic DNA sequences from a Neandertal. Proceedings of the National Academy of Sciences 104:14616-14621.

Capranico G, Tinelli S, Austin CA, Fisher ML, Zunino F (1992) Different patterns of gene expression of topoisomerase II isoforms in differentiated tissues during murine development. Biochimica et Biophysica Acta (BBA)-Gene Structure and Expression 1132:43-48.

Champoux JJ (2001) DNA topoisomerases: structure, function, and mechanism. Annual review of biochemistry 70:369-413.

Chen S, Krinsky BH, Long M (2013) New genes as drivers of phenotypic evolution. Nature Reviews Genetics 14:645-660.

Chrencik JE, Staker BL, Burgin AB, Pourquier P, Pommier Y, Stewart L, Redinbo MR (2004) Mechanisms of camptothecin resistance by human topoisomerase I mutations. Journal of molecular biology 339:773-784.

Conant GC, Wagner A (2003) Asymmetric sequence divergence of duplicate genes. Genome research 13:2052-2058.

Conant GC, Wolfe KH (2008) Turning a hobby into a job: how duplicated genes find new functions. Nature Reviews Genetics 9:938-950.

Cowell IG et al. (2012) Model for MLL translocations in therapy-related leukemia involving topoisomerase IIβ-mediated DNA strand breaks and gene proximity. Proceedings of the National Academy of Sciences 109:8989-8994.

Cretaio E, Pattarello L, Fontebasso Y, Benedetti P, Losasso C (2007) Human DNA topoisomerase IB: structure and functions. The Italian journal of biochemistry 56:91-102.

Delsuc F, Brinkmann H, Chourrout D, Philippe H (2006) Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature 439:965-968.

Denoeud F et al. (2010) Plasticity of animal genome architecture unmasked by rapid evolution of a pelagic tunicate. Science 330:1381-1385.

Didier G, Landès C, Hénaut A, Torrésani B (2021) Four billion years: the story of an ancient protein family.

Enz R (2012) Structure of metabotropic glutamate receptor C-terminal domains in contact with interacting proteins. Frontiers in molecular neuroscience 5:52.

Forterre P, Gadelle D (2009) Phylogenomics of DNA topoisomerases: their origin and putative roles in the emergence of modern organisms. Nucleic Acids Research 37:679-692.

Forterre P, Gribaldo S, Gadelle D, Serre M-C (2007) Origin and evolution of DNA topoisomerases. Biochimie 89:427-446.

Gellert M, Mizuuchi K, O'Dea MH, Nash HA (1976) DNA gyrase: an enzyme that introduces superhelical turns into DNA. Proceedings of the National Academy of Sciences 73:3872-3876.

Goh K-I, Cusick ME, Valle D, Childs B, Vidal M, Barabási A-L (2007) The human disease network. Proceedings of the National Academy of Sciences 104:8685-8690.

Goto-Ito S, Yamagata A, Takahashi TS, Sato Y, Fukai S (2017) Structural basis of the interaction between Topoisomerase IIIβ and the TDRD3 auxiliary factor. Scientific reports 7:1-10.

Green RE et al. (2010) A draft sequence of the Neandertal genome. science 328:710-722.

Hansen JC, Lu X, Ross ED, Woody RW (2006) Intrinsic protein disorder, amino acid composition, and histone terminal domains. Journal of Biological Chemistry 281:1853-1856.

Jumper J et al. (2021) Highly accurate protein structure prediction with AlphaFold. Nature 596:583-589.

Kato J-i, Nishimura Y, Imamura R, Niki H, Hiraga S, Suzuki H (1990) New topoisomerase essential for chromosome segregation in E. coli. Cell 63:393-404.

Kondrashov FA, Rogozin IB, Wolf YI, Koonin EV (2002) Selection in the evolution of gene duplications. Genome biology 3:1-9.

Kriventseva EV, Kuznetsov D, Tegenfeldt F, Manni M, Dias R, Simão FA, Zdobnov EM (2018) OrthoDB v10: sampling the diversity of animal, plant, fungal, protist, bacterial and viral genomes for evolutionary and functional annotations of orthologs. Nucleic Acids Research 47:D807-D811.

Kryazhimskiy S, Plotkin JB (2008) The population genetics of dN/dS. PLoS genetics 4:e1000304.

Kuhlwilm M, Boeckx C (2019) A catalog of single nucleotide changes distinguishing modern humans from archaic hominins. Scientific reports 9:1-14.

Kumar R, Thompson EB (2003) Transactivation functions of the N-terminal domains of nuclear hormone receptors: protein folding and coactivator interactions. Molecular Endocrinology 17:1-10.

Lane AB, Giménez-Abián JF, Clarke DJ (2013) A novel chromatin tether domain controls topoisomerase IIα dynamics and mitotic chromosome formation. Journal of Cell Biology 203:471-486.

Logsdon GA, Vollger MR, Eichler EE (2020) Long-read human genome sequencing and its applications. Nature Reviews Genetics 21:597-614.

Lyu YL, Lin C-P, Azarova AM, Cai L, Wang JC, Liu LF (2006) Role of Topoisomerase IIβ in the Expression of Developmentally RegulatedGenes. Molecular and cellular biology 26:7929-7941.

Lyu YL, Wang JC (2003) Aberrant lamination in the cerebral cortex of mouse embryos lacking DNA topoisomerase IIβ. Proceedings of the National Academy of Sciences 100:7123-7128.

Mankouri HW, Hickson ID (2007) The RecQ helicase–topoisomerase III–Rmi1 complex: a DNA structure-specific ‘dissolvasome’? Trends in biochemical sciences 32:538-546.

Miller MP, Kumar S (2001) Understanding human disease mutations through the use of interspecific genetic variation. Human molecular genetics 10:2319-2328.

Mirski SE, Gerlach JH, Cummings HJ, Zirngibl R, Greer PA, Cole SP (1997) Bipartite nuclear localization signals in the C terminus of human topoisomerase IIα. Experimental cell research 237:452-455.

Nosenko T et al. (2013) Deep metazoan phylogeny: when different genes tell different stories. Molecular phylogenetics and evolution 67:223-233.

Ohno S (2013) Evolution by gene duplication. Springer Science & Business Media,

Overballe-Petersen S, Orlando L, Willerslev E (2012) Next-generation sequencing offers new insights into DNA degradation. Trends in biotechnology 30:364-368.

Pál C, Papp B, Lercher MJ (2006) An integrated view of protein evolution. Nature reviews genetics 7:337-348.

Pommier Y (2006) Topoisomerase I inhibitors: camptothecins and beyond. Nature Reviews Cancer 6:789-802.

Pommier Y, Leo E, Zhang H, Marchand C (2010) DNA topoisomerases and their poisoning by anticancer and antibacterial drugs. Chemistry & biology 17:421-433.
Pommier Y, Nussenzweig A, Takeda S, Austin C (2022) Human topoisomerases and their roles in genome stability and organization. Nature Reviews Molecular Cell Biology:1-21.
Pommier Y, Sun Y, Shar-yin NH, Nitiss JL (2016) Roles of eukaryotic topoisomerases in transcription, replication and genomic stability. Nature reviews Molecular cell biology 17:703-721.
Putnam NH et al. (2008) The amphioxus genome and the evolution of the chordate karyotype. Nature 453:1064-1071.
Reich D et al. (2010) Genetic history of an archaic hominin group from Denisova Cave in Siberia. Nature 468:1053-1060.
Sakaguchi A, Kikuchi A (2004) Functional compatibility between isoform α and β of type II DNA topoisomerase. Journal of cell science 117:1047-1054.
Saleem A, Edwards TK, Rasheed Z, Rubin EH (2000) Mechanisms of resistance to camptothecins. Annals of the New York Academy of Sciences 922:46-55.
Satoh N, Rokhsar D, Nishikawa T (2014) Chordate evolution and the three-phylum system. Proceedings of the Royal Society B: Biological Sciences 281:20141729.
Scannell DR, Wolfe KH (2008) A burst of protein sequence evolution and a prolonged period of asymmetric evolution follow gene duplication in yeast. Genome research 18:137-147.
Shykind BM, Kim J, Stewart L, Champoux JJ, Sharp PA (1997) Topoisomerase I enhances TFIID-TFIIA complex assembly during activation of transcription. Genes & development 11:397-407.
Smith JJ, Keinath MC (2015) The sea lamprey meiotic map improves resolution of ancient vertebrate genome duplications. Genome research 25:1081-1090.
Steinke D, Salzburger W, Braasch I, Meyer A (2006) Many genes in fish have species-specific asymmetric rates of molecular evolution. BMC genomics 7:1-18.
Takahashi DT et al. (2022) Topoisomerase I (TOP1) dynamics: conformational transition from open to closed states. Nature communications 13:1-11.
Tauber M, Fischle W (2015) Conserved linker regions and their regulation determine multiple chromatin-binding modes of UHRF1. Nucleus 6:123-132.
Treangen TJ, Salzberg SL (2012) Repetitive DNA and next-generation sequencing: computational challenges and solutions. Nature Reviews Genetics 13:36-46.
Tsagkogeorga G, Turon X, Galtier N, Douzery EJ, Delsuc F (2010) Accelerated evolutionary rate of housekeeping genes in tunicates. Journal of molecular evolution 71:153-167.
Tua A, Wang J, Kulpa V, Wernette C (1997) Mitochondrial DNA topoisomerase I of Saccharomyces cerevisiae. Biochimie 79:341-350.
Uusküla-Reimand L et al. (2016) Topoisomerase II beta interacts with cohesin and CTCF at topological domain borders. Genome biology 17:1-22.
Van de Peer Y, Maere S, Meyer A (2009) The evolutionary significance of ancient genome duplications. Nature Reviews Genetics 10:725-732.
Van de Peer Y, Taylor JS, Braasch I, Meyer A (2001) The ghost of selection past: rates of evolution and functional divergence of anciently duplicated genes. Journal of molecular evolution 53:436-446.
Wang J, Kearney K, Derby M, Wernette CM (1995) On the Relationship of the ATP-Independent, Mitochondrial-Associated DNA Topoisomerase of Saccharomyces cerevisiae to the Nuclear Topoisomerase I. Biochemical and biophysical research communications 214:723-729.
Wang JC (2002) Cellular roles of DNA topoisomerases: a molecular perspective. Nature reviews Molecular cell biology 3:430-440.
Wilson TM, Chen AD, Hsieh T-s (2000) Cloning and Characterization of DrosophilaTopoisomerase IIIβ: relaxation of hypernegatively supercoiled DNA. Journal of Biological Chemistry 275:1533-1540.
Xu B et al. (2012) De novo gene mutations highlight patterns of genetic and neural complexity in schizophrenia. Nature genetics 44:1365-1369.
Zhang H, Barceló JM, Lee B, Kohlhagen G, Zimonjic DB, Popescu NC, Pommier Y (2001) Human mitochondrial topoisomerase I. Proceedings of the National Academy of Sciences 98:10608-10613.
Zhang H, Meng L-H, Pommier Y (2007) Mitochondrial topoisomerases and alternative splicing of the human TOP1mt gene. Biochimie 89:474-481.

Zhang H, Meng LH, Zimonjic DB, Popescu NC, Pommier Y (2004) Thirteen-exon-motif signature for vertebrate nuclear and mitochondrial type IB topoisomerases. Nucleic acids research 32:2087-2092.
Zhang H, Zhang Y-W, Yasukawa T, Dalla Rosa I, Khiati S, Pommier Y (2014) Increased negative supercoiling of mtDNA in TOP1mt knockout mice and presence of topoisomerases IIα and IIβ in vertebrate mitochondria. Nucleic acids research 42:7259-7267.

Chapter 7
Conclusion

7. Conclusion

The tools of evolutionary biology are important to understand the function and complexity of proteins. The long-term divergence of proteins is limited by the constraints imposed to maintain specific molecular functions and structures. This feature is particularly relevant for proteins that play essential cellular roles, such as DNA topoisomerases. Our comparison of topoisomerases across the animal kingdom revealed significant differences that can teach us a lot on how these proteins operate. For example, the identification of conserved regions within highly variable protein domains can be used to guide future experimental work for the detection of possible cellular interactions. Alternatively, such regions could relate to yet to be discovered structural features that are not evident on current structural models. These models are snapshots that try to reconstruct the 3D shape of the proteins, but are limited by making predictions outside their native environment. Perhaps the future merging of the artificial intelligence tools with evolutionary information will provide a much better picture of the dynamic structures of topoisomerases.

But not only conservation patterns across species can be useful for understanding topoisomerases activities. Our detection of variable regions across some putatively relevant protein domains or motifs can suggest that they are not relevant at all or that different species evolved different mechanisms that are still to be discovered. In this regard, the identification of positively selected sites in proteins that evolve under strong purifying selection should deserve the attention of future works.

Because evolution shapes the functions of proteins, the resulting diversity can be used to understand the evolution of the organisms where it occurred. In this regard, animals are a particularly interesting group to study this phenomenon. Even though they represent only a small fraction of the eukaryotic tree of life, they show a bewildering diversity of forms. The recent advances in sequencing methods and phylogenetic inference tools have led to a substantial advance toward the reconstruction of the tree of life. Our phylogenetic analyses place topoisomerases among the universal markers that could provide useful information for resolving deep phylogenies, at least within Metazoa. The use of topoisomerases together with other informative genes can reduce stochastic errors that are known to cause incorrect topologies. For instance, the Type IIA and IB paralogues can be particularly useful for a better understanding of the events that took place at the radiation of vertebrates. The different evolutionary histories of the two sets

of paralogues should be further explored when more genomes become available, particularly of the jawless fishes.

Gene duplications provide the raw material for the creation of novel functions, facilitating adaptive evolutionary innovations. Therefore, future investigations could try to determine if (or how) the duplication of topoisomerase genes may have contributed to the evolution of new developmental and morphological features in vertebrates. At a stricter evolutionary scale, we identified a few missense mutations amongst the topoisomerases of modern and archaic humans. These enzymes are not obvious candidates to explain the unique human features, as for example the *FOXP2* gene is for the evolution of language. However, the role of topoisomerases on cell division and neuronal differentiation may influence the development of the brain and cognitive traits. Although speculative, the differences that we identified here may have contributed to the distinctive traits of humans and Neanderthals, and should be tested in future experimental validations.

The discovery of compounds that convert DNA topoisomerases to DNA-damaging agents highlights the dual nature of these enzymes. Their action is crucial to solve the topological problems that come with DNA's double-helical structure, but by doing it, they pose a risk for creating vulnerable spots in the DNA, with potentially disastrous results. This latter feature can result in disorders caused by *de novo* or inherited mutations, which we found not always conserved across species. Our findings highlight the limitations of using animal models for understanding the pathogenesis of human genetic diseases, when there are genetic background differences within species that may influence the outcome of studies. The potential of topoisomerases to cause damage in DNA has been also used to fight cancer and bacterial infections in a clever way. However, the power of natural selection acting on the replicative cancer cells and bacteria eventually resulted in the emergence of mutations that resist the action of topoisomerase poisons. As perhaps expected, we found that these mutated sites are highly conserved across species, supporting their critical role in the activity of topoisomerases to cleave DNA. Perhaps more unexpected was the identification of several differences in drug-resistance sites between the topoisomerase paralogues. Assuming those sites are relevant for the action of the drugs, they can be used to design therapeutic approaches that will only poison one paralogue, reducing the side effects caused by inhibiting the other. These new approaches can rely on ‘smarter’ drugs that can suppress drug resistance and reduce side effects.

Overall, this dissertation contributes with new insights into the evolutionary trajectories of topoisomerases in animals and the selective forces that shaped them over millions of

years. Hopefully, our findings will help design new experimental approaches for a better comprehension of these fascinating proteins.