Étude structurale et fonctionnelle de la fidélité des ADN polymérases X spécialisées dans la réparation des cassures doubles brins programmées chez Paramecium tetraurelia

Antonin Nourisson

Sorbonne Université

École doctorale 515 : Complexité du Vivant

THÈSE

Pour obtenir le grade de

DOCTEUR DE SORBONNE UNIVERSITE

Biochimie et Biologie structurale

Étude structurale et fonctionnelle de la fidélité des ADN polymérases X spécialisées dans la réparation des cassures doubles brins programmées chez *Paramecium tetraurelia*

Par Antonin Nourisson

Dirigée par le Dr Marc Delarue

Présentée et soutenue publiquement le 15 avril 2024

Devant un jury composé de :

Dr Mireille Bétermier I2BC Gif-sur-Yvette	Rapporteure
Dr Bertrand Castaing CNRS Orléans, CBM	Rapporteur
Dr Ghislaine Henneke Ifremer Brest	Examinatrice
Dr Catherine Vénien-Bryan Sorbonne Université, IMPMC	Présidente du jury
Dr Marc Delarue Institut Pasteur	Directeur de thèse

Remerciements

Tout d'abord, je tiens à remercier les rapporteurs de ce manuscrit de thèse, le Dr Mireille Bétermier et le Dr Bertrand Castaing. Je remercie également le Dr Ghislaine Henneke et le Dr Catherine Vénien-Bryan d'avoir accepté de faire partie du jury de ma soutenance de thèse.

Les travaux décrits dans cette thèse ont été réalisés au sein de l'unité Architecture et Dynamique des Macromolécules Biologiques (anciennement Dynamique Structurale des Macromolécules) à l'Institut Pasteur. Les six derniers mois ont été financés par la Fondation pour la Recherche Médicale.

Je tiens à remercier le Dr Marc Delarue, directeur de l' unité, pour m'avoir fait confiance en 2020 à la suite de mon (court) stage de Master 2, et pour s'être battu pour que je puisse obtenir un financement pour ces trois ans et six mois de doctorat. Merci pour ce que vous m'avez apporté tout au long de ma thèse, et pour la rigueur que vous m'avez inculqué. Merci également de m'avoir donné l'opportunité de faire partie d'autres travaux, en étroite collaboration avec Dariusz, qui ont inspiré toute la première partie de mon manuscrit de thèse et des travaux associés, en éveillant ma curiosité pour les techniques de classification des protéines. Je vous dois beaucoup d'enseignements, scientifiques ou non, pendant ces trois ans et demi, et je vous en remercie chaleureusement.

Un très grand merci également au Dr Sophia Missoury, pour m'avoir accompagné de près pendant ces trois ans et demi. Merci pour tes conseils, tes coups de main et toutes les leçons que tu m'as transmises (et désolé de ne pas les avoir toutes écoutées ou suivies spontanément). Tu m'as beaucoup appris, et je pense être devenu un meilleur scientifique à ton contact.

Merci au Dr Dariusz Czernecki et au Dr Rémi Sieskind, qui ont participé à l'amélioration de ce manuscrit de thèse.

Je souhaite remercier également les personnes avec qui je partage ou ai partagé le laboratoire. Merci au Dr Ludovic Sauguet pour sa bonne humeur, ses blagues, mais aussi ses bons conseils et son soutien pendant la rédaction de cette thèse. Merci aussi pour ce conseil prodigué en 2019 à la suite d'un entretien pour un stage de M2 : « Tu devrais essayer de prendre contact avec Marc ». Merci au Dr Soizick Lucas-Staat, pour son extrême gentillesse, sa bienveillance et son écoute. Merci pour ton côté « maman » pour toute l'équipe, très rassurant

à tout moment. J'ai été très content de te transmettre mes connaissances sur les tests d'activité à ton arrivée, et cela fait un moment déjà que l'élève a dépassé le maître ! Merci au Dr Rémi Sieskind pour tous ses excellents conseils, pour sa gentillesse, et pour toutes les passionnantes discussions que nous avons pu avoir. Merci à mes trois co-doctorants, Marie Imbert, Eloi Vincent et Léo Betancurt, pour leur bonne humeur et toutes les discussions que nous avons pu avoir. Merci au Dr Markel Martinez pour son apprentissage de la langue française qui nous a permis à tous de nous demander « mais ça se dit vraiment ça ? » dès qu'Isciane trouvait une nouvelle expression. Merci justement à Isciane Commenge, pour sa gentillesse, pour nos accords commerciaux (d'échange de cônes et de boîtes de tubes Eppendorf), et pour toutes les discussions que nous avons pu avoir : je n'aurais pas pu rêver meilleure voisine de paillasse. Enfin, merci à Murielle Seif El Dahan, fraîchement arrivée au labo, pour ton soutien pendant la rédaction de cette thèse et tes conseils. Enfin, merci aux anciens de l'unité, en particulier Camille Samson, Dariusz Czernecki et Clément Madru, pour tout ce que vous m'avez appris, et pour m'avoir donné envie de rester dans cette unité. Je tiens également à remercier les personnes qui ont apporté leur soutien scientifique et technique à ces travaux, en particulier le Dr Ahmed Haouz, qui m'a offert un magnifique cadeau d'anniversaire pour mes 25 ans (mes premières données de diffraction !), mais aussi Patrick Weber et Cédric Pissis pour leur aide et leur gentillesse. Un grand merci à l'ensemble de la plateforme de Biophysique, en particulier Sébastien Brûlé, Bertrand Raynal, Sandrine Rosario, Sylviane Hoos et Maelenn Chevreuil.

J'ai une attention particulière pour ma famille et surtout mes parents. Merci à vous deux pour votre soutien depuis toujours et pendant ces 8 ans et demi d'études. Merci de n'avoir jamais cessé de croire en moi, même quand je n'y arrivais plus, et merci pour votre amour, que je ne vous rendrai jamais assez.

Merci à Olympe. Merci à toi pour ton sourire, pour tes idées plus folles les unes que les autres (que je suivrai toujours), pour ta patience, et pour ta compréhension. Désolé pour le temps passé à m'attendre, pour mon indisponibilité, et pour tout le reste. Merci pour ton adhésion à mes idées farfelues et à mon amour des jeux vidéos. J'ai très hâte de notre avenir ensemble.

Enfin, merci à toutes les personnes qui m'ont aidé, accompagné, de près ou de loin, pendant ces trois ans et demi. Sans vous, le résultat n'aurait pas été le même. Et merci à toutes les personnes qui m'ont permis d'arriver là où je suis actuellement, bien avant tout ça.

Table des matières

Liste des abréviations

aa : acide aminé
Abs : Absorbance
ADN : Acide Désoxyribonucléique
AP (site) : APurinique/ APyrimidique
ARN/RNA : Acide RiboNucléique
ARNm : ARN messager
ARNt : ARN de transfert
ATP : Adénosine TriPhosphate
BER : *Base excision Repair* (Réparation par Excision de Base)
BLAST : Basic Local Alignment Research Tool
BRCT : *BReast cancer Carboxy Terminal associated.*
BSA : *Bovine Serum Albumine*
C-ter : Carboxyl terminal
CDB : Cassure Double Brins
CLANS : *CLuster Analysis of Sequences*
CTP : Cytosine TriPhosphate
Da : Dalton (équivalent des g/mol)
db : double brin
dHJ : *double Holliday Junction*
DLS : *Dynamic Light Scaterring*
DNA PK(cs) : (sous unité catalytique de la) Protéine Kinase dépendante de l'ADN
dNTPs : désoxyNucléotides TriPhosphates
DO : densité optique
dRP : désoxyRibose Phosphate
DTT : DiThioThréitol
dU : désoxyuridine
EDTA : éthylènediaminetétraacétique
FA : *Fanconi Anemia* (anémie de Fanconi)
FAM : 6-Carboxyfluorescéine
FL : *Full-Length*
GF : Gel filtration (chromatographie par exclusion stérique)
GTP : Guanosine TriPhosphate
ICL : *Interstrand CrossLink* (liaison inter brins)
IES : *Internal Eliminated Sequence(s)* (Séquence(s) interne(s) éliminée(s))
iesRNA : ARN d'IES
Ig : Immunoglobuline
IPTG : isopropyl β-D-1-thiogalactopyranoside
KLD : Kinase, Ligase, DpnI
LB : *Lysogeny Broth*
Lig : ADN ligase
LR : *Long Range*
MAC : macronoyau
MALDI-TOF : *Matrix Assisted Laser Desorption Ionization – Time Of Flight*
Mb : mégabase (1 000 000 bases)
MIC : micronoyau
MMEJ : *Microhomology Mediated End Joining*
MMR : *MisMatch Repair* (réparation des mésappariements)
N-ter : Amino terminal
NCBI : *National Center for Biotechnology Information*
NER : *Nucleotide Excision Repair* (Réparation par Excision de Nucléotide)
NHEJ : *Non Homologous End Joining*
nt : nucléotide
PAE : *Predicted Aligned Error*
PAGE : *PolyAcrylamide Gel Electrophoresis* (Electrophorèse en gel de polyacrylamide)
PAXX : PAralogue de XRCC4 et XLF
pb : paire de bases
PCR : *Polymerase Chain Reaction*
PEG : PolyÉthylèneGlycol
Pgm : PiggyMac
PgmL : PiggyMac *Like*
PHP : Polymérase / Histidinol Phosphatase
pLDDT : *predicted Local Distance Difference Test*
PNKP : PolyNucléotide Kinase Phosphatase
Pol : Polymérase (ici, ADN polymérase)
PSI-BLAST : *Position-Specific Iterated BLAST*
RMSD : *Root-Mean-Square-Deviation*
NTPs : riboNucléotides TriPhosphates
RSS : *Recombination Signal Sequence* (Séquence signal de recombinaison)
Sar : *Stramenopiles-Alveolata-Rhizarians*
sb : simple brin
scnRNA : ARN scan
SD : *Sequence Determinant*
SDSA : S*ynthesis Dependant Strand Annealing* (hybridation de brins dépendante de la synthèse)
SOC : *Super optimal medium with catabolic repressor*
SR : *Short Range*
TCR : *T Cell Receptor* (Récepteur des lymphocytes T)
Tdt : Terminal Deoxynucleotidyltransférase
TEV : *Tobacco Etch Virus* (virus de la gravure du tabac)
TTP : Thymidine TriPhosphate
UV : UltraViolets
XLF : *XRCC4-like Factor*
XRCC4 : X-*ray Repair Cross-Complementing Protein 4*

Introduction

1 Avant-Propos

L'ADN (Acide DésoxyriboNucléique) est le principal support de l'information génétique au sein du vivant. Après la découverte de sa structure en 1953 par Watson et Crick sur la base d'expériences réalisées par Rosalind Franklin (Watson and Crick, 1953), l'ADN a longtemps été perçu comme une molécule très stable. Cependant, 21 ans plus tard, un des découvreurs de sa structure souleva la question des dommages qu'il peut rencontrer et des mécanismes lui permettant d'être réparé (Crick, 1974):

« We totally missed the possible role of [DNA] repair although, […] I later came to realize that DNA is so precious that probably many distinct repair mechanisms would exist. »

"Nous avons totalement ignoré le rôle possible de la réparation [de l'ADN], bien que [...] j'ai réalisé plus tard que l'ADN est si précieux qu'il existe probablement de nombreux mécanismes de réparation."

En effet, chez tous les organismes vivants (Friedberg, 2003), l'ADN subit des dommages: chaque cellule humaine fait face à plus de 10000 lésions par jour (Yousefzadeh *et al.*, 2021). Leurs causes sont variées, mais certaines sont dues au fonctionnement normal de la cellule, comme certaines erreurs de réplication qui peuvent incorporer des mésappariements (Ganai and Johansson, 2016). Certains processus métaboliques dans la cellule peuvent aussi former des espèces réactives de l'oxygène ou de l'azote, qui peuvent créer des sites abasiques, des cassures simple brin ou double brin, ou des mutations (Van Houten *et al.*, 2018). Des lésions peuvent aussi être crées par des agents extérieurs comme l'exposition à des rayonnements ionisants (rayons UV ou X) (de Gruijl, 1999), ou à des produits chimiques mutagènes (comme les hydrocarbures polycycliques aromatiques issus de la consommation de tabac) (Hakem, 2008). L'ADN subit également des dommages chimiques qui peuvent mener à des mutations si ils ne sont pas réparés (Boiteux *et al.*, 2017). Tous ces dommages peuvent mener à la mort des cellules par l'apparition de mutations, qui peuvent entrainer une cancérisation des cellules (Alhmoud *et al.*, 2020).

Comme proposé par Francis Crick en 1974, les cellules vivantes présentent des mécanismes spécialisés leur permettant de détecter et réparer tous ces types de dommages (1, 2)(Hoeijmakers, 2001). Il est fondamental d'en apprendre davantage sur les mécanismes mis en

place pour réparer les dommages de l'ADN : cela peut par exemple permettre de mettre au point de nouvelles approches thérapeutiques alliant radiothérapie et inhibition des mécanismes de la réparation de l'ADN des cellules cancéreuses (Feng *et al.*, 2022; Srivastava and Raghavan, 2015) ; ou de mieux comprendre d'autres maladies comme les immunodéficiences combinées sévères (Woodbine *et al.*, 2014). Une meilleure compréhension de ces mécanismes pourrait également aider au développement d'outils biotechnologiques (Chu *et al.*, 2015; Maruyama *et al.*, 2015; Robert *et al.*, 2015).

Dans cette thèse, je détaille mes travaux ayant visé à mieux comprendre le fonctionnement d'une forme particulière d'un système de réparation des cassures doubles brins.

L'introduction aux travaux que j'ai réalisés permettra dans un premier temps d'établir le contexte des dommages de l'ADN et des mécanismes qu'utilisent les cellules pour les réparer, en se focalisant sur les cassures double brins et le NHEJ (*Non Homologous End Joining* ou réparation des extrémités non homologues). Une seconde partie sera consacrée à un organisme particulier, *Paramecium tetraurelia*, qui présente une forme spécialisée de NHEJ, qui est au centre des travaux de cette thèse.

2 Les acides nucléiques

2.1 Les briques fondamentales des acides nucléiques : les nucléotides

Comme la plupart des composants moléculaires des cellules, les acides nucléiques sont des polymères. Ils sont formés par un assemblage de nucléotides. Ces derniers sont formés de trois parties : une base azotée qui varie selon le nucléotide, un sucre qui varie selon le type d'acide nucléique, et des groupements phosphates (Bowater and Gates, 2015) (Figure 1).

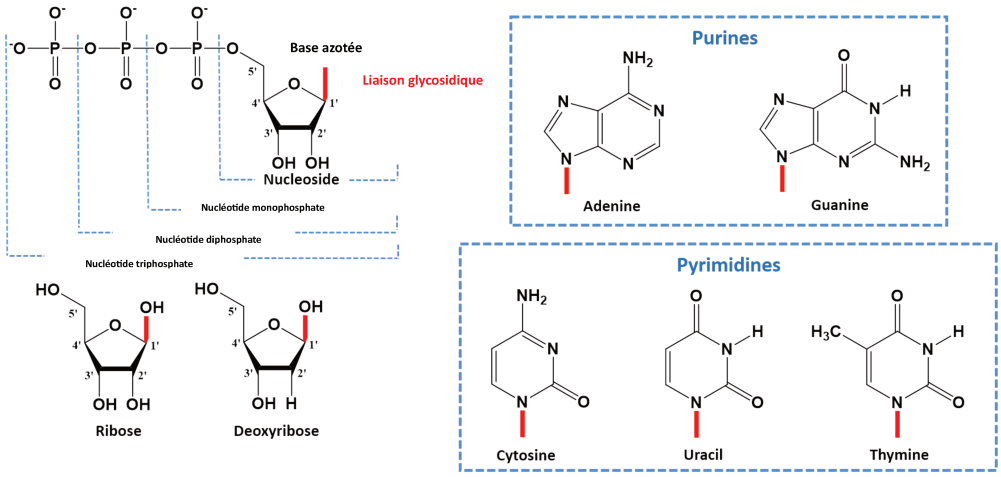


Figure 1 : Structure chimique des nucléotides
À gauche : structure d'un nucléotide, détaillant les différents groupements chimiques présents. La liaison osidique lie le pentose (en bas, ribose ou désoxyribose) à la base azotée. Le pentose est lié en 5' à un ou plusieurs groupements phosphates (en haut).
À droite : Structure des 5 bases azotées, séparées en fonction de leur famille (purine ou pyrimidine). La lettre R symbolise la liaison au pentose.
D'après https://wou.edu/chemistry/files/2019/07/nucleotide-structure.png

Il existe deux familles de bases azotées : les purines et les pyrimidines. Les premières sont l'adénine et la guanine, et les secondes sont la cytosine, la thymine et l'uracile (Bowater and Gates, 2015). Si un sucre est ajouté à la base azotée, cela forme un nucléoside. Il existe deux types de pentoses qui forment les acides nucléiques canoniques : le désoxyribose (qui forme l'ADN, acide **désoxyribo**nucléique) et le ribose (qui forme l'ARN, acide **ribo**nucléique). Leur seule différence est la présence d'un groupement 2'hydroxyl (OH) pour le ribose, là où le désoxyribose ne porte qu'un atome d'hydrogène (H) (Bowater and Gates, 2015). Enfin, c'est l'ajout d'un ou plusieurs groupements phosphate sur l'atome d'oxygène 5' du sucre qui forme un nucléotide, qui peut être mono-, di- ou triphosphate (Bowater and Gates, 2015).

2.2 L’assemblage des désoxynucléotides triphosphate : la structure de l’ADN

L’ADN est un polymère de désoxynucléotides monophosphate (dNMPs). Ces polymères se forment grâce à une liaison phosphodiester entre le phosphate le plus proche du sucre (appelé phosphate α) et l’oxygène en 3’ du nucléotide suivant (le sens de lecture étant par convention du nucléotide portant un groupement 5’P libre vers le nucléotide portant un groupement 3’OH libre) (Figure 2).

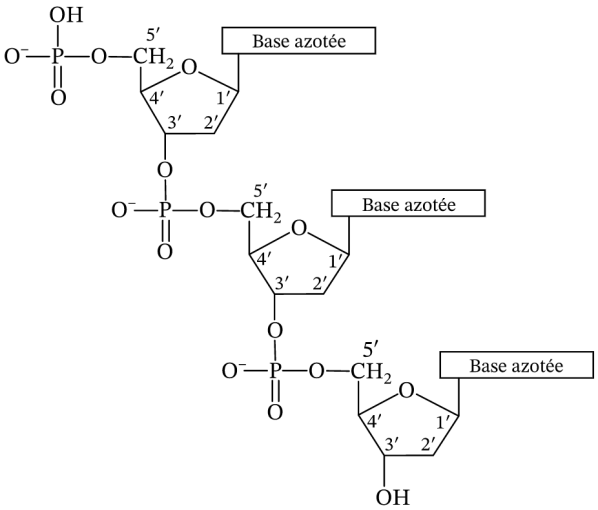


Figure 2 : Assemblage des nucléotides en chaîne d’acide nucléique.
Chaque nucléotide est relié via son groupement 3’OH au suivant, via son groupement 5’Phosphate : c’est la liaison phosphodiester.

Cependant, la structure des acides nucléiques est plus complexe, car ils peuvent être formés d’un seul brin ou de deux. Concernant l’ADN, James Watson et Francis Crick, sur la base d’expériences faites par Rosalind Franklin, ont montré en 1953 que l’ADN a canoniquement une structure en double hélice antiparallèle (Watson and Crick, 1953), c’est-à-dire que deux brins d’ADN se font face en sens inverse : l’un a une orientation 5’-3’, et l’autre est orienté de 3’ vers 5’. Les deux brins qui se font face doivent être hybridés, c’est-à-dire que les bases qu’ils portent doivent pouvoir s’associer *via* des liaisons hydrogène (H) spécifiques : l’adénine A s’associe avec la thymine T avec 2 liaisons H, et la guanine G et la cytidine C s’associent avec 3 liaisons H (Figure 3). Ces liaisons canoniques sont dites « Watson-Crick », mais il peut en exister d’autres (comme les liaisons « Hoogsteen » (Zhou *et al.*, 2015), les G quadruplex (Xu and Komiyama, 2023) ou les I-motifs (Gehring *et al.*, 1993)).

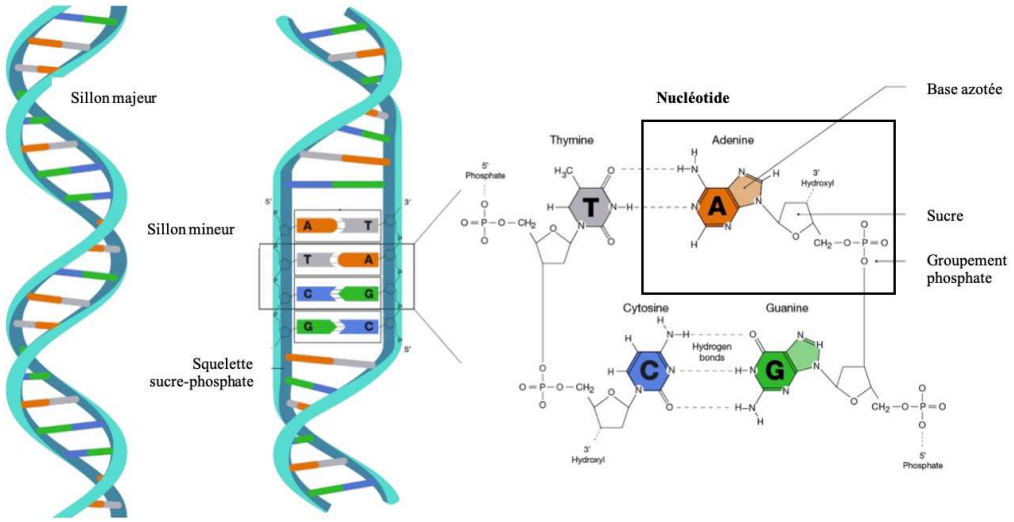
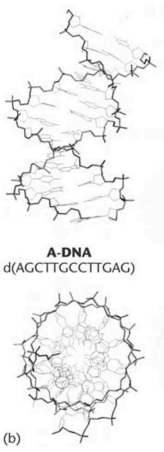
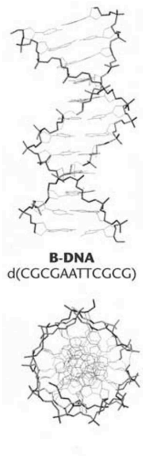
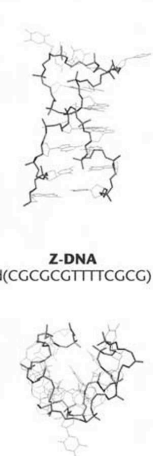


Figure 3 : Structure générale de l'ADN.
À gauche : schéma de la structure en double hélice B de l'ADN. Les deux brins d'ADN sont placés de façon antiparallèle selon leur polarité : l'un est orienté de 5' vers 3', et l'autre de 3' vers 5'. Ils sont liés entre eux par des liaison hydrogènes entre leurs nucléotides. Cette structure impose la formation de deux sillons dans l'ADN : un sillon majeur, de 11,7 Å, et un mineur de 5,7 Å. À droite : Les deux brins se lient entre eux par des liaison hydrogènes entre des paires de nucléotides. Les adénines forment deux liaisons hydrogènes avec les thymines, et les cytosines forment trois liaisons hydrogènes avec les guanines. D'après https://www.genome.gov/genetics-glossary/Deoxyribonucleic-Acid.

Il existe également plusieurs formes de doubles(1)-hélices d'ADN (Tableau 1) : hélices A, B (la plus fréquente), qui tournent à droite, ou Z, qui tourne à gauche. La forme de l'ADN varie selon les conditions (séquence, état de l'ADN, etc.) (Ussery, 2002).

Tableau 1 : Paramètres et structures des hélices A, B et Z de l'ADN.

	Hélice A	**Hélice B**	**Hélice Z**
Sens de l'hélice	Droite	Droite	Gauche
Motif répété	1 pb	1 pb	2 pb
Nombre de paires de bases par tour d'hélice	11 pb	10 pb	12 pb
Pas de l'hélice par tour	2,82 nm	3,32 nm	4,56 nm
Diamètre	2,3 nm	2 nm	1,8 nm
Structure	 **A-DNA** d(AGCTTGCCTTGAG) (b)	 **B-DNA** d(CGCGAATTCGCG)	 **Z-DNA** d(CGCGCGTTTTCGCG)

2.3 Les fonctions des acides nucléiques

Dans la cellule, les acides nucléiques peuvent avoir plusieurs fonctions. Cependant, leurs fonctions principales sont le stockage de l'information génétique, sa transmission aux cellules filles lors des divisions cellulaires, et l'utilisation de cette information génétique, qui permet aux cellules de synthétiser les protéines nécessaires à leur métabolisme (Minchin and Lodge, 2019).

Cependant, tandis que l'ADN se limite principalement à ces fonctions, l'ARN peut aussi avoir d'autres fonctions, parfois enzymatique : la maturation d'autres ARN (snRNA, RNase P, RNase MRP, snoRNA), la régulation de l'expression des gènes (eRNA, ceRNA, miRNA, siRNA, shRNA), la modification épigénétique du génome (lncRNA), la défense contre des pathogènes chez les bactéries (crRNA), et d'autres fonctions encore inconnues (Dai *et al.*, 2020).

3 Le dogme central de la biologie moléculaire

Dès 1958, et jusqu'à la publication de son article « *Central Dogma of Molecular Biology* » (Crick, 1970), Francis Crick a proposé que les acides nucléiques sont des éléments d'un processus permettant aux cellules de synthétiser les protéines utiles à leur métabolisme. Il a proposé que l'information contenue dans la séquence de l'ADN et de l'ARN était exprimée sous la forme de protéines, et que les fonctions de ces protéines étaient dépendantes de leur séquence, issue de la traduction des acides nucléiques. Plus tard, James Watson précisera ces mécanismes en indiquant que l'ADN permet la synthèse de l'ARN par la transcription, et que cet ARN est ensuite traduit en protéine. De nos jours, ces mécanismes ont largement été confirmés et affinés. Par exemple, on peut citer la découverte de la notion de transcription inverse, c'est-à-dire le passage de l'information de l'ARN vers l'ADN, qui existe chez certains virus (Temin, 1985).

Depuis l'énoncé de ce dogme en 1970 (Crick, 1970), chacune de ces étapes a été étudiée en détails afin de comprendre son fonctionnement. Également, tout le processus allant de la réplication de l'ADN à la synthèse de protéines actives a été complété par de très nombreuses connaissances (Ille *et al.*, 2022) sur l'expression des gènes, la régulation de cette expression, la modification de certains gènes en cours d'expression (édition d'ARNm, épissage alternatif,…), et la modification des protéines après leur expression (Figure 4).

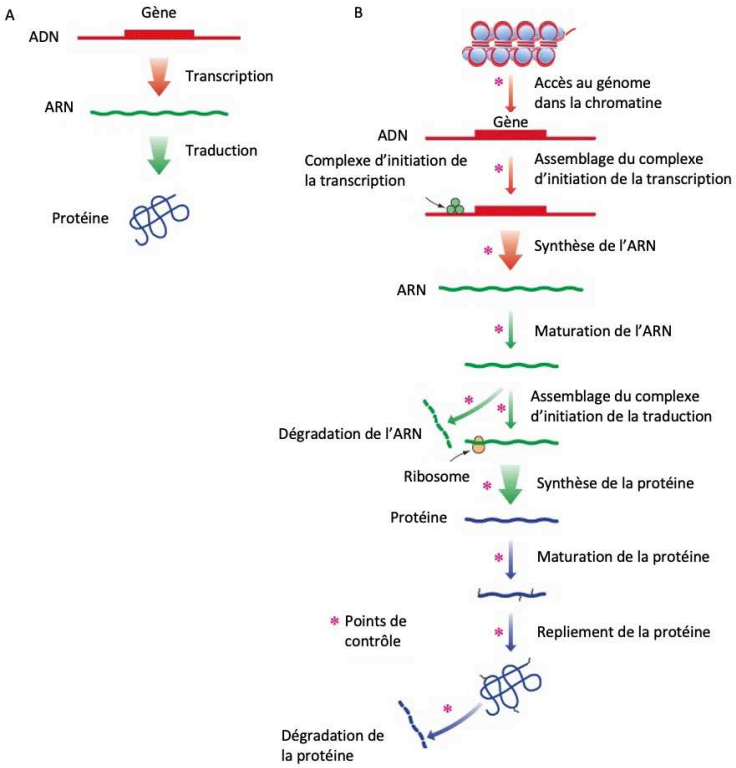


Figure 4 : Évolution du dogme de la biologie moléculaire.
A : vision originale du dogme de la biologie moléculaire. Ce dogme se compose de deux grandes étapes : Transcription d'un gène (ADN) en ARN, et traduction du transcrit ARNm en protéine.
B : vision actualisée du dogme de la biologie moléculaire. Le génome doit être décompacté avant de permettre tout assemblage de machinerie de transcription sur l'ADN. Cette machinerie, impliquant de nombreux partenaires, permet la synthèse d'un ARN, qui doit ensuite être maturé pour être traduit en protéine par le ribosome (qui est lui aussi un assemblage de nombreux éléments d'ARN et de protéines). L'ARN messager (ARNm) est également dégradé dans le même temps. Une fois la protéine synthétisée par le ribosome, elle peut être modifiée, et doit prendre son repliement final, qui lui permet de réaliser sa fonction. Les protéines sont enfin dégradées par des systèmes cellulaires.
D'après Brown TA. Genomes. 2nd ed. Oxford: Wiley-Liss; 2002.

3.1 L'expression des gènes : les protéines

Les protéines sont les actrices du métabolisme de la cellule, que ce soit d'un point de vue structurel ou fonctionnel (Whitford, 2013). Ces molécules sont composées de combinaisons de 20 briques élémentaires appelées acides aminés (aa), donc leur diversité repose sur la combinatoire des acides aminés utilisés dans leur séquence. Les protéines peuvent ainsi prendre de très nombreuses formes, comme le suggère l'origine du mot *protéine*, qu'on peut rapprocher du dieu grec Protée, dont le principal attribut serait de prendre différentes formes (Whitford, 2013). Au 24 Janvier 2024, la base de données UniProtKB/TrEMBL dénombrait 250322721 séquences protéiques, allant du plus petit peptide de 7 aa (A0A1Y8EMM2_HUMAN) à la plus grande protéine de 45 354 aa (A0A5A9P0L4_9TELE).

3.2 La synthèse des protéines

La séquence des protéines découle de celle des gènes contenus dans l'ADN des cellules. Ces gènes sont transcrits en ARN par une enzyme (une protéine « catalyseur » qui a pour fonction d'accélérer une réaction chimique) appelée ARN polymérase, ce qui donne un ARN messager (ARNm). Cet ARNm est une copie simple brin du gène, mais ici les thymines T sont remplacées par des uraciles U (Nelson *et al.*, 2008). C'est ensuite le ribosome, un assemblage de protéines et d'ARN, qui traduit l'ARNm en protéine en suivant le code génétique (Fox, 2010) (Figure 5). Celui-ci, qui varie légèrement entre les organismes, permet au ribosome d'associer chaque triplet de 3 nucléotides (appelée codon) à un acide aminé en particulier. C'est une famille d'ARN particulière, les ARN de transfert (ou ARNt, synthétisés par les aminoacyl-ARNt-synthétases (Delarue, 1995)) qui est chargée dans le ribosome de « lire » un codon et de transférer l'acide aminé correspondant au ribosome, qui le lie à l'acide aminé précédent, et ainsi de suite jusqu'à la lecture d'un codon « stop » qui entraine la libération de la chaine d'acides aminés (Nelson *et al.*, 2008).

Il devrait exister 64 codons possibles, puisqu'il y a 4 nucléotides possibles qui se combinent en triplets. Il devrait donc y avoir dans chaque organisme 64 ARNt, mais ce n'est pas exactement le cas : chez les eucaryotes, on en dénombre entre 41 et 55, dont certains peuvent reconnaitre plusieurs codons différents codant pour un même acide aminé (Goodenbour and Pan, 2006) : ils sont dits isoaccepteurs (Staehelin, 1973). Cependant, afin d'avoir suffisamment d'ARNt en permanence dans la cellule, les gènes codants pour ces ARNt sont présents en plusieurs copies : il existe 497 gènes d'ARNt qui sont identifiés chez l'humain (Lander *et al.*, 2001).

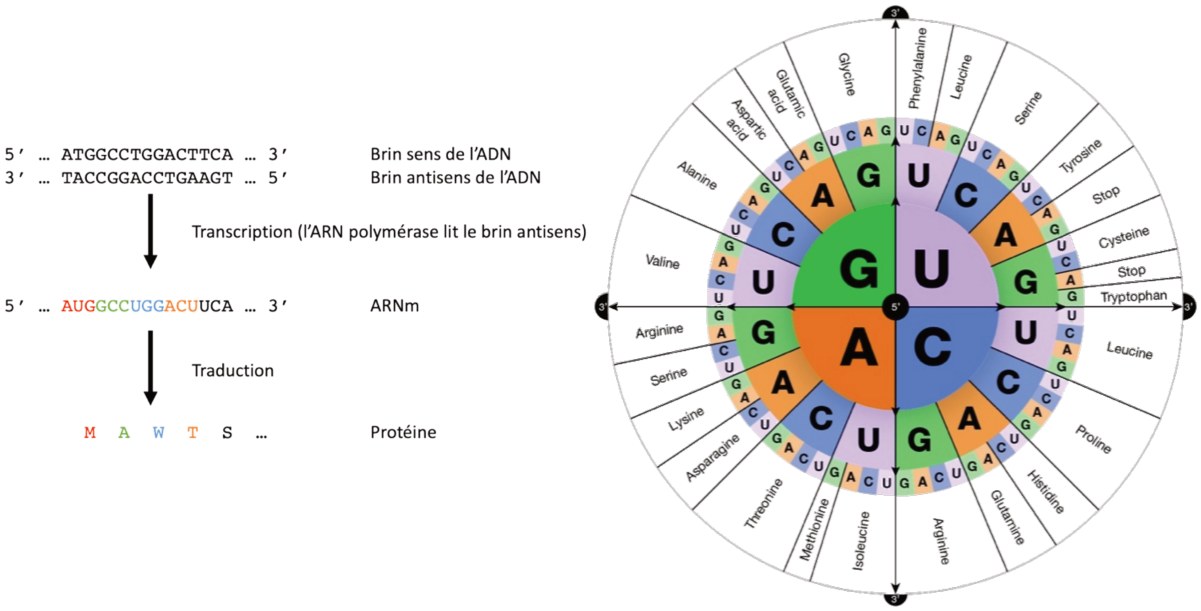


Figure 5 : Transcription et traduction de l'ADN en protéines
À gauche : résumé des étapes de la synthèse des protéines. Dans un premier temps, une ARN polymérase synthétise un ARNm à partir de l'ADN. L'étape de traduction, réalisée par le ribosome, traduit la séquence d'ARNm en séquence d'acides aminés : chaque suite de 3 nucléotides (appelée codon) permet de synthétiser un acide aminé ; c'est l'enchainement de ces acides aminés qui forme les protéines. À droite : code génétique. C'est ce code que suivent les ARNt dans le ribosome pour traduire chaque codon en acide aminé. Ici, le premier nucléotide du codon est au centre. A partir de ce nucléotide, le second nucléotide du codon peut être lu, et à partir de celui-ci, on peut lire le dernier nucléotide du codon. Ainsi, on arrive à l'extérieur du cercle avec une suite de 3 nucléotides : un codon. Correspondant à un acide aminé. Par exemple, le codon GAC permet de former un acide aspartique (ou aspartate). D'après https://www.genome.gov/genetics-glossary/Genetic-Code

3.3 Les acides aminés

Les acides aminés sont des molécules qui ont une base commune, dite amino-acide, qui leur permet d'être associés en chaînes par le ribosome (Nelson *et al.*, 2008), et une partie variable, dite chaine latérale. Comme leur nom l'indique, leur base commune est constituée d'un groupement amine et d'un groupement carboxyle reliés par un atome de carbone (carbone α ou Cα), lui-même relié à un atome d'hydrogène et à la chaine latérale de l'acide aminé (R)(Lopez and Mohiuddin, 2023). C'est cette chaîne latérale qui définit l'acide aminé et qui lui confère ses spécificités (Figure 6). Celle-ci est constituée d'atomes de carbone (appelés Cβ, Cγ, …) sur lesquels sont branchés des groupements chimiques particuliers. Selon leurs caractéristiques chimiques, les acides aminés peuvent être classés par groupes (figure 6). Tous ces acides aminés et les protéines qu'ils forment peuvent également être modifiés après la traduction : on parle de modifications post-traductionnelles (Khoury *et al.*, 2011).

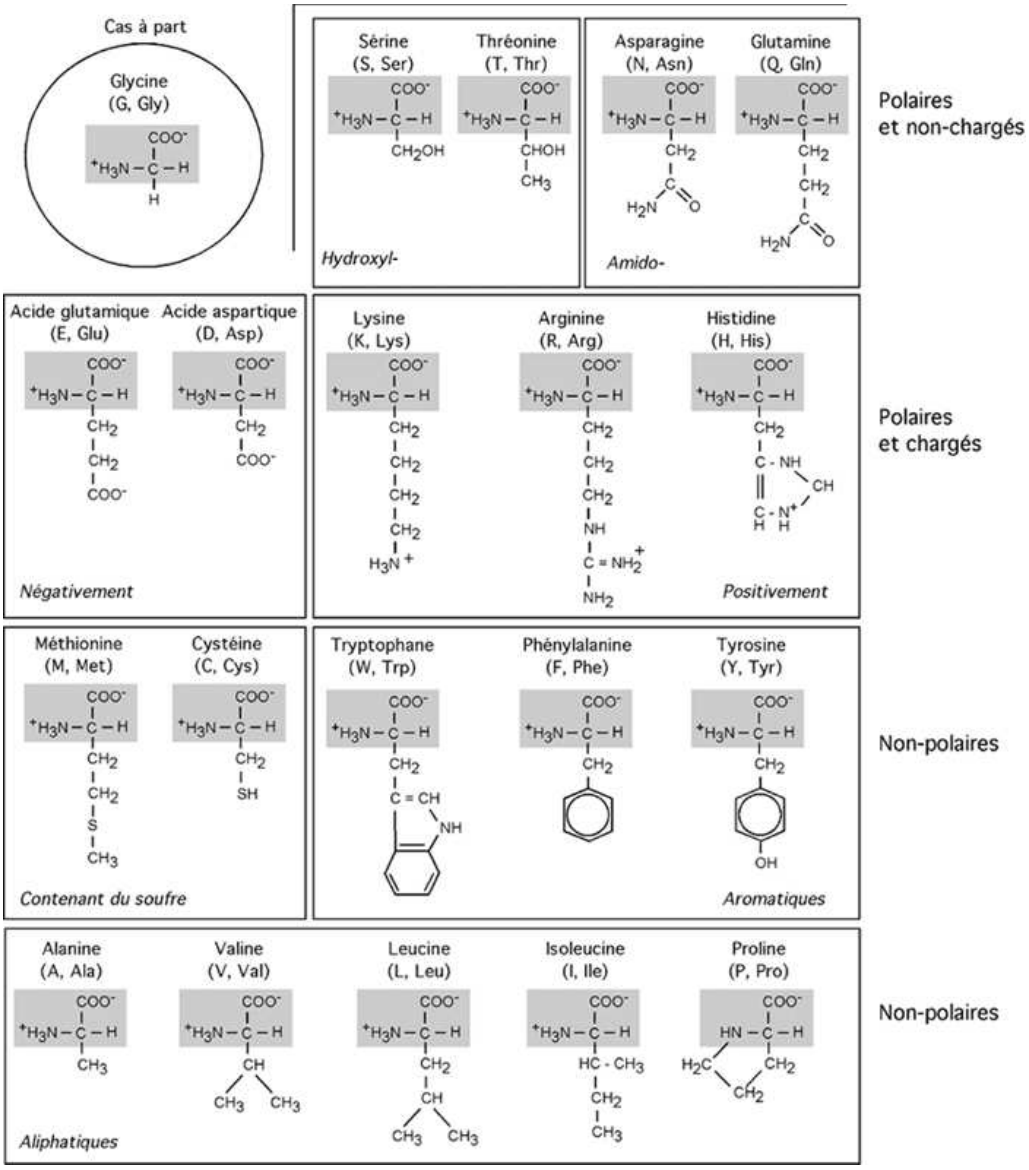


Figure 6 : Classification et structures des acides aminés.
Ici, les acides aminés sont représentés par leur nom et leur codes 1 lettre et 3 lettres, et organisés selon leurs caractéristiques physico chimiques.

3.4 La structure des protéines

Lors de la traduction, le ribosome lie les acides aminés entre eux *via* une réaction de condensation entre le groupement carboxyle du premier acide aminé et le groupement amine du suivant, et ainsi de suite (Figure 7).

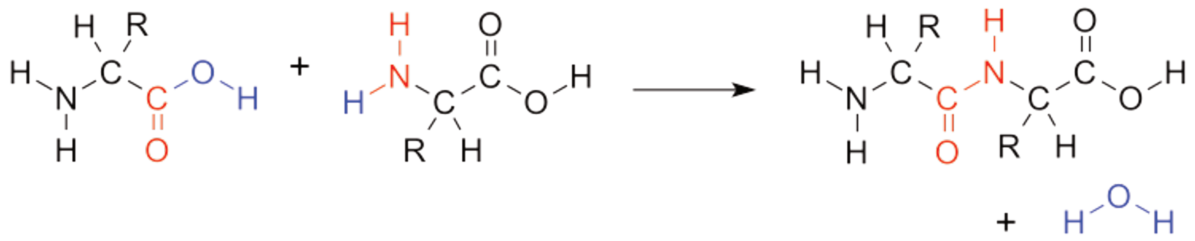


Figure 7 : Schéma de la réaction de condensation entre deux acides aminés.
Deux acides aminés (les chaînes latérales sont représentées par R) se lient par condensation de l'extrémité COOH de l'un et du groupement NH2 de l'autre. Cela forme une chaine de deux acides aminés polarisée d'une extrémité N-terminale (NH_2) à une extrémité C-terminale (COOH), ainsi qu'une molécule d'eau (H_2O, en bleu). Les deux résidus d'acides aminés sont liés par une liaison peptidique, en rouge. D'après https://chem.libretexts.org/Bookshelves/Introductory_Chemistry

Cette liaison plane, rigide et polaire est appelée liaison peptidique, et on appelle résidu un acide aminé impliqué dans cette liaison. Cet ensemble forme un squelette peptidique sur lequel sont branchées les chaines latérales des acides aminés (Whitford, 2013). Ainsi, tout comme les acides nucléiques ont des extrémités 5' et 3', les protéines ont des extrémités N terminales (N-ter) et C terminales (C-ter) (Whitford, 2013).

Cet enchainement séquentiel des acides aminés reliés par les liaisons peptidiques forme la structure primaire des protéines. La structure secondaire des protéines décrit quant à elle l'agencement local des acides aminés entre eux selon les rotations de la chaine polypeptidique (Ramachandran *et al.*, 1963). Cependant, toutes les interactions entre résidus ne sont pas possibles : la liaison peptidique étant rigide, seules deux types de rotations sont possibles : entre le Cα et l'azote (angle ϕ), et entre le Cα et le groupement carboxyle (angle φ)(Ramachandran *et al.*, 1963). De plus, l'encombrement stérique et les charges pouvant être présentes sur les chaines latérales ne permettent pas d'adopter toutes les conformations imaginables, mais quelques conformations sont favorisées et stabilisées par des liaisons hydrogène, ce qui donne des hélices (α, 3_{10}, π) et des brins β (Ramachandran *et al.*, 1963). La structure tertiaire des protéines résulte des interactions entre les structures secondaires formées localement dans la protéine (Godbey, 2022). Elle se forme par l'interaction des résidus de ces structures secondaires, *via* des interactions hydrophobes (qui vont rassembler au cœur de la structure les résidus hydrophobes et exposer les résidus polaires au solvant), des interactions électrostatiques, des liaisons de Van der Waals ou encore des liaisons covalentes comme les ponts disulfures, qui se forment sous certaines conditions entre les atomes de soufre des chaines latérales de deux cystéines (Godbey,

2022). Enfin la structure quaternaire décrit quant à elle l'interaction de plusieurs protéines pour former un complexe multimérique (Skipper, 2005).

La fonction des protéines découle de leurs structures : en effet, c'est l'agencement des résidus dans le site actif d'une enzyme qui définit son activité (Orengo *et al.*, 1999). Par conséquent, les quatre niveaux de structure sont fondamentaux pour que les protéines ou complexes protéiques puissent être actifs, c'est pourquoi il existe dans les cellules des mécanismes qui permettent d'aider les protéines à atteindre un repliement correct (Jackson, 2013). Cependant, dans la plupart des cas, la structuration des protéines *in vivo* découle des nombreuses interactions des résidus entre eux, qui vont former les structures secondaires puis tertiaires en permettant d'arriver à un minimum énergétique (Onuchic *et al.*, 1997). Cependant, il faut également noter l'existence de protéines ou de domaines protéiques intrinsèquement désordonnés mais fonctionnels (Uversky, 2011).

4 La rencontre entre protéines et ADN : le maintien de l'information génétique

On l'a vu précédemment, l'ADN est le support de l'information génétique des cellules vivantes. Cependant, au cours de la vie de la cellule, cet ADN traverse de nombreux processus, dont deux vont nous intéresser particulièrement ici : la réplication et la réparation.

4.1 La réplication de l'ADN chez les eucaryotes

Au sein des cellules eucaryotes, la réplication de l'ADN commence toujours en un point nommé origine de réplication (Nelson *et al.*, 2008). Il peut y en avoir une seule ou plusieurs sur un chromosome, et c'est en ce point que le complexe d'initiation s'assemble. Ce complexe recrute une hélicase, qui a pour rôle de séparer les brins d'ADN. Les tensions structurelles formées sont relâchées par une topoisomérase, et les brins séparés sont maintenus séparés par des protéines de fixation à l'ADN simple brin appelées RPA (*Replication Protein A*) chez les eucaryotes. De chaque côté de l'origine de réplication, une fourche de réplication progresse (l'ensemble des protéines impliquées est appelé le réplisome) (Pellegrini, 2023).

Chez les eucaryotes, au sein de chaque fourche de réplication, de l'ADN est synthétisé sur chacun des deux brins : l'ADN synthétisé sur le brin parental orienté de 3' vers 5' est

synthétisé directement de 5’ vers 3’ par une ADN polymérase, après qu’une ADN primase ait créé une amorce ; et l’ADN synthétisé sur le brin parental orienté de 5’ vers 3’ ne peut pas être synthétisé directement de 3’ vers 5’ (les ADN polymérases ayant besoin d’une extrémité 3’OH libre). Par conséquent, une ADN primase crée des amorces d’ARN sur le brin parental, et deux autres ADN polymérases synthétisent le second brin appelé brin tardif ou indirect, par morceaux de 150 à 200 paires de bases (pb) appelés les fragments d’Okazaki (Nelson *et al.*, 2008). Enfin, les amorces d’ARN sont éliminées par une enzyme portant une activité exonucléase, une ADN polymérase complète les trous, et une ligase rétablit les liaisons phosphodiesters (Nelson *et al.*, 2008) (Figure 8).

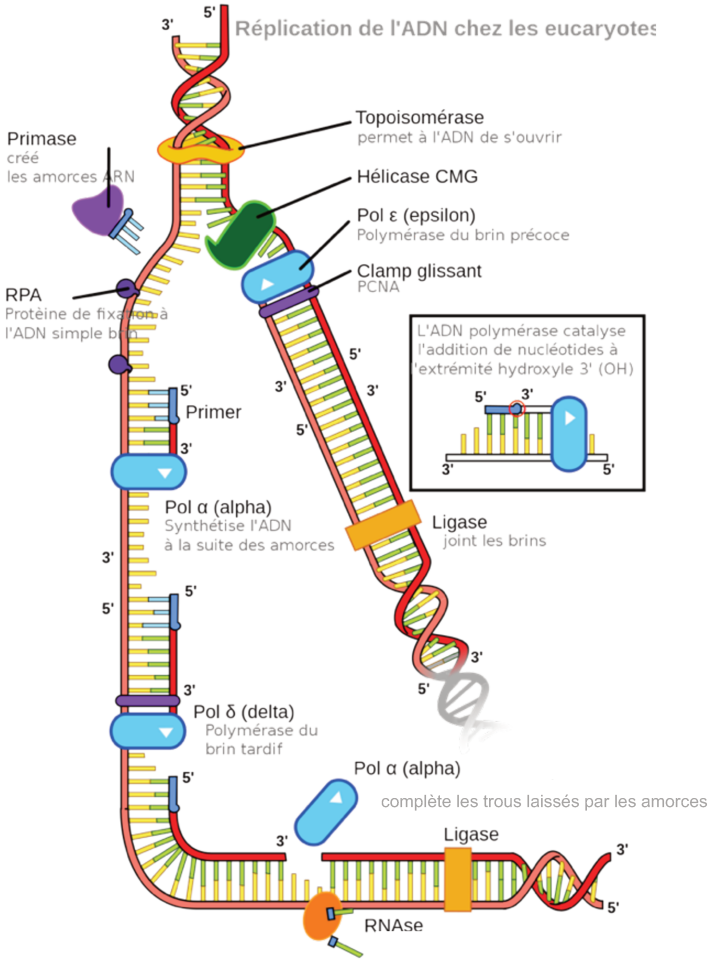


Figure 8 : La fourche de réplication chez les eucaryotes.
Chez les eucaryotes, la fourche de réplication est formée d’une topoisomérase et d’une hélicase, qui ouvrent la double hélice d’ADN et séparent les deux brins. Chaque ADN simple brin est stabilisé par le RPA. Sur le brin direct, une ADN primase synthétise une amorce, qui est ensuite étendue par une ADN polymérase. Sur le brin indirect, une ADN primase synthétise des amorces d’ARN sur l’ADN, pour permettre leur élongation par une ADN polymérase. Celle-ci synthétise l’ADN par morceaux de 150 à 200 pb, appelés fragments d’Okazaki, qui sont reliés entre eux par une ADN ligase.
D’après Wikipedia (https://fr.wikipedia.org/wiki/R%C3%A9plication_de_l%27ADN).

4.2 L'organisation spatiale du génome

4.2.1 La chromatine

Dans le noyau des cellules eucaryotes, qui mesure environ 5 à 10 µm, l'ADN doit être compacté, car sa longueur totale atteint environ 2 m chez l'homme (Cooper, 2000). Les structures protéiques permettant de compacter cet ADN s'appellent des histones, et l'ensemble forme la chromatine. La chromatine est formée d'éléments appelés nucléosomes, qui sont des unités couvrant 146 pb d'ADN, enroulé 1,65 fois autour d'un cœur d'histones. Il existe 5 types d'histones : H1, H2A, H2B, H3 et H4. L'histone H1 couvre l'ADN à l'entrée de chaque nucléosome, et chaque nucléosome est composé d'une association de 2 autres histones. L'ensemble incluant l'histone H1 s'appelle un chromatosome, et cette structure couvre 166 pb (Cooper, 2000) (Figure 9).

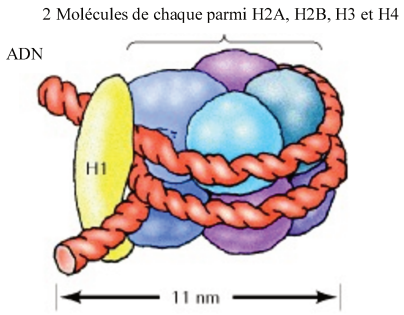


Figure 9 : Le chromatosome.
Le chromatosome est constitué de 2 molécules de chaque histone (H2A, H2B, H3, H4), autour desquelles s'enroulent un ADN long de 146 paires de bases.
D'après Cooper GM. The Cell: A Molecular Approach. 2nd edition. 2000. Chromosomes and Chromatin. Disponible sur: https://www.ncbi.nlm.nih.gov/books/NBK9863/

Ces structures forment des fibres de chromatine d'environ 10 nm de diamètre, composée de chromatosomes séparés d'environ 80 pb, et ces fibres permettent de réduire la longueur de l'ADN environ 6 fois (Cooper, 2000). La chromatine est elle aussi condensée en fibres de 30 nm

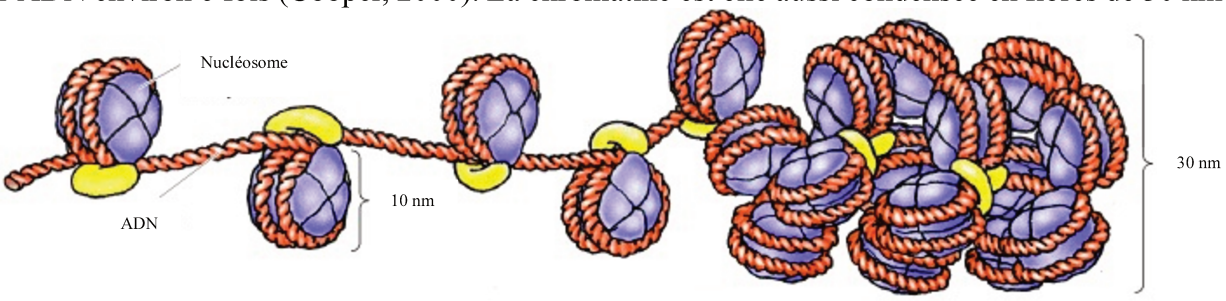


Figure 10 : La fibre de chromatine
L'empilement des nucléosomes de 10 nm forme une fibre de chromatine d'environ 30 nm de diamètre.
D'après Cooper GM. The Cell: A Molecular Approach. 2nd edition. 2000. Chromosomes and Chromatin. Disponible sur: https://www.ncbi.nlm.nih.gov/books/NBK9863

L'état de condensation de la chromatine varie dans le noyau en fonction de l'étape du cycle cellulaire (Cooper, 2000) :

- en interphase, 90% de la chromatine est décondensée sous forme de fibres de 30 nm, et 10% de cette euchromatine est sous forme de fibres de 10 nm, permettant l'expression des gènes. Les 10% restants sont très condensés, on parle alors d'hétérochromatine (Cooper, 2000) ;
- en mitose, tout le génome est très condensé (10000 fois), sous forme de chromosomes.

4.2.2 Les modifications d'histones

Pour permettre aux protéines ayant pour rôle d'interagir avec l'ADN de le faire, l'ADN doit être rendu accessible malgré sa condensation par la chromatine. Pour permettre cela, les histones présents peuvent être modifiés de nombreuses façons (Allfrey *et al.*, 1964) sur des queues chargées positivement situées en N-terminal, qui peuvent interagir avec d'autres nucléosomes. La modification des histones peut donc altérer la structure de la chromatine, en modifiant directement ces interactions, ou en recrutant des enzymes pouvant repositionner les nucléosomes. Les principales modifications d'histones sont l'acétylation, la phosphorylation et la méthylation.

Les méthylations d'histones sont faites sur les chaines latérales des lysines et arginines (Allfrey *et al.*, 1964). Contrairement aux autres modifications, elles ne modifient pas la charge des histones, mais les lysines peuvent être mono / bi / tri méthylées, et les arginines peuvent être mono méthylées ou bi méthylées. Les lysines méthyltransférases sont spécifiques du nombre de modifications à ajouter sur une lysine (Zhang *et al.*, 2003). Ces modifications sont reconnues par des protéines partenaires, pouvant avoir des rôles variés : remodelage de la chromatine et des nucléosomes (Sims *et al.*, 2005), recrutement d'autres modificateurs d'histones étant associés avec la prolifération cellulaire (Shi *et al.*, 2006). Leur présence est caractéristique de certains états de la chromatine : par exemple l'hétérochromatine constitutive (qui englobe les gènes inactivés de façon permanente) est caractérisée par un haut niveau de triméthylation sur la lysine 9 des histones H3 (H3K9me3) (Allfrey *et al.*, 1964), alors que l'hétérochromatine facultative est caractérisée par un enrichissement en H3K27me3 (Allfrey *et al.*, 1964). Ces modifications permettent entre autres le recrutement des enzymes de la transcription (Xiao *et al.*, 2003).

4.2.3 Autres éléments d'organisation du génome

En plus des éléments décrits ci-dessus, le génome des cellules eucaryotes en utilise d'autres lui permettant de se replier dans le noyau, comme la méthylation de l'ADN ou l'incorporation de variants d'histones (Talbert and Henikoff, 2017). Certains de ces éléments définissent ensemble le repliement des nucléosomes (Zhou *et al.*, 2011). *In fine*, le génome est organisé dans le noyau en compartiments regroupant des gènes avec le même état de chromatine (Nichols and Corces, 2021). Un autre niveau d'organisation du génome repose sur la formation de domaines topologiquement associés (TAD pour *Topologically Associated Domains*), qui sont des boucles de chromatine. Ces boucles partiellement isolées du reste du génome permettent de favoriser les interactions entre les séquences activatrices et promotrices de la transcription (Nichols and Corces, 2021).

5 La réparation de l'ADN

L'ADN de toutes les cellules vivantes subit des dommages. Les causes de ces lésions sont variées, mais certaines sont dues au fonctionnement normal de la cellule, comme certaines erreurs de réplication qui peuvent incorporer des mésappariements (lorsque la règle de complémentarité A-T / G-C n'est plus respectée) (Ganai and Johansson, 2016). Ce sont des évènements rares mais néanmoins statistiquement significatifs. Certains processus métaboliques dans la cellule peuvent aussi créer des espèces réactives de l'oxygène ou de l'azote, qui peuvent créer sur l'ADN des sites abasiques (AP pour apurinique/apyrimidique), des cassures simple brin ou double brin, des mutations (Van Houten *et al.*, 2018). Cependant, des lésions peuvent aussi être crées par des agents extérieurs comme l'exposition à des rayonnements ionisants (rayons UV ou X), à des produits chimiques mutagènes (comme les hydrocarbures polycycliques aromatiques issus de la consommation de tabac) (Hakem, 2008). L'ADN subit également des dommages chimiques comme l'oxydation, l'alkylation, la déamination, qui peuvent mener à des mutations si ils ne sont pas réparés (Boiteux *et al.*, 2017). Enfin, l'ADN peut subir des cassures simple brin à cause de l'activité de certaines enzymes lors de la réplication ou de la transcription, ou à cause de certaines radiations, et l'accumulation de ce type de lésions peut mener à des cassures doubles brins (Hossain *et al.*, 2018, p. 2). Tous ces dommages peuvent mener à la mort de la cellule par l'apparition de mutations, qui peuvent entrainer une cancérisation des cellules (Alhmoud *et al.*, 2020).

Face à tous ces dommages, les cellules doivent être en mesure de réagir pour réparer les éventuelles lésions, sans quoi l'intégrité du génome pourrait être compromise. Ainsi, pour chaque type de dommage, les cellules possèdent des voies de réparation. Nous allons en voir quelques-unes ici, et nous nous attarderons sur l'une d'entre elles, qui est centrale dans les travaux de cette thèse.

5.1 Les bases endommagées

Comme indiqué précédemment, les nucléotides composant l'ADN peuvent eux aussi être endommagés et modifiés (par oxydation, alkylation, etc.) (Boiteux *et al.*, 2017). Dans ce cas, ces nucléotides peuvent ne plus s'hybrider normalement avec ceux de l'autre brin d'ADN. C'est le cas par exemple avec la 8-oxo-7,8-dihydroguanine (ou 8-oxoG), formée à partir de la guanine, et qui peut lors de la réplication être associée à une adénine (via une liaison de type Hoogsteen), et qui sera lors de la réplication suivante associée à une thymine (Boiteux *et al.*, 2017) (Figure 11).

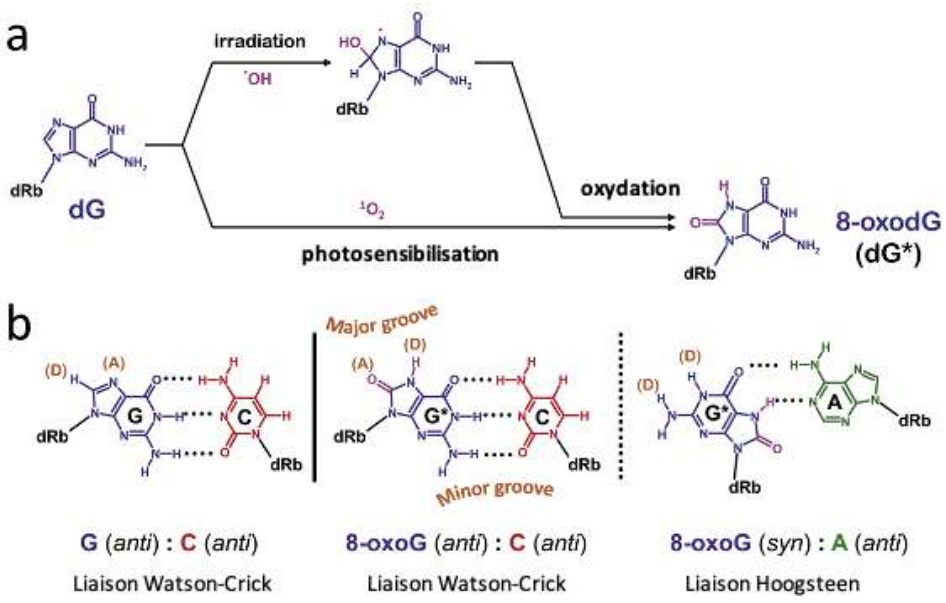


Figure 11 : Formation de 8-oxoG et de paires de bases incorrectes
a : voies de formation de la 8-oxoG
b : liaisons possibles avec (de gauche à droite) une guanine et une cytosine, une 8-oxoG et une cytosine, et une 8-oxoG et une adénine. D'après Boiteux et al. Free radical biology & medicine 2017

Ce type de dommage est donc très mutagène, puisqu'il peut muter une paire G-C en T-A (Boiteux *et al.*, 2017). Le principal système de réparation utilisé par les cellules est la réparation par excision de base (ou BER pour *Base Excision Repair*, Figure 12). Cette réparation se fait en

plusieurs étapes : la suppression de la base endommagée (laissant un site AP) par une glycosylase ; l'élimination du sucre par une AP lyase ; l'ouverture de la liaison phosphodiester par une AP endonucléase; puis soit l'ajout d'un nucléotide correct par une ADN polymérase (l'ADN polymérase β chez les eucaryotes) et le rétablissement de la liaison phosphodiester par une ligase (Boiteux *et al.*, 2017) ; soit la synthèse d'un morceau d'ADN plus long, l'élimination des bases sortantes par une flap endonucléase, et le rétablissement de la liaison phosphodiester par une ADN ligase.

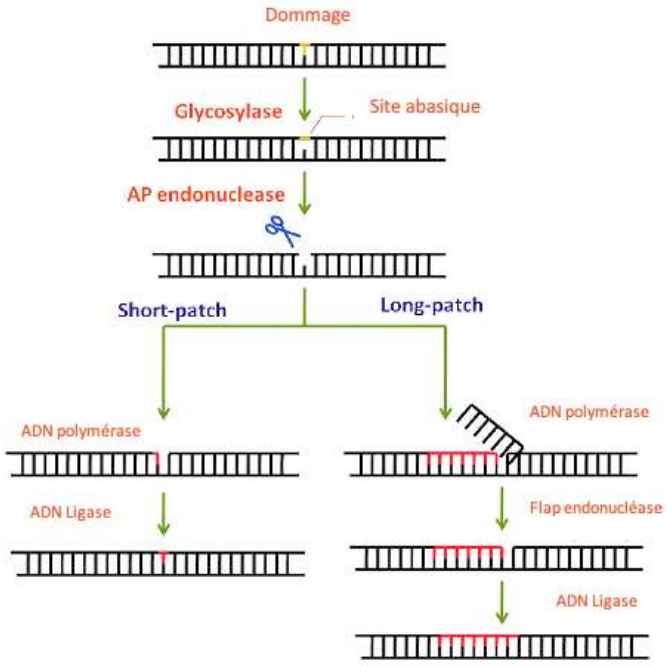


Figure 12 : Système de réparation par excision de base
Lorsqu'une base endommagée est présente dans l'ADN, elle est éliminée par une ADN glycosylase, puis le site abasique est éliminé par une endonucléase. Ensuite, soit une ADN polymérase ajoute le nucléotide manquant, et une ADN ligase rétablit la liaison phosphodiester, soit l'ADN polymérase synthétise plusieurs nucléotides, les nucléotides sortants sont éliminés, et une ADN ligase rétablit la liaison. D'après Barve A et al. DNA Repair Repertoire of the Enigmatic Hydra. Front Genet. 2021.

5.2 Les lésions déformant l'ADN

Parmi les dommages que l'ADN peut subir, certains peuvent aussi déformer sa structure (Figure 13), comme les dimères de pyrimidines (les plus connus étant les dimères de thymine (Beukers *et al.*, 2008)), les cyclopurines (Brooks, 2017), et d'autres dommages causés par des agents mutagènes ou l'exposition aux UV.

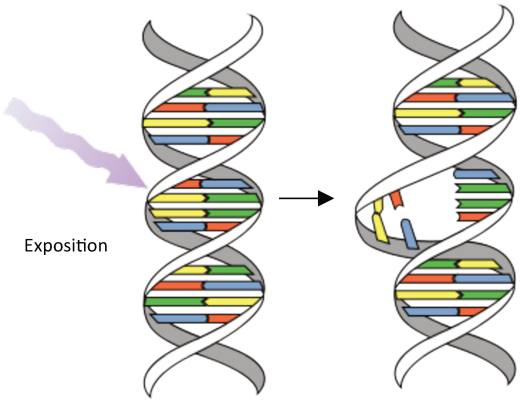


Figure 13 : Schéma de déformation de la double hélice d'ADN par un dommage du type « dimère de pyrimidines » provoqué par l'exposition aux UV. D'après Wikipedia (https://fr.wikipedia.org/wiki/Dim%C3%A8re_de_pyrimidine)

Le principal point commun de ces dommages est la déformation de la double hélice, et ce changement de conformation induit sa reconnaissance par le système de réparation par excision de nucléotides (NER pour *Nucleotide Excision Repair*, Figure 14) (Schärer, 2013). Après avoir reconnu le dommage, ce système composé de nombreux partenaires ouvre l'ADN autour de la lésion grâce à l'activité hélicase de TFIIH (un facteur de transcription), pour former une « bulle » (Schärer, 2013). Lorsque cette activité hélicase est bloquée par la lésion, d'autres partenaires sont recrutés, et coupent l'ADN en 5' et en 3' de la lésion. Enfin, des ADN polymérases resynthétisent l'ADN retiré, et des ADN ligases relient les extrémités 3'OH et 5'P du brin resynthétisé (Schärer, 2013).

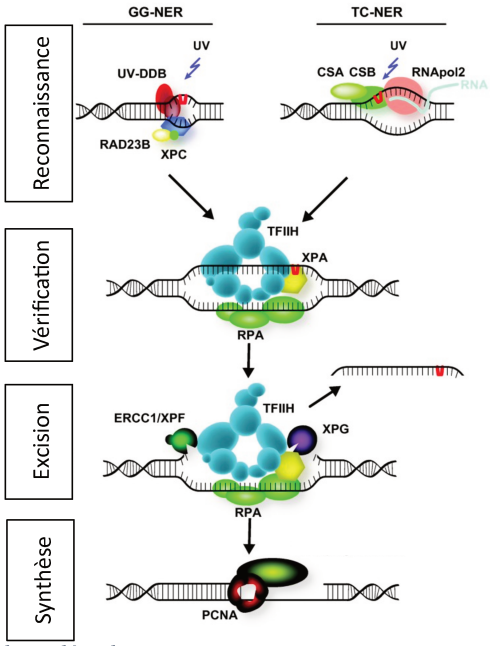


Figure 14 : Voie de réparation par excision de nucléotides
Après la reconnaissance du dommage (par des systèmes de réparation ou de transcription), la double hélice d'ADN est ouverte par TFIIH pour que le dommage soit reconnu. Le brin portant le dommage est excisé, ce qui laisse la place à des ADN polymérases et leurs partenaires pour synthétiser l'ADN de façon correcte. D'après Lans, H. et al. Epigenetics & Chromatin (2012).

5.3 Les liaisons covalentes entre les brins d'ADN (*Interstrand Crosslink*)

Certaines molécules (comme le gaz moutarde utilisé lors de la première guerre mondiale (Castaño *et al.*, 2017)) ont la capacité de lier les deux brins d'ADN, qui ne peuvent alors plus être séparés, ce qui bloque la réplication et la transcription (Deans and West, 2011). Ces dommages peuvent aussi avoir des causes endogènes comme le stress oxydatif qui peut former des molécules pouvant créer ces ICL (*Interstrand CrossLinks*). La voie de réparation des ICL porte le nom de la maladie génétique humaine dans laquelle cette réparation fait défaut : l'anémie de Fanconi (FA). La voie FA regroupe 19 protéines, et implique également d'autres systèmes de réparation (NER, synthèse translésionnelle et recombinaison homologue) (Lopez-Martinez *et al.*, 2016).

5.4 Les mésappariements

Bien que les ADN polymérases humaines réplicatives soient très fidèles (à raison d'une erreur tous les 10^8 à 10^{10} paires de bases (Bębenek and Ziuzia-Graczyk, 2018)) il peut leur arriver d'apparier incorrectement des nucléotides : on parle alors de mésappariement. Ces mésappariements peuvent mener à des mutations au fil des réplications, ainsi ces erreurs doivent être rapidement corrigées. Ces dommages sont réparés par le système de réparation des mésappariements (MMR pour *MisMatch Repair*, Figure 15) (Hsieh and Zhang, 2017). Dans un premier temps (chez les eucaryotes), l'erreur est reconnue par MutS qui recrute ensuite l'hétérodimère MutL (Hsieh and Zhang, 2017). Cette étape régit la reconnaissance des dommages de l'ADN dans cette voie de réparation (Figure 15). Le complexe MutS-MutL se déplace autour du mésappariement et introduit des coupures simple brin avec l'aide d'autres partenaires comme le PCNA (Hsieh and Zhang, 2017). Cela permet à la nucléase EXO1, grâce à son activité 5'-3' exonucléase (Keijzers *et al.*, 2015), de dégrader le brin d'ADN sur 150 pb autour du site de l'erreur, laissant un ADN simple brin qui est protégé par le RPA (Hsieh and Zhang, 2017) : c'est l'étape d'excision. Enfin, une ADN polymérase resynthétise l'ADN dégradé par l'exonucléase avec l'aide de PCNA et RFC, et l'ADN ligase I rétablit la continuité du squelette phosphodiester (Hsieh and Zhang, 2017) : c'est l'étape de resynthèse (Figure 15).

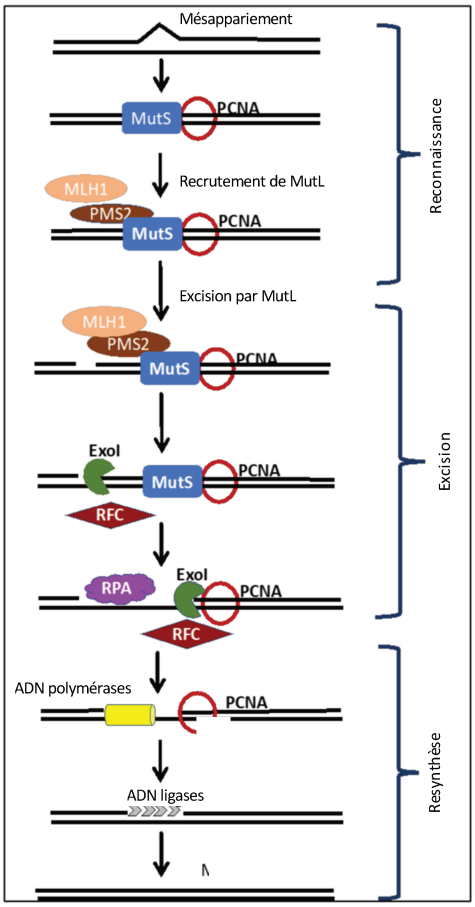


Figure 15 : Système de réparation des mésappariements
Après la reconnaissance d'un mésappariement par MutS, MutL est recruté et excise l'ADN, ce qui permet à ExoI de dégrader l'ADN jusqu'au dommage. L'ADN simple brin est protégé par RPA, et des ADN polymérases resynthétisent l'ADN manquant, avant qu'une ADN ligase ne rétablisse la liaison phosphodiester. D'après Deshpande M, et al. Cancers. 2020

5.5 Les cassures double-brins (CDB)

Parmi les dommages que subit l'ADN, celui sur lequel ces travaux se focalisent est la cassure double-brin. Comme son nom l'indique, il s'agit d'une cassure de l'ADN sur les deux brins. Les causes des CDB sont variées : causes exogènes (les radiations ionisantes et certains produits chimiques), endogènes (radicaux libres, cassures simple-brin lors de la réplication, fourches de réplication interrompues) (Cannan and Pederson, 2016), thérapeutiques ou encore dues à des mécanismes physiologiques que nous allons voir (van Gent *et al.*, 2001).

5.5.1 Les causes physiologiques de cassures double-brin

5.5.1.1 La recombinaison V(D)J

Au cœur du système immunitaire des vertébrés (Buchmann, 2014), de nombreux types de cellules sont impliquées, dont les lymphocytes B et T (Crotty, 2015). Les premiers ont pour fonction de synthétiser et libérer des immunoglobulines (Ig), qui servent à déclencher la réponse immunitaire. Les lymphocytes T ont pour fonction de reconnaitre les cellules étrangères pathogènes, grâce à des récepteurs appelés récepteurs des cellules T (ou TCR pour *T Cell Receptor*) (Crotty, 2015).

Les immunoglobulines et les TCR étant impliqués dans la reconnaissance de pathogènes variés, leur diversité est un avantage (Brandt and Roth, 2008). Le mécanisme utilisé pour générer cette diversité s'appelle la recombinaison V(D)J (Roth, 2014). Les TCR et les Ig sont constitués de façon semblable (Figure 16) : une chaine lourde qui porte une région variable (VH) et trois régions constantes (CH 1, 2 et 3) et une chaine légère constitués d'une région variable (VL) et une constante (CL) (Schroeder and Cavacini, 2010).

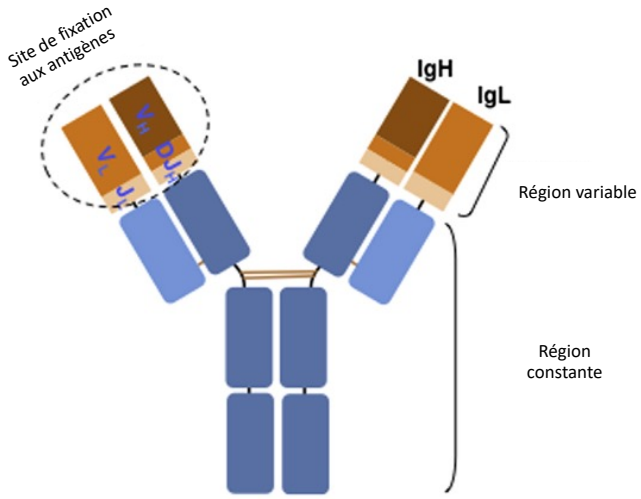


Figure 16 : Représentation schématique d'une immunoglobuline.

Les cellules B et T portent dans leur génome de nombreux allèles (versions de gènes) permettant de constituer les domaines variables VH et VL (Hoch and Schwaber, 1996). Chaque domaine variable est divisé en 3 segments : variable (V), de jonction (J) et de diversité (D). La recombinaison V(D)J a donc pour but d'associer des allèles variés de chacun de ces segments,

afin de créer des domaines variables nouveaux et diversifiés pour les Ig et les TCR, qui permettront de reconnaitre de nouveaux pathogènes (Roth, 2014).

Pour cela, la proximité des différents allèles sur chaque segment permet à la recombinaison V(D)J de les associer de façon aléatoire (Figure 17) (van Gent *et al.*, 2001). Pour chaque segment, un allèle est sélectionné au hasard et les trois sont assemblés ensemble : c'est ce nouveau gène assemblé qui une fois traduit formera les régions VH ou VL (van Gent *et al.*, 2001) des Ig et des TCR. Ce système de recombinaison repose sur l'existence aux extrémités de chacun de ces gènes de séquences qui servent de signaux de recombinaison (RSS pour *Recombination Signal Sequence*) (Brandt and Roth, 2008). Les protéines RAG1 et RAG2 introduisent une CDB à l'extrémité de deux gènes et retirent la séquence située entre les deux (Brandt and Roth, 2008). Cela forme une coupure franche aux extrémités des RSS et une structure en épingle à cheveux aux extrémités des gènes, formée par liaison covalente entre les deux brins de l'ADN. Ce processus introduit donc une CDB, qui sera réparée ensuite par le NHEJ (pour *Non Homologous End Joining* soit réparation des extrémités non homologues) (van Gent *et al.*, 2001).

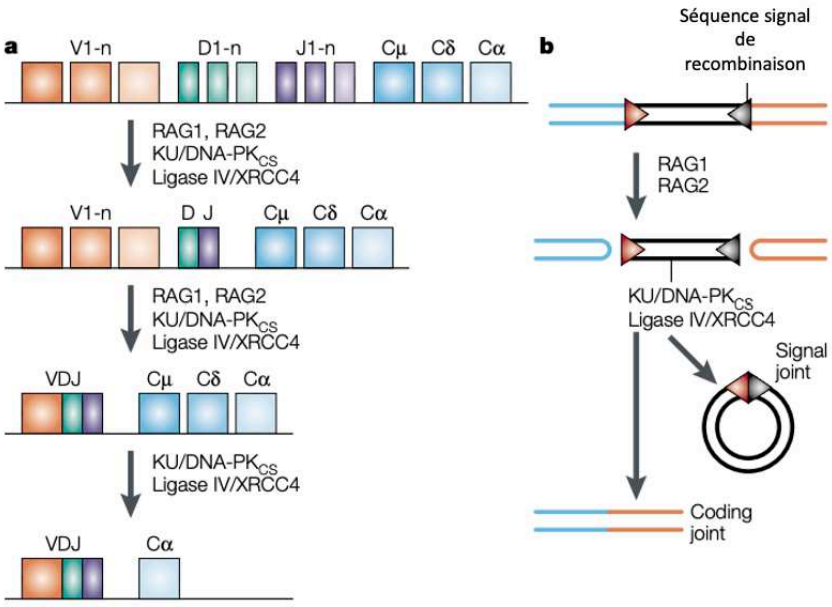


Figure 17 : La recombinaison V(D)J
a : L'objectif de la recombinaison V(D)J est de combiner des allèles variés V, D et J pour former des anticorps nouveaux. Pour cela, les protéines RAG1/2, Ku, DNA-PKcs, Ligase IV et XRCC4 (entre autres) sont nécessaires. À la fin, une combinaison unique d'allèles est obtenue, ce qui maximise la diversité des Ig et des TCR.
b : Aux extrémités des segments V, D et J, des séquences appelées RSS sont présentes. Celles-ci sont reconnues par les recombinases RAG1 et 2 qui introduisent des cassures doubles brins à chaque extrémité du segment. Ces cassures doubles brins sont ensuite réparées : les deux RSS sont assemblés en « signal joint », et les deux extrémités de segments sont assemblées en « coding joint ». Au cours de cet assemblage, une étape d'addition aléatoire de nucléotides est réalisée par la polymérase Tdt (on parle de N-addition), et permet de maximiser la diversité des gènes. D'après van Gent, D.et al. Nat Rev Genet (2001).

5.5.1.2 Les commutations isotypiques et l'hypermutation somatique

Il existe plusieurs classes d'immunoglobulines : IgG, IgM, IgA, IgD et IgE, chacune jouant un rôle différent dans l'immunité des organismes (Schroeder and Cavacini, 2010). La classe d'une Ig est définie par sa région constante (CH 1, 2 et 3 ou CL). Pour la chaine lourde, il existe 8 isotypes possibles, et 2 pour la chaine légère : et selon leur combinatoire, l'immunoglobuline formée sera une IgG, une IgM, etc (Schroeder and Cavacini, 2010). Lors de la fixation d'un antigène sur un lymphocyte B, le locus (la localisation dans le génome) duquel dépend l'expression des Ig subit deux types d'altérations : la commutation isotypique (pour les chaines lourdes constantes) et l'hypermutation somatique pour les chaines légères et lourdes variables (Chi *et al.*, 2020). Dans ces deux cas, les CDB sont programmées.

La commutation isotypique a pour but d'échanger un gène codant pour une chaine lourde constante avec un autre gène dans un ensemble présent dans le génome (Xu *et al.*, 2012). Pour cela (Figure 18), une désaminase appelée AID reconnait spécifiquement certaines cytidines sur des régions répétées précédant les gènes des différents isotypes, et les modifie pour former des déoxyuraciles dU (Xu *et al.*, 2012). Le recrutement du MMR ou du BER pour corriger la présence de ce U dans l'ADN entraine la formation d'une CDB qui sera réparée par le NHEJ, car les modifications peuvent être proches les unes des autres et leur réparation entraine momentanément des cassures double brin (Xu *et al.*, 2012).

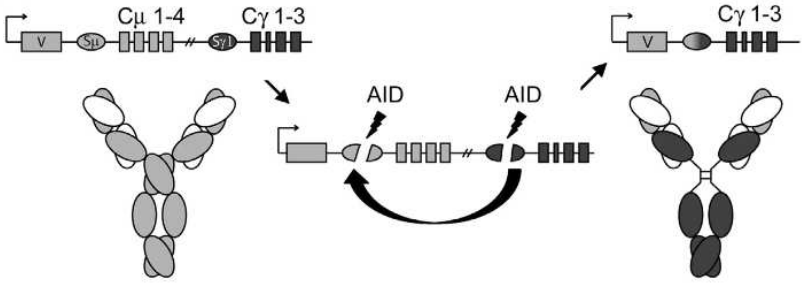


Figure 18 : La commutation isotypique
Une séquence codant pour une chaine lourde constante (Cγ ici) en remplace une autre (Cμ), en passant par l'insertion de CDB aux extrémités des gènes. D'après Martin, A. et al. Chapter 20 - Somatic Hypermutation: The Molecular Mechanisms Underlying the Production of Effective High-Affinity Antibodies. Molecular Biology of B Cells (Second Edition) 2015, Pages 363-388.

Dans le cas de l'hypermutation somatique (Figure 19), AID convertit en uraciles des cytidines situées dans les gènes V, D et J, ce qui forme des mésappariements (Martin *et al.*, 2015). Le système de réparation MMR ou BER est donc recruté pour corriger l'erreur, pouvant insérer une mutation qui enrichit la diversité des sites de fixation aux antigènes (Martin *et al.*, 2015).

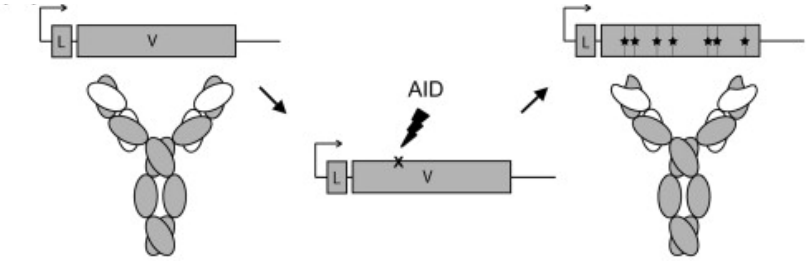


Figure 19 :L'hypermutation somatique
La région V codant pour un anticorps accumule très rapidement des mutations, ce qui permet de créer des anticorps variants. D'après Martin, A. et al. Chapter 20 - Somatic Hypermutation: The Molecular Mechanisms Underlying the Production of Effective High-Affinity Antibodies. Molecular Biology of B Cells (Second Edition) 2015, Pages 363-388.

5.5.1.3 Les recombinaisons méiotiques

Durant la méiose, qui est une division de cellules qui permet la formation de gamètes haploïdes (ne portant qu'un chromosome de chaque paire), un brassage génétique a lieu, pour permettre de diversifier les gènes présents dans les gamètes. Le processus qui permet cette diversification est la recombinaison (Baudat *et al.*, 2013).

Lors de la méiose, des chromosomes homologues se rencontrent le long de l'axe chromosomique. Cette rencontre permet la formation de cassures double-brin par la protéine Spo11 (et d'autres protéines et complexes comme MRX, Sae et ExoI) sur un des chromosomes, et cet ADN endommagé est pris en charge par le système de recombinaison homologue. La résolution de ces cassures et mélanges de brins entre les deux chromosomes peut soit insérer une séquence d'un chromosome dans l'autre, soit échanger une partie des chromosomes : on parle de *crossing-over* (Baudat *et al.*, 2013).

5.5.1.4 Les cassures double brins à but thérapeutique

Il peut sembler surprenant de vouloir causer des dommages tels que des CDB à des cellules vivantes, étant donnée leur nature destructrice. Cependant, dans certains cas, il est utile de pousser les cellules vers leur mort (l'apoptose), comme dans le cas des cancers, où leur fonctionnement est déréglé. C'est pourquoi une partie des traitements anticancéreux actuels repose sur l'idée de créer des dommages de l'ADN (comme la radiothérapie ou les inhibiteurs de topoisomérase (Pommier, 2006)), y compris des CDB (Moon *et al.*, 2023). D'autres traitements visent plutôt à empêcher la réparation des CDB, comme les inhibiteurs de ligase (Greco *et al.*, 2016).

Les cellules cancéreuses sont particulièrement sensibles à certains types de stress (métabolique, protéotoxique, oxydatif, mitotique…) dont le stress génotoxique : celui lié aux dommages à l'ADN. C'est sur cela que reposent les traitements anticancéreux : la création de nombreuses CDB, de telle façon que les systèmes de réparation de la cellule soient saturés, ce qui entraine ces cellules vers l'apoptose (Moon *et al.*, 2023).

Le traitement le plus classique face au cancer à l'heure actuelle est la radiothérapie, qui consiste à irradier la tumeur. D'autres traitements sont les molécules radiomimétiques (qui miment l'action des rayonnements). Celles-ci ont pour fonction d'attaquer le désoxyribose de l'ADN sur des séquences spécifiques pour l'oxyder et le rendre très réactif chimiquement, ce qui rompt la liaison glycosidique entre la base et le sucre ou la liaison phosphodiester. Lorsque cela se produit sur les deux brins de l'ADN, cela peut mener à des CDB (Povirk, 1996).

D'autres molécules utilisées en traitement anticancéreux sont les inhibiteurs de ligases ou de topoisomérases. L'idée de ce type de molécule est de bloquer le fonctionnement d'une enzyme indispensable à la cellule pour maintenir son génome (Moon *et al.*, 2023).

5.6 La réparation des cassures double brin

Au sein des cellules eucaryotes, il existe deux systèmes de réparation des cassures double-brin : la recombinaison homologue et le NHEJ. Les deux peuvent être utilisés pour la réparation de toute une variété de CDB (exemples en partie gauche de la figure 25). Bien que le NHEJ entre souvent en action en premier (Karanam *et al.*, 2012), le choix entre les deux est lié à différents facteurs, tels que la structure des extrémités lésées de l'ADN (Reynolds *et al.*, 2012) ou l'étape du cycle cellulaire (Scully *et al.*, 2019).

5.6.1 La recombinaison homologue

Le premier système de réparation et aussi le plus efficace des deux (car le moins mutagène) est la recombinaison homologue (Wright *et al.*, 2018). Celle-ci repose sur le concept du « copier-coller ». En effet, en résumé ce système utilise une copie saine de l'ADN endommagé pour permettre à ce dernier de retrouver sa séquence exacte (Wright *et al.*, 2018). Cependant, cela nécessite que deux copies du génome soient présentes dans la cellule, ce qui ne permet à la recombinaison homologue d'être active que dans les phases où deux chromatides

sœurs sont présentes : les phases S, G2 et M du cycle cellulaire (Hustedt and Durocher, 2017) (Figure 20).

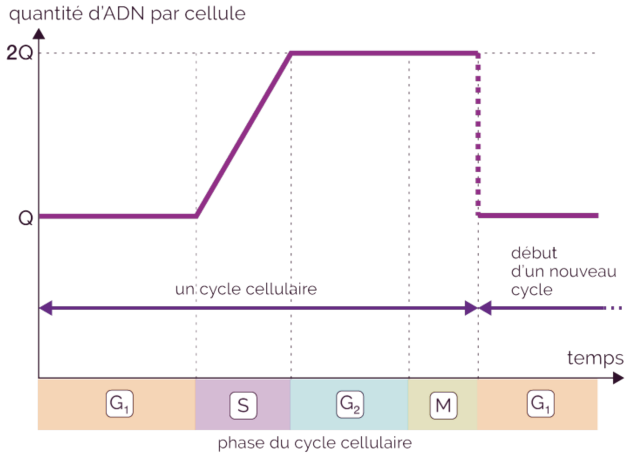


Figure 20 : Évolution de la quantité d'ADN au cours des différentes phases du cycle cellulaire. D'après https://www.maxicours.com/se/cours/la-mitose-reproduction-conforme-de-l-information-genetique--terminale--svt/

Ainsi, lorsqu'une CDB apparait, la première étape de la recombinaison homologue est la dégradation des extrémités lésées de 5' vers 3' sur un brin, de façon à obtenir des extrémités 3'OH simple brin (Wright *et al.*, 2018) (Figure 21). C'est dès cette étape que le choix du système de réparation se fait : en effet, une des protéines impliquées dans cette résection, CtIP doit, pour être activée, subir une phosphorylation par une protéine kinase (Anand *et al.*, 2016). Une fois phosphorylée, CtIP agit comme un cofacteur du complexe MRN (Mre11-Rad50-NBS1), qui a des activités endo- et exonucléase (Anand *et al.*, 2016). Mre11 utilise son activité 5'-3' endonucléase pour couper un brin d'ADN près du site de cassure, puis son activité 3'-5' exonucléase pour dégrader ce même brin d'ADN (Wright *et al.*, 2018). Ce traitement de l'ADN bloque les facteurs de reconnaissance impliquées dans le NHEJ, afin d'assurer une réparation par la recombinaison homologue. Ensuite, la protéine Exo1 dégrade l'ADN (Wright *et al.*, 2018). Une fois les brins d'ADN simple brin formés, ils sont pris en charge par des filaments formés par Rad51 et ses paralogues (Taylor *et al.*, 2015), après déstabilisation des protéines RPA fixées sur l'ADN simple brin pour le protéger, par Rad52 ou avec l'aide de BRCA2 (Jensen *et al.*, 2010; New *et al.*, 1998) (Figure 21). Le filament nucléoprotéique formé scanne ensuite l'ADN sain jusqu'à trouver une homologie suffisante pour tout l'ADN simple brin. Pour cela, des protéines partenaires ont pour rôle de « retourner » les bases de l'ADN scanné afin de les faire sortir de leur double hélice et essayer de les associer à l'ADN simple brin (Dray *et al.*, 2010; Modesti *et al.*, 2007). La formation d'homologies d'au moins 8 pb stabilise le complexe entre Rad51, l'ADN

simple brin (sb) et l'ADN double brin (db) (Qi *et al.*, 2015). Une fois qu'une homologie a été trouvée, un appariement est formé entre l'ADN sb et son brin complémentaire dans l'ADN db : on parle de synapse ou complexe synaptique (Figure 21). Celle-ci forme une boucle D : l'extrémité 3' s'entrelace dans son ADN db complémentaire, ce qui forme l'équivalent d'une jonction amorce-modèle, comme dans la réplication (Wright *et al.*, 2018). Cela déplace le brin complémentaire de l'ADN db pour former un hétéroduplexe entre un brin de l'ADN db et l'ADN sb. La formation de ces boucles D est réalisée par Rad54, qui retire Rad51 pour permettre l'hybridation de l'ADN sb avec le db (Wright and Heyer, 2014). Une fois l'hybridation réalisée et l'extension du brin à réparer finie (grâce à l'ADN polymérase δ et PCNA (Li *et al.*, 2009)), il existe deux façons de terminer la réparation : l'hybridation de brins synthèse dépendante (SDSA pour *synthesis dependant strand annealing*) ou la formation de doubles jonctions de Holliday (dHJ) (Wright *et al.*, 2018)(Figure 21).

Dans le cas du SDSA (Figure 21), une fois que le brin à réparer sort de la boucle D après plusieurs cycles de formations de boucles D et élongations (McVey *et al.*, 2004), il s'hybride à son brin complémentaire (en aval de la cassure). À partir de là, la réparation peut être terminée par une ADN polymérase (McVey *et al.*, 2016) et une ADN ligase (Wright *et al.*, 2018). Cette voie est celle qui est favorisée *in vivo* (Zapotoczny and Sekelsky, 2017), et elle permet d'éviter la formation de *crossing over* entre les chromosomes, et donc l'échange de matériel génétique entre les deux ou le passage de matériel génétique de l'une à l'autre.

Dans le cas des doubles jonctions de Holliday (Figure 21), le second brin de l'ADN à réparer s'hybride avec le brin déplacé de l'ADN db homologue, ce qui permet l'extension des deux brins dans la boucle D (Wright *et al.*, 2018). Cette double jonction de Holliday est ensuite résolue soit par dissolution, ce qui permet d'éviter les crossing-over, soit par résolution nucléolytique, ce qui peut créer des *crossing-overs* (comme dans le cadre de la recombinaison méiotique) (Wright *et al.*, 2018).

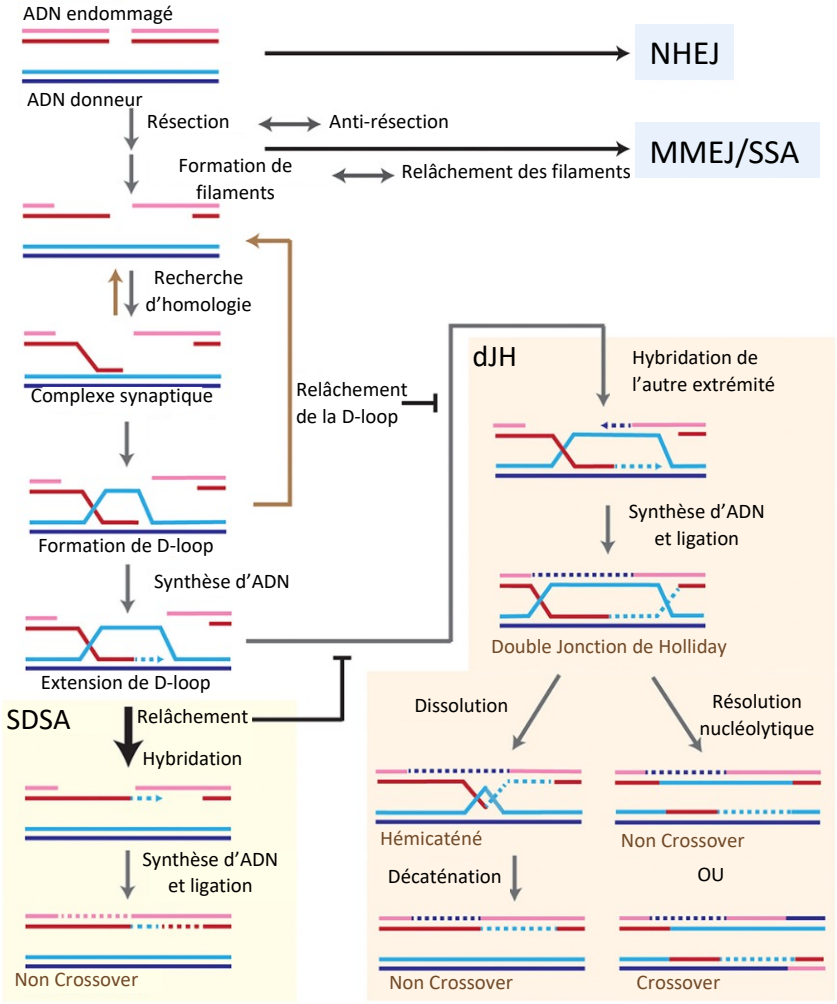


Figure 21 : Modèle de réparation par le système de recombinaison homologue.
À la suite de l'apparition de CDB dans l'ADN, si les systèmes NHEJ et MMEJ ne peuvent pas être utilisés, c'est la recombinaison homologue qui répare ce dommage. Pour cela, un ADN homologue est utilisé : une homologie de séquence est recherchée, et lorsqu'elle est trouvée l'ADN endommagé envahit la double hélice intacte pour former une « boucle D ». La majorité de ces boucles sont interrompues et réparées par hybridation des brins dépendante de la synthèse (ou SDSA) : le brin endommagé ayant commencé à être synthétisé est réhybridé à son brin initial, et la réparation se termine par la synthèse de l'ADN manquant et la ligation. Si la boucle est maintenue, la réparation se fait via la formation d'une double jonction de Holliday. Le brin ayant envahi son homologue continue d'être synthétisé, et le second brin endommagé est réparé grâce au brin homologue déplacé lors de la formation de la boucle. Cette double jonction de Holliday peut ensuite soit être rompue par une topoisomérase; soit par coupure de l'ADN entremêlé, ce qui peut ou non former un crossing-over . D'après Wright WD, et al. J Biol Chem.

5.6.2 Le NHEJ

Le second système de réparation des CDB, central dans le cadre de cette thèse, est le NHEJ. Là où la recombinaison homologue fait un « copier-coller », le NHEJ vise quant à lui à recoller les extrémités lésées. Son fonctionnement est donc plus mutagène, car il ne se base pas sur la reproduction d'une information existante mais en rajoute si il en manque, sans modèle (ce système n'a pas forcément besoin de micro-homologies entre les extrémités lésées : 40% des réparations par le NHEJ se font sans micro-homologie (Pannunzio *et al.*, 2014)).

En résumé, le NHEJ est composé de quatre grandes étapes suivant la cassure de l'ADN : la reconnaissance de la CDB, la mise en place de la synapse NHEJ, le traitement des extrémités lésées et la ligation (Figure 22).

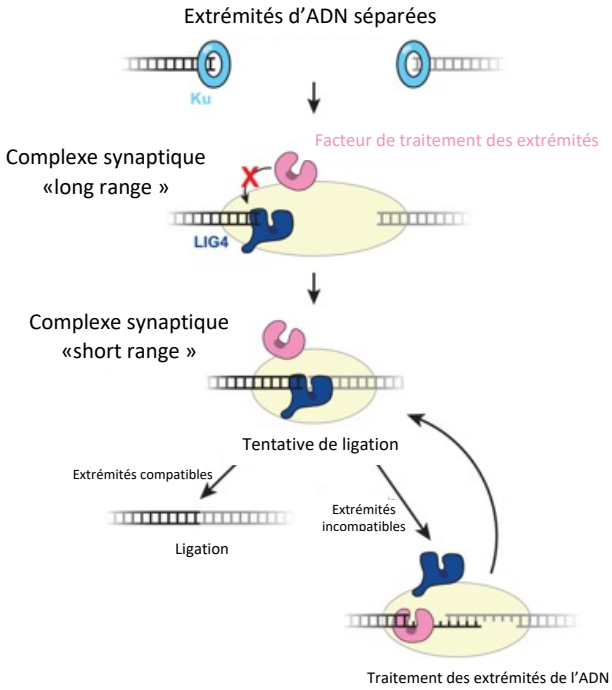


Figure 22 : Résumé schématique des principales étapes du NHEJ. Après la reconnaissance des extrémités d'ADN par Ku70/80, DNA-PKcs est recruté, pour former le complexe « Long Range », complété entre autres par l'ADN ligase IV. À cette étape, les extrémités d'ADN ne peuvent pas être traitées par d'autres facteurs. Le complexe synaptique « Short Range » permet ensuite le rapprochement des extrémités d'ADN, et le recrutement de facteurs de traitement. Si la ligation est possible, elle est réalisée par la ligase IV. Au besoin, des facteurs de traitement peuvent interagir avec l'ADN, pour le rendre apte à la ligation. D'après Carney et al. eLife 2020.

Le NHEJ est principalement utilisé hors des phases G2 et S du cycle cellulaire, bien que même en G2 il corrige jusqu'à 80% des cassures générées par des radiations ionisantes, si ces cassures sont éloignées des fourches de réplication (Karanam *et al.*, 2012). C'est également le système de réparation utilisé dans le cadre de la recombinaison V(D)J et de la commutation isotypique.

5.6.2.1 La reconnaissance des CDB et la mise en place de la synapse (complexe Long Range)

La première étape du NHEJ, comme évoqué précédemment, est la reconnaissance de la CDB. Les protéines impliquées dans cette reconnaissance sont celles du dimère Ku70/80. Ce dimère identifié en 1981 (Mimori *et al.*, 1981) forme un anneau dans lequel l'ADN s'insère, il a

une grande affinité pour les extrémités de l'ADN (avec un K_d de l'ordre du nM (Blier *et al.*, 1993)) peu importe leur état ou leur séquence, et il est relativement abondant dans le noyau des cellules (environ 500 000 molécules par cellule humaine) (Walker *et al.*, 2001). La structure de cet hétérodimère a été résolue chez *Homo sapiens et Saccharomyces cerevisiae* (Walker *et al.*, 2001), et semble avoir été conservée au cours de l'évolution des eucaryotes. Ku70/80 n'interagit qu'avec le squelette sucre-phosphate de l'ADN, mais avec une orientation spécifique : Ku70 est plus proche de l'extrémité libre de l'ADN (Yoo and Dynan, 1999). Lors de l'obtention de la structure du dimère, les auteurs ont suggéré que certains résidus en contact avec l'ADN, de par leurs chaines latérales, pourraient limiter le glissement de Ku le long de l'ADN, sans quoi l'ADN serait recouvert de dimères Ku70/80 ayant glissé après s'être fixé à une extrémité libre (Grob *et al.*, 2012). De plus, Ku70/80 présente une activité 5'dRP/AP lyase, qui lui permet d'éliminer les sites AP dans certains cas (Roberts *et al.*, 2010).

Le rôle premier de Ku 70/80 est de détecter les cassures (en quelques secondes, le dimère commence à s'accumuler auprès de l'ADN endommagé (Mari *et al.*, 2006)) et de protéger les extrémités d'ADN (Mimitou and Symington, 2010). Pour cela, l'hétérodimère a besoin d'un minimum de 14 pb pour s'installer (Frit *et al.*, 2019). Son rôle est ensuite de servir de plateforme de recrutement pour les autres partenaires du NHEJ.

À la suite de la reconnaissance d'une CDB par Ku70/80, l'apoenzyme DNA-PKcs (sous unité catalytique de la protéine kinase dépendante de l'ADN) est recrutée, l'ensemble forme DNA-PK (Hennequin *et al.*, 1999), et Ku70/80 est « poussé » sur l'ADN, s'éloignant des extrémités de la CDB et laissant sa place à DNA-PKcs à l'extrémité de l'ADN (Yoo *et al.*, 1999) (Figure 23).

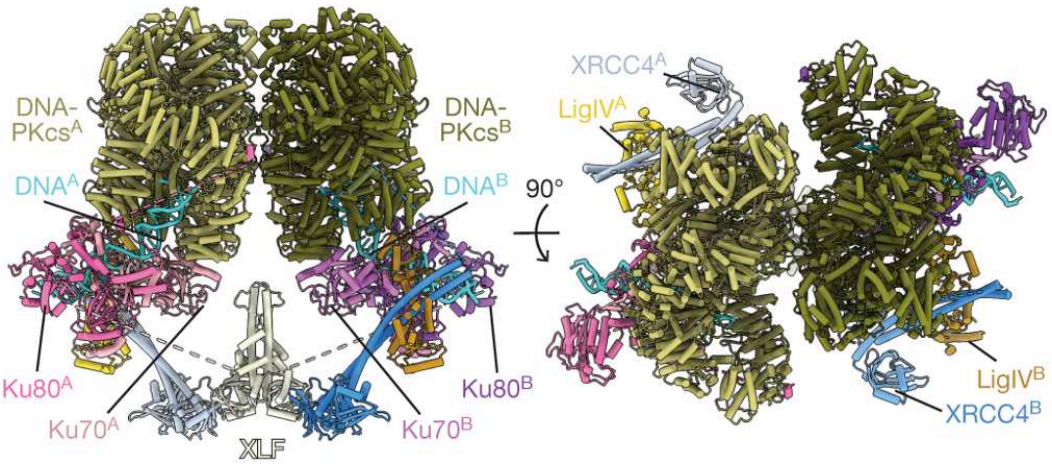


Figure 23 : Structure du complexe NHEJ "Long Range". D'après Chen Set al. Nature. 2021

Le rôle de ce complexe *Long Range* (LR) est de stabiliser les deux extrémités lésées de l'ADN côte à côte en formant la synapse du NHEJ, en les tenant à distance l'une de l'autre (>80 angströms (Å)) (Graham *et al.*, 2016). Ce complexe est stabilisé par d'autres protéines : PAXX, XRCC4, et XLF (Figure 23).

XLF et XRCC4 sont deux protéines assez similaires d'un point de vue structural, bien que leurs séquences soient différentes (Ahnesorg *et al.*, 2006). Elles sont toutes deux formées d'une tête N-terminale globulaire, d'un *coiled coil* central permettant leur homo-dimérisation, et d'un domaine C-terminal non structuré (Li *et al.*, 2008). Dans les complexes du NHEJ, XLF et XRCC4 interagissent via leurs têtes N-terminales, mais XLF interagit en plus avec Ku70/80 via un motif d'interaction (KBM pour *Ku Binding Motif*) situé dans son domaine C-terminal (Yano *et al.*, 2011). L'ensemble de ces interactions entre un dimère XLF, un complexe XRCC4-Ligase IV (X4L4) et Ku-ADN, semblent permettre à elles seules la formation d'un autre complexe appelé *Short-Range* (voir plus bas).

Un autre partenaire, PAXX, est très proche de XLF d'un point de structural, et sa fonction n'est pas encore totalement comprise (Seif-El-Dahan *et al.*, 2023). Sa présence est accessoire mais ce partenaire semble pouvoir stabiliser le dimère de DNA-PK, et partiellement se substituer à XRCC4 ou le compléter en interagissant avec Ku *via* un autre site d'interaction (Ochi *et al.*, 2015).

5.6.2.2 *La formation du complexe Short Range et le traitement des extrémités de l'ADN*

Après la formation du complexe LR, les deux DNA-PK se phosphorylent mutuellement et se détachent de la synapse, ce qui permet aux autres facteurs de changer de conformation pour rapprocher les extrémités de l'ADN, formant le complexe *Short Range* (Chen *et al.*, 2023), ce qui permet à d'autres protéines d'être recrutées pour maturer les extrémités de l'ADN *(Vogt et al.*, 2023) (Figure 24).

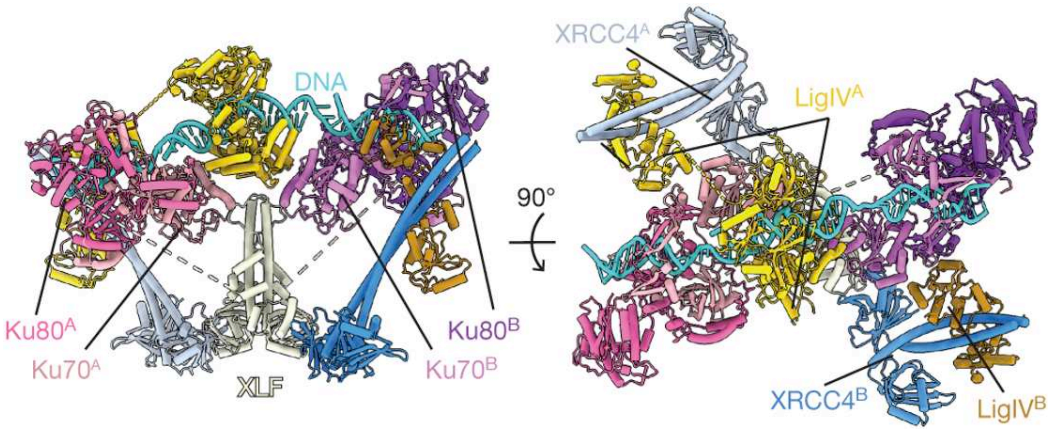


Figure 24 : Structure du complexe de ligation "Short Range". D'après Chen S et al. Nature. 2021.

Si les extrémités de l'ADN sont compatibles, la ligation peut être réalisée directement par l'ADN ligase IV. Mais dans le cas où les extrémités de l'ADN ne peuvent pas être rattachées par une ligation directe, une étape de modification de ces extrémités est nécessaire. Il a été montré que l'ADN ligase IV favorise certaines voies de traitement des extrémités de l'ADN, selon les besoins (Waters *et al.*, 2014). L'état de ces extrémités pouvant être très varié, de nombreuses fonctions peuvent être requises (Figure 25).

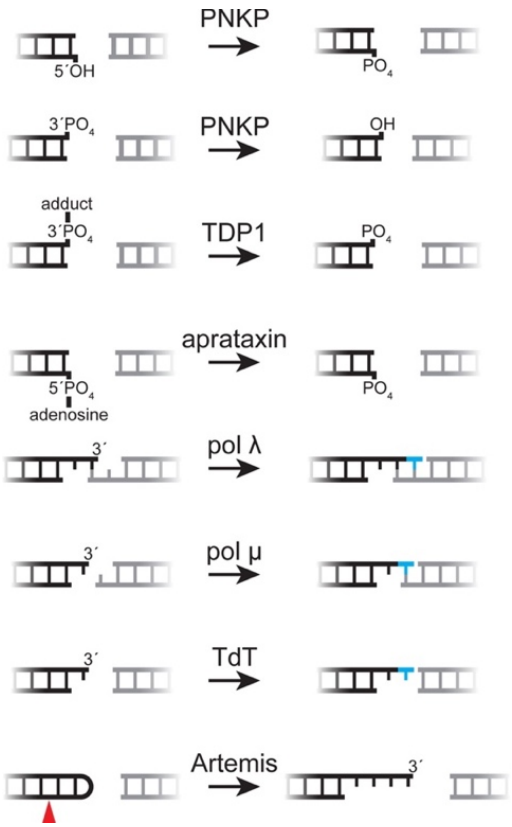


Figure 25 : Les enzymes de traitement des extrémités de l'ADN et leurs contextes d'activation. D'après Stinson BM et al. Annu Rev Biochem. 2021.

Dans le cas où l'extrémité 5' de l'ADN endommagé aurait perdu un groupement phosphate, ou bien si l'extrémité 3' porte un groupement phosphate, la ligation ne peut pas avoir lieu. Dans ces cas, la PNKP (PolyNucléotide Kinase Phosphatase) peut soit ajouter soit retirer des groupements phosphates, de manière à permettre la ligation (Weinfeld *et al.*, 2011). Cette protéine interagit avec XRCC4 *via* son domaine FHA N-terminal.

Dans certains cas, pouvant être programmés (comme dans la recombinaison V(D)J)) ou non, les extrémités de l'ADN peuvent former une structure en épingle à cheveux (lorsque les groupements 3'OH et 5'P sont liés par les transposases RAG par exemple). Dans ce cas, ces extrémités ne peuvent pas être utilisées pour une ligation, donc cette liaison doit être rompue par une nucléase. Ce rôle est rempli par la nucléase appelée Artemis (Ma *et al.*, 2002).

Lorsqu'une réaction de ligation commence et ne peut être finie, un groupement adénosine peut être transféré sur le 5'P. Ce groupement peut alors être retiré par l'aprataxine (Ahel *et al.*, 2006).

Si l'activité d'une topoisomérase est requise et qu'elle est bloquée (par un inhibiteur de topoisomérases par exemple), elle peut rester liée covalement aux extrémités 5' ou 3' de l'ADN, ce qui empêche tout traitement de l'ADN (Cho and Jinks-Robertson, 2018). Les protéines TDP1 et TDP2 peuvent alors être recrutées pour supprimer cette liaison covalente (Cortes Ledesma *et al.*, 2009). TDP1 peut également supprimer d'autres types de groupements ou adduits sur les extrémités de l'ADN (Interthal *et al.*, 2005).

D'autres protéines ayant des activités de nucléases (APLF), hélicases (WRN (Rossi *et al.*, 2010)), ou même dRP/AP lyases (Ku) peuvent également être nécessaires (Menon and Povirk, 2016).

Enfin, des ADN polymérases peuvent également être impliquées. Il s'agit en particulier des ADN polymérases de la famille X ayant un domaine BRCT N-terminal : les ADN polymérase λ, μ et Tdt (Terminal déoxynucléotidyltransférase) (Nick McElhinny *et al.*, 2005; Pryor *et al.*, 2015; Uchiyama *et al.*, 2009) (Figure 26), qui seront au cœur des travaux de cette thèse.

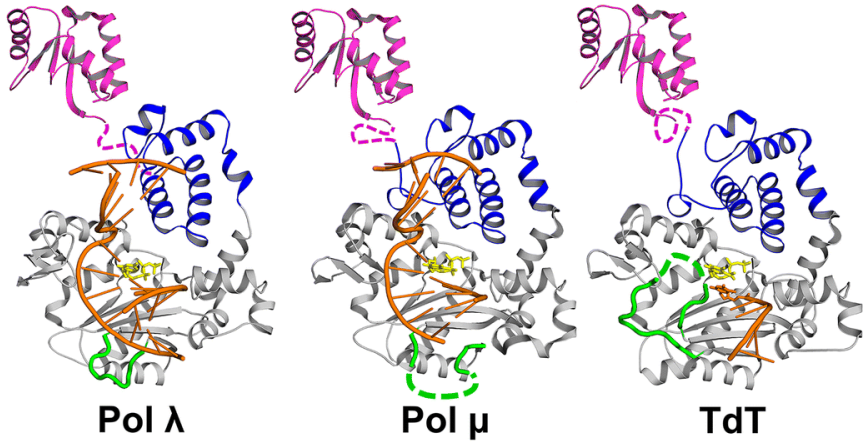


Figure 26 : Structures des ADN polymérases de la famille X impliquées dans le NHEJ. Les trois ont un domaine BRCT (en rose) permettant l'interaction avec Ku70/80, un domaine de 8-kDa (en bleu) servant à la reconnaissance des groupements 5'P en aval des cassures de l'ADN, et à l'activité dRP lyase chez l'ADN polymérase λ. Au sein du domaine catalytique (en gris), les principales différences se situent dans la boucle 1 (en vert) et dans le site catalytique, où l'ADN polymérase λ a un steric gate bloquant l'incorporation des NTPs. La Tdt peut grâce à sa boucle 1 incorporer des nucléotides de façon indépendante d'un brin template. D'après Hoitsma NM et al. Cell Mol Life Sci. 2020

Ce domaine N-terminal leur permet d'interagir avec Ku70/80 au sein du NHEJ. Il a d'ailleurs été montré que ces ADN polymérases ont la priorité sur les nucléases en cas d'ADN simple brin non compatible, ce qui permet de perdre moins d'information génétique et donc réduit la mutagénicité (Stinson *et al.*, 2020). L'ADN polymérase μ et la Tdt ont la capacité d'introduire des ribonucléotides dans l'ADN (Martin *et al.*, 2013; Pryor *et al.*, 2018). Cette

possibilité est avantageuse car les NTPs sont bien plus abondants que les dNTPs dans la cellule (environ 100 µM de NTPs pour 10 µM de dNTPs) (Traut, 1994), et la ligation de ces brins ADN/ARN est plus tolérante à certains dommages de l'ADN (Nick McElhinny and Ramsden, 2003).

5.6.2.3 La ligation

Dans les cas où aucun traitement de l'ADN n'est nécessaire, ou bien quand ce traitement est fini, la ligation peut avoir lieu. Cette étape est réalisée par l'ADN ligase IV (U. Grawunder *et al.*, 1998), qui est formée d'un domaine catalytique commun aux ADN ligases (sous divisé en trois domaines : un domaine de fixation à l'ADN (DBD), un domaine nucléotidyltransférase et un domaine OBD (Pascal, 2008)) et de deux domaines BRCT (1 et 2) en C-terminal (Sibanda *et al.*, 2001). La ligase IV interagit au sein du NHEJ via son domaine BRCT 2 avec XRCC4 (Ulf Grawunder *et al.*, 1998). Elle interagit avec Ku70/80 via son premier domaine BRCT (Costantini *et al.*, 2007), et est présente en deux copies lors de la réparation d'une CDB (Chen *et al.*, 2021).

Une particularité de l'ADN ligase IV est qu'elle peut utiliser l'ATP ou le NAD+ pour son adénylation initiale (Chen and Yu, 2019). De plus, contrairement aux autres ADN ligases humaines (LigI, Lig III), la ré-adénylation post catalytique est peu efficace chez la ligase IV (Chen *et al.*, 2009), et est aidée d'autres partenaires comme XLF (Riballo *et al.*, 2009).

Bien qu'elle soit surtout active après le traitement des extrémités de l'ADN par d'autres enzymes, elle peut tout de même s'accommoder de certains dommages de l'ADN non réparés (Gu *et al.*, 2007a) comme des bases oxydées ou des gaps (Gu *et al.*, 2007b), et peut même assurer la ligation en présence de NTPs (Nick McElhinny and Ramsden, 2003).

5.6.2.4 La fin du NHEJ

Enfin, après que la ligation a eu lieu, le complexe SR est dissocié de l'ADN. Le mécanisme de cette dissociation n'est pas encore tout à fait compris, mais il semble que cela repose sur l'ubiquitinylation de Ku70/80, qui entraine sa dégradation par le protéasome (Ishida *et al.*, 2017), ou sur la phosphorylation de Ku70 (Lee *et al.*, 2016).

5.6.2.5 L'influence de la chromatine sur le NHEJ

Il a été montré que tout le génome n'est pas soumis au même risque de rencontrer des CDB. En effet, les gènes transcriptionnellement actifs sont particulièrement à risque d'être touchés par ces dommages (Canela *et al.*, 2017; Marnef *et al.*, 2017). Il a également été montré que la chromatine est momentanément remodelée par la protéine PARP1 suite à l'introduction de CDB dans le génome (Smith *et al.*, 2018), et ce de façon spécifique selon le système de réparation utilisé. La composition en histones des nucléosomes est différente selon le système de réparation utilisé, tout comme les modifications d'histones présentes (Iacovoni *et al.*, 2010). Ces modifications pourraient permettre à la chromatine des zones endommagées de se regrouper sous forme de *clusters* dans le noyau (Arnould *et al.*, 2023). Cependant, ces modifications sont bien plus marquées pour les CDB réparées par la recombinaison homologue que par le NHEJ (Clouaire *et al.*, 2018). En plus de ces modifications permettant de regrouper les zones endommagées du génome, il a été montré que des modifications d'histones spécifiques du NHEJ (Clouaire and Legube, 2019), comme l'ajout aux nucléosomes du variant d'histone H3.3 ou H2A.Z, permettent de favoriser le recrutement des protéines du NHEJ, de favoriser ce système de réparation, de déplacer les histones et nucléosomes pour faciliter le NHEJ, voire de bloquer la résection de l'ADN pour empêcher la recombinaison homologue (Clouaire and Legube, 2019).

5.6.2.6 Le NHEJ chez les levures

Le NHEJ est un système conservé chez les eucaryotes en général, mais certains organismes présentent quand même des spécificités. Chez les levures, en particulier l'organisme modèle *Saccharomyces cerevisiae*, le NHEJ et les protéines impliquées ont été caractérisés. En grande majorité, les protéines humaines trouvent des homologues, sauf concernant DNA-PKcs qui est absent chez la levure (Emerson and Bertuch, 2016). L'équivalent de Ku70/80 s'appelle chez la levure YKu70/80, et n'est pas essentiel à la survie des cellules (Ribes-Zamora *et al.*, 2007). Comme pour Ku70/80, le rôle de cet hétérodimère est de recruter d'autres partenaires du NHEJ. L'homologue de l'ADN ligase IV humaine s'appelle chez la levure Dnl4, et est indissociable de Lif1, qui est un homologue de XRCC4, et sans lequel Dnl4 n'est pas stable. Lif1 est construite comme XRCC4, avec une tête globulaire et un *coiled-coil* central lui permettant d'interagir avec le second BRCT de Dnl4 (Deshpande and Wilson, 2007). Le premier BRCT permet quant à lui le recrutement de Dnl4 par YKu70/80 (Chiruvella *et al.*, 2014). Cette

interaction entre Dnl4/Lif1 et YKu70/80 est indispensable au NHEJ, et chaque partenaire permet de stabiliser la liaison à la CDB des autres, y compris une protéine additionnelle : Nej1 (Chen and Tomkinson, 2011). Nej1 est une protéine indispensable au NHEJ chez la levure, qui interagit avec Lif1 (Deshpande and Wilson, 2007) et est probablement un homologue de XLF (Mojumdar *et al.*, 2022). Les deux protéines interagissent *via* leurs têtes globulaires, et cette interaction n'a pas d'impact sur le complexe de Lif1 avec Dnl4. Nej1 peut aussi interagir directement avec certains types de dommages et avec YKu70/80, et permet même de faciliter la ligation par Dnl4 (Chen and Tomkinson, 2011).

D'autres protéines sont également présentes chez la levure, en particulier pour traiter les extrémités de l'ADN (Emerson and Bertuch, 2016). On peut citer l'ADN polymérase 4 (ou IV), une ADN polymérase X équivalente aux ADN polymérases λ et μ humaines, qui a une activité de gap-filling, et de synthèse en situation de CDB, malgré une faible processivité et une faible fidélité (Bebenek *et al.*, 2005). D'autres facteurs peuvent porter des activités nucléases (Rad27, Pol2), polymérases (Pol3, qui a une activité proche de celle de Pol4), ou phosphodiestérase (Tdp1, qui bloque l'activité de l'ADN polymérase IV en formant des extrémités 3' phosphate) (Emerson and Bertuch, 2016).

Les études concernant la régulation de la réparation des CDB et le choix du système de réparation ont été très nombreuses chez les levures, et ont grandement participé à la connaissance des effets de la chromatine et de ses modifications sur le NHEJ (Frigerio *et al.*, 2023).

5.6.2.7 Le NHEJ bactérien

Il a longtemps été considéré que le NHEJ était une particularité des eucaryotes, et que les procaryotes n'avaient que la recombinaison homologue pour réparer les CDB. En 2001, des études *in silico* ont apporté les premières preuves de l'existence du NHEJ chez certaines bactéries, en trouvant dans leurs génomes des séquences codant pour un homologue de Ku70/80 et une ADN ligase, regroupées sous forme d'opérons (ensemble de gènes contigus avec des fonctions reliées, très fréquent chez les bactéries) (Aravind and Koonin, 2001; Doherty *et al.*, 2001). Plus précisément, il a été montré que chez ces bactéries, il existe une seule protéine Ku, formant un homodimère, et que l'ADN ligase porte des fonctions variées. La protéine Ku bactérienne est bien plus petite que son homologue eucaryote (30-40 kDa chez les bactéries, 70-

85 kDa chez les eucaryotes), car elle ne possède pas les domaines vWA et SAP (Walker *et al.*, 2001). Il a été proposé que l'acquisition de ces domaines s'est faite plus tard dans l'évolution (Pitcher *et al.*, 2007a), et même que le Ku bactérien pourrait venir d'un phage (phage Mu), portant une protéine proche appelée Gam (d'Adda di Fagagna *et al.*, 2003), dont le génome aurait été intégré comme prophage par certaines bactéries (Figure 27). Cependant, il n'a jamais été montré de lien entre Gam et la réparation des CDB, son rôle étant plutôt de protéger les extrémités de l'ADN viral de la destruction par la bactérie. Bien que les séquences bactériennes et eucaryotes soient assez différentes, il a été montré que la structure de l'homodimère bactérien est très proche de celle de l'hétérodimère eucaryote (Pitcher *et al.*, 2007a).

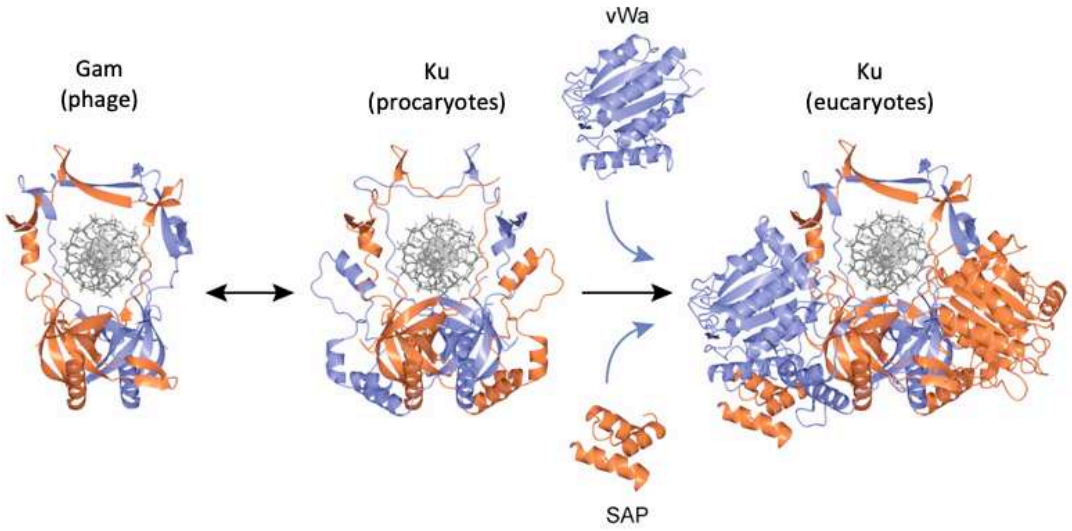


Figure 27 : Évolution des protéines Ku. Le scénario évolutif actuel propose que les protéines Ku actuelles proviennent de protéines ancestrales de phages appelées Gam. Celles-ci ont pu évoluer chez les hôtes bactériens pour obtenir la capacité d'interagir avec les ADN ligases, ce qui a formé les protéines Ku procaryotes. C'est enfin par l'ajout de domaines additionnels (SAP et vWA) qu'ont pu apparaitre les protéines Ku70/80 eucaryotes. D'après Pitcher RS et al. Annu Rev Microbiol. 2007.

L'ADN ligase retrouvée chez ces bactéries, appelée LigD, est composée de 3 domaines, pouvant être organisés dans des ordres différents (Shuman and Glickman, 2007) : un domaine ligase (LigDom), un domaine polymérase (PolDom), et un domaine portant une activité nucléase (NucDom) (Gong *et al.*, 2004). Le domaine polymérase fait partie de la famille AEP (*archaeo-eukaryotic primase*) (Guilliam *et al.*, 2015), et est actif sur de nombreux types d'extrémités d'ADN. En effet, ce domaine a une activité indépendante d'un brin instructeur, une activité d'extension d'amorce, de gap-filling, de synthèse face à des bases endommagées (8-oxoguanine par exemple), de dépassement de site abasique, et peut même utiliser des ribonucléotides *(Pitcher et al.*, 2007b). C'est également ce PolDom qui est à l'origine de l'interaction avec Ku et de la stabilisation de la synapse NHEJ (Pitcher *et al.*, 2005). Le domaine nucléase a des activités 3'

nucléase et 3' phosphatase, dépendantes du manganèse, et n'appartient pas à une famille connue de nucléases (Zhu and Shuman, 2005). Enfin, le domaine ligase représente la seule ADN ligase bactérienne spécialisée dans la ligation post-CDB, qui est stimulée par l'interaction avec Ku (Weller, 2002).

Ce n'est que plus tard que d'autres travaux ont montré que ces protéines permettaient ensemble la réparation de CDB chez les bactéries, formant un système NHEJ fonctionnel (Stephanou *et al.*, 2007). Par la suite de nombreux travaux ont permis de proposer le modèle suivant pour le NHEJ bactérien (Zhu and Shuman, 2010, 2007) (Figure 28): lors d'une CDB, l'homodimère Ku reconnait les extrémités de l'ADN lésé, et recrute LigD. Cette dernière, grâce à ses activités variées, a la capacité de traiter les extrémités de l'ADN au besoin avant de réaliser la ligation.

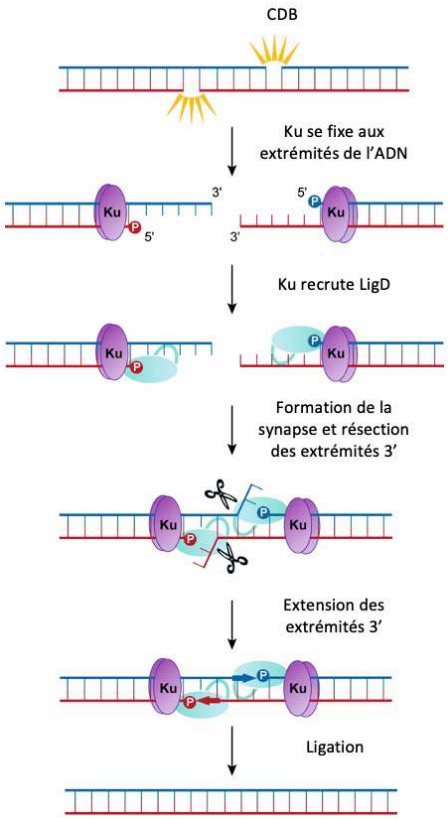


Figure 28 : Modèle des étapes du NHEJ bactérien. Après l'apparition d'une CDB, le dimère Ku se fixe sur les extrémités d'ADN, et recrute LigD en interagissant avec son domaine polymérase. LigD permet la formation de la synapse, ainsi que la maturation des extrémités de l'ADN et la ligation. D'après Pitcher RS et al. Annu Rev Microbiol. 2007

Plus récemment, il a été montré que le NHEJ bactérien est vraisemblablement plus complexe que cela, car il peut chez certaines espèces bactériennes impliquer d'autres partenaires protéiques, comme des hélicases (UvrD1 chez les mycobactéries (Sinha *et al.*, 2007)), ou des

facteurs pouvant réguler ce système de réparation (comme Sir2, une désacétylase (Li *et al.*, 2011)).

Certains des homologues de Ku sont parfois même codés par des gènes présents sur des plasmides, ce qui suggère qu'ils ont été acquis par transfert horizontal (Bertrand *et al.*, 2019). De plus, certaines espèces bactériennes (particulièrement des mycobactéries) portent plusieurs allèles codant pour des LigD, auxquelles il peut manquer un domaine ; voire même des gènes codant uniquement pour des ADN polymérases homologues du domaine polymérase de LigD (Bhattarai *et al.*, 2014; Gong *et al.*, 2005). Ces protéines peuvent se complémenter, et il est même possible que certaines servent de plateforme de recrutement pour les autres. C'est le cas par exemple chez les bactéries du genre *Streptomyces*, qui portent trois protéines différentes ayant les activités de ligase, polymérase et nucléase (Hoff *et al.*, 2016).

Il a été montré que certaines espèces bactériennes présentent, en plus des gènes de Ku et LigD, des gènes codant pour des ADN polymérases de la famille X. Celles-ci, plus proches de l'ADN polymérase β humaine que des ADN polymérases λ, μ et Tdt, n'ont pas de domaine BRCT mais portent en C-terminal un domaine additionnel, portant plusieurs activités nucléases (Prostova *et al.*, 2022). Elles peuvent cependant être séparées en 2 catégories, comme nous le verrons dans le chapitre 1 : un groupe canonique, avec les activités ADN polymérase et nucléase ; et un groupe non canonique qui n'a pas d'activité polymérase. L'ensemble de ces ADN polymérases X bactériennes semblent être davantage présentes chez les bactéries présentant l'ensemble Ku-LigD. Cependant, certaines bactéries possédant ces ADN polymérases X n'ont tout simplement pas les gènes codant pour Ku et LigD, ou un seul des deux, ou uniquement une partie des domaines de LigD. De plus, les groupes de bactéries présentant un système NHEJ sans activité nucléase semblent enrichis en ADN polymérases X, ce qui indique que ces dernières pourraient complémenter ce système grâce à leur activité nucléase. Pourtant, leur activité ADN polymérase semble accessoire pour le NHEJ chez ces bactéries (Nakane *et al.*, 2012a), et pas chez d'autres comme *Deinococcus radiodurans*, une bactérie connue pour sa résistance à l'irradiation et donc aux dommages de l'ADN (Bentchikou *et al.*, 2007).Ces ADN polymérases X ne sont pas présentes chez toutes les bactéries, leur présence est actuellement expliquée par des transferts horizontaux plutôt que verticaux (Prostova *et al.*, 2022), et leur association avec un système NHEJ est fréquente mais différente selon les *phyla*. De la même façon, l'association des gènes codant pour ces ADN polymérases X dans des opérons avec d'autres protéines liées

au NHEJ est inconstante, mais ces gènes sont fréquemment retrouvés proches d'autres gènes impliqués dans la maintenance du génome. Des séquences proches sont également retrouvées chez des archées, ce qui suggère que les transferts horizontaux ont même pu se faire jusqu'à ces organismes, un mécanisme déjà décrit (Nelson-Sathi *et al.*, 2015). Chez les bactéries, le NHEJ implique des facteurs variables. Cela suggère que ces mécanismes ont aussi pu être intégrés les uns après les autres par des transferts horizontaux de plasmides porteurs de gènes nouveaux.

5.6.3 Le NHEJ fait-il toujours des erreurs ?

De manière générale, le NHEJ est considéré comme le système de réparation des CDB faisant le plus d'erreurs. Cette idée repose sur plusieurs éléments :

- L'existence d'un système proche pour la réparation des CDB, le MMEJ ou alternative-EJ (Liang *et al.*, 1996; Sfeir and Symington, 2015), pouvant faire de nombreuses erreurs puisque son fonctionnement est basé sur la résection de l'ADN lésé jusqu'à trouver une micro-homologie pouvant permettre une ligation ;
- L'adaptabilité du NHEJ classique aux différentes formes d'ADN lésé, qui peut entrainer une perte de quelques nucléotides, ou l'ajout d'autres nucléotides selon le traitement des extrémités de l'ADN ;
- Le NHEJ peut être responsables de translocations et de réarrangements illégitimes, entre des régions éloignées du génome, que ce soit de façon intra- ou interchromosomique (Guirouilh-Barbat *et al.*, 2004), bien que en théorie, dès lors que Ku70/80 et DNA-PKcs prennent place sur l'ADN, les extrémités lésées sont maintenues ensemble.

Il existe cependant des organismes chez lesquels la réparation des CDB par le NHEJ est très efficace et très fidèle. C'est le cas du NHEJ impliqué dans la réparation des CDB programmées chez *Paramecium tetraurelia* (Bétermier *et al.*, 2000), que nous allons voir ci-après et qui est le sujet principal des travaux de cette thèse.

6 Les paramécies

Les paramécies ont été parmi les premiers organismes unicellulaires à être observés au microscope, au XVIIème siècle (Dobell *et al.*, 1932). Ces organismes sont des eucaryotes unicellulaires ciliés. Leurs cils sont sensoriels et vibratoires et leur permettent de se déplacer dans leur milieu (eau douce ou eau stagnante) (Abello, 2019). Une cellule de paramécie est ovoïde et grande (plus de 100 μm^3), beaucoup plus grande que les bactéries, algues unicellulaires ou levures dont elle se nourrit *via* sa « bouche », une structure lui permettant de générer des vacuoles digestives et de se nourrir par endocytose. Un cycle cellulaire de *Paramecium* dure 5 heures à sa température optimale de croissance de 27°C (Abello, 2019).

La paramécie présente une particularité notable : un dimorphisme nucléaire. Elle porte en effet des noyaux diploïdes appelés micronoyaux (MIC) et un noyau polyploïde (contenant jusqu'à 1600 copies du génome) appelé macronoyau (MAC) (Abello, 2019) (Figure 29). Ces deux types de noyaux ont des fonctions bien distinctes : le MIC est germinal, c'est-à-dire qu'il ne sert pas à l'expression des protéines mais à la transmission du génome aux cellules filles lors de la reproduction ; et le MAC est le noyau somatique, qui sert quant à lui à l'expression des gènes. Il est perdu par fragmentation à chaque cycle sexuel et remplacé par un nouveau noyau généré à partir des MIC (Abello, 2019). Ce remplacement, comme nous le verrons après, génère de nombreux réarrangements programmés du génome, dont des CDB.

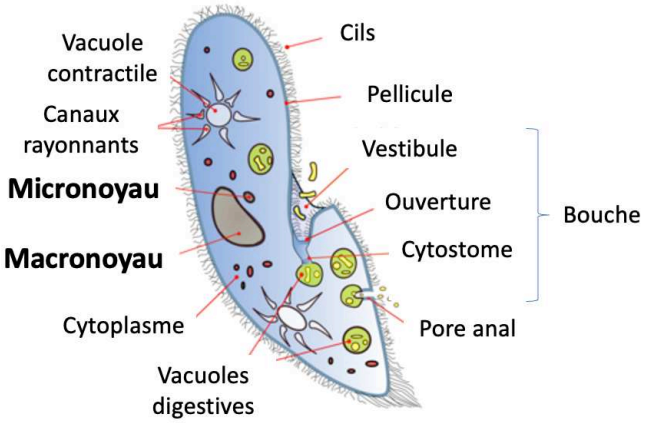


Figure 29 : Anatomie de Paramecium. D'après Wikipedia (https://en.wikipedia.org/wiki/Paramecium).

Ici, nous nous intéresserons à une espèce en particulier de paramécie : *Paramecium tetraurelia*.

6.1 Le cycle de vie de la paramécie

6.1.1 Le cycle végétatif

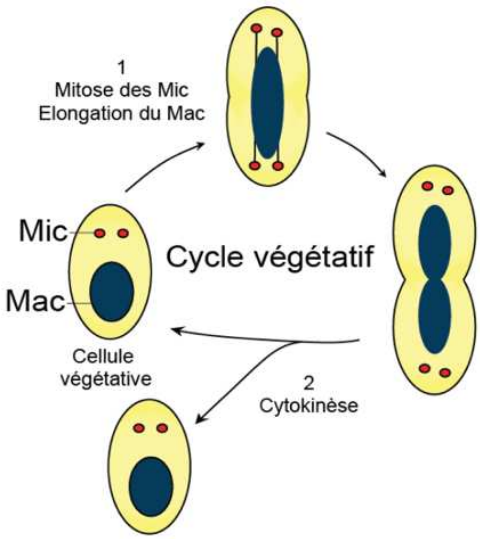


Figure 30 : Cycle végétatif de P. tetraurelia. Le cycle végétatif de P. tetraurelia, qui se produit en conditions normales (sans stress ou carences alimentaires, se compose de deux grandes étapes : une mitose des MIC et une division asymétrique du MAC, puis une étape de cytokinèse, c'est à dire la formation de deux cellules filles par séparation des membranes. D'après Spécialisation de Ku80c dans le couplage entre coupure et réparation de l'ADN lors des réarrangements programmés du génome chez Paramecium tetraurelia. Arthur Abello. 2019.

À sa température optimale de croissance (27°C), la paramécie réalise en laboratoire une division végétative toutes les 5 à 6 heures. Lors de cette division (Figure 30), les MIC réalisent une mitose en gardant leur enveloppe nucléaire, tandis que le MAC se divise autrement : après la réplication de l'ADN, il s'allonge puis se scinde en deux en séparant de façon aléatoire les copies des chromosomes (Abello, 2019). Ce mode de fonctionnement, risqué, peut entrainer un manque de copies de certains gènes dans un des deux noyaux descendants, mais la forte ploïdie de la paramécie garantit qu'aucun gène ne manque dans les noyaux descendants (Abello, 2019). La division cellulaire prend fin lors de la séparation des deux cellules filles. Comme toutes les cellules, les paramécies ne peuvent pas se diviser indéfiniment. Elles peuvent se diviser environ 200 fois, jusqu'à ce que leur vitesse de division diminue jusqu'à un arrêt total (Gilley and Blackburn, 1994) ou à un déséquilibre chromosomique entre les MAC néoformés qui pourrait finir par survenir.

6.1.2 Le cycle sexuel

Il semble cependant qu'un mécanisme ait été favorisé par l'évolution comme solution aux problèmes du génome du MAC : la reproduction sexuée (Abello, 2019). Celle-ci a pour avantage de générer de nouveaux MAC dans les cellules filles, pour lesquelles le « compteur » de divisions est donc remis à zéro (Abello, 2019). Il existe deux types de cycles sexuels : la conjugaison et l'autogamie, qui semblent être déclenchées par les situations de carences alimentaires (du moins en laboratoire).

6.1.2.1 La conjugaison

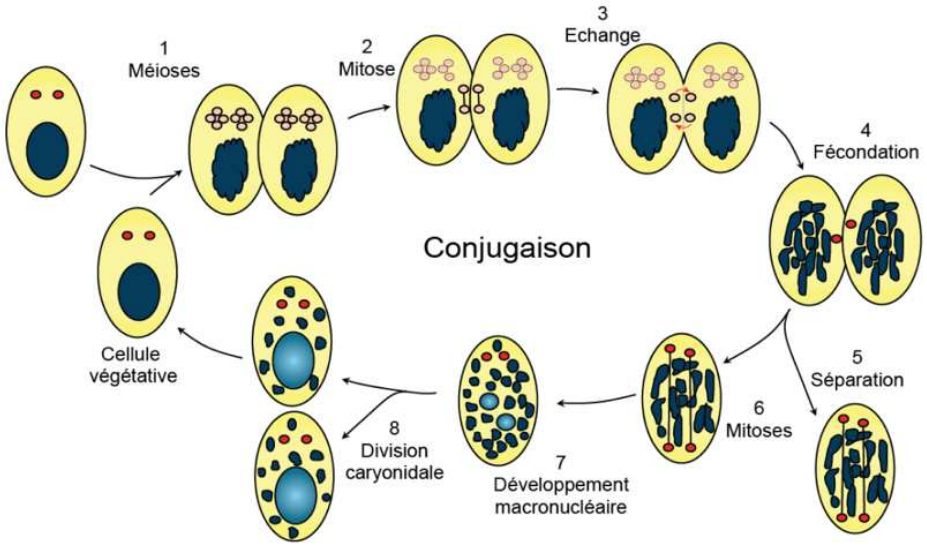


Figure 31 : Conjugaison chez P. tetraurelia. Ce système de reproduction sexuée se produit en cas de stress ou de carence alimentaire. Il se déroule en 8 étapes : méiose des MIC, mitose d'un des MIC résultants, échange de MIC entre les deux cellules, fécondation entre les deux MIC, séparation des deux cellules, puis pour chaque cellule : mitoses des MIC, développement du macronoyau, division caryonidale. D'après Spécialisation de Ku80c dans le couplage entre coupure et réparation de l'ADN lors des réarrangements programmés du génome chez Paramecium tetraurelia. Arthur Abello. 2019.

Chez les paramécies, il existe deux types sexuels : le type E (*even*) et le type O (*odd*), découverts en 1937 (Sonneborn, 1937). Lors de la conjugaison (Figure 31), une cellule E et une cellule O s'associent *via* leurs bouches après avoir réalisé une méiose des MIC, formant 8 noyaux haploïdes. L'un de ces noyaux est sélectionné (de façon encore inconnue) et répliqué par mitose. Un pont cytoplasmique est ensuite formé entre les deux cellules et leur permet d'échanger leurs noyaux gamétiques. Ceux-ci fusionnent, formant un noyau diploïde. Les deux cellules se séparent alors, et dans chacune, deux mitoses donnent ensuite lieu aux deux MIC des cellules filles et à leurs deux MAC. Les MAC peuvent ensuite se développer par de nombreuses

réplications du génome et de nombreux réarrangements. Dans le même temps, la cellule subit une division caryonidale : les nouveaux MIC passent par une mitose et sont ségrégés (avec les nouveaux MAC en développement) dans les deux cellules filles, appelées caryonides. Pendant tout ce processus, le MAC de chacune des cellules se détruit, par débobinage de l'ADN, qui est ensuite dégradé de façon inconnue. Par la suite, les caryonides, par divisions végétatives, vont donner lieu à des populations caryonidales.

6.1.2.2 L'autogamie

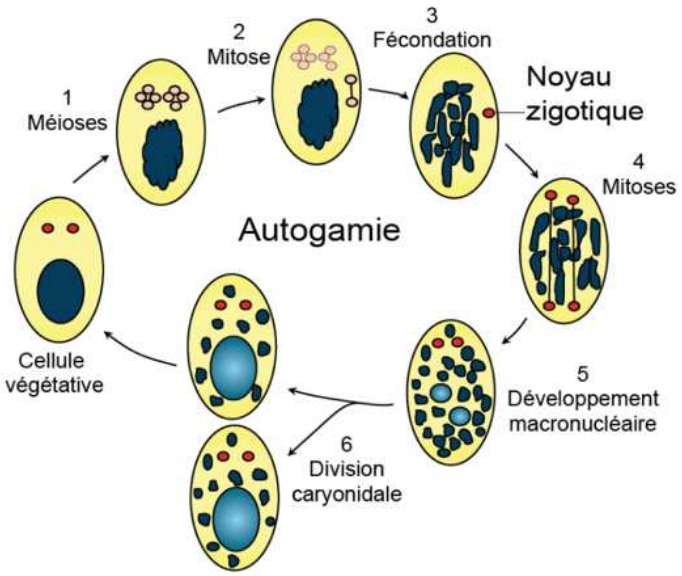


Figure 32 : Autogamie chez P. tetraurelia. Ce système de reproduction sexuée se produit en cas de stress ou de carence alimentaire. Il se déroule en 6 étapes : méiose des MIC, mitose d'un des MIC résultants, fécondation entre les deux MIC pour former un noyau zygotique (et début de la fragmentation des MAC), mitoses des MIC, développement du macronoyau, division caryonidale. D'après Spécialisation de Ku80c dans le couplage entre coupure et réparation de l'ADN lors des réarrangements programmés du génome chez Paramecium tetraurelia. Arthur Abello. 2019.

L'autogamie (Figure 32), quant à elle, commence comme la conjugaison : les MIC subissent une méiose qui donne lieu à 8 nouveaux MIC haploïdes. Un de ces noyaux est sélectionné et répliqué à l'identique par mitose, tandis que les 7 autres sont détruits. Les deux noyaux haploïdes sont ensuite fusionnés, ce qui donne lieu à une autofécondation, donc à un noyau diploïde strictement homozygote. A partir de là, tout se passe comme lors de la conjugaison : mitoses, développement des MAC, division caryonidale (et destruction des MAC parentaux) (Sonneborn, 1937).

6.2 Le génome des paramécies

La paramécie a un génome compact : 78 % de son génome est codant, les régions intergéniques sont courtes (352 bases en moyenne) et les introns sont petits (25 bases en moyenne) (Aury *et al.*, 2006). Une des particularités de la paramécie est aussi qu'elle est un contre-exemple de l'universalité du code génétique : elle ne porte qu'un codon stop (opale): UGA ; les codons stops canoniques ambre et ocre (UAG et UAA) codent tous deux pour la glutamine (Preer *et al.*, 1985), ce qui pourrait être un mécanisme de défense contre les invasions du génome par les transposons (Klobutcher and Herrick, 1997) : pour se propager, un transposon devrait utiliser cette variation du code génétique, qui modifierait ses protéines, en y ajoutant des extensions C-terminales.

6.2.1 L'importance de la polyploïdie

La paramécie étant unicellulaire, on s'attendrait à une certaine simplicité dans son fonctionnement et son organisation. Cependant, comme indiqué précédemment, le génome utile à l'expression des protéines contenu dans le MAC y est présent en de très nombreuses copies (jusqu'à 1600). Il semble que les 40 000 gènes de *Paramecium tetraurelia* (le nombre le plus élevé en comparaison avec tous les métazoaires et les plantes) trouvent leur origine dans quatre duplications globales du génome. C'est ce qui a été proposé en 2006 (Aury *et al.*, 2006) lors du séquençage du génome de cet organisme. Ces duplications ont entrainé l'apparition de gènes paralogues (on parle ici d'ohnologues) dans le génome de *P. tetraurelia*, pouvant avoir des fonctions proches (voire des spécialisations) et des niveaux d'expression variables : au lieu d'avoir deux copies d'un gène avec les mêmes fonctions, ces deux copies se sont spécialisées, avec chacune une fonction distincte. Ces duplications seraient également liées aux événements de spéciations parmi les paramécies : l'ensemble des *Paramecium aurelia* (*tetraurelia*, *primaurelia*, *octaurelia*, …) seraient issues de la duplication la plus récente, et les duplications les plus anciennes auraient entrainé l'apparition d'autres espèces comme *Tetrahymena thermophila* (Figure 33).

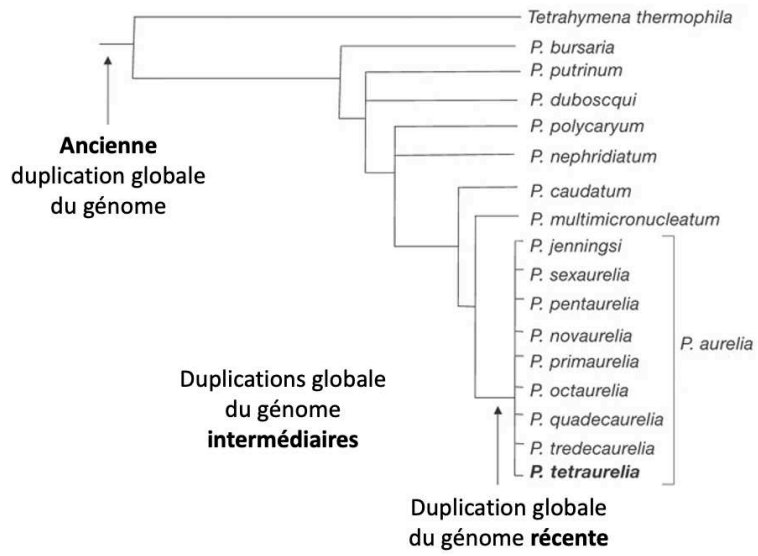


Figure 33 : Arbre phylogénétique retraçant les grandes duplications globales du génome ayant mené à l'apparition des Paramecium aurelia. La duplication la plus ancienne a entrainé une divergence entre les espèces Paramecium et Tetrahymena, et les plus récentes ont permis l'apparition des différentes espèces de Paramecium, puis des sous espèces de P. aurelia. D'après Aury et al. Nature. 2006

On peut cependant se demander quel est l'avantage évolutif pour ces organismes à avoir autant de copies du génome à répliquer, réparer et exprimer, et à posséder deux types de noyaux. D'après Klobutcher et Herrick (Klobutcher and Herrick, 1997), l'apparition de deux noyaux pourrait provenir d'invasions successives du génome par des éléments transposables. D'après cette hypothèse, ces éléments transposables auraient envahi ce qui est aujourd'hui le MIC (étapes d'invasion puis de *bloom* au cours de laquelle ils se seraient multipliés dans tout le génome), et auraient dégénéré pour devenir ce qu'on appelle des IES (séquences internes éliminées, ou *Internal Eliminated Sequences*). Par conséquent, la possibilité d'éliminer ces séquences pour former un noyau somatique autonome (MAC) est un avantage, qui permet de se débarrasser du problème de ces transposons dans le cadre de l'expression du génome. Le mode de reproduction sexuée de *Paramecium* est lui aussi un avantage, puisqu'il permet de générer de la diversité tout en se débarrassant des IES, et des éventuelles mutations délétères.

La forte ploïdie de ces organismes pourrait quant à elle être un avantage évolutif. En effet, dans un noyau contenant de 800 à 1600 copies du génome, l'apparition de mutations est favorisée : si une mutation sur un allèle est délétère, ses effets seront noyés dans la masse des autres allèles. Ce mode de fonctionnement est particulièrement intéressant au cours du cycle de vie végétatif de *Paramecium*. En effet, dans ce cadre, seul le MIC est répliqué par mitose , alors que le MAC est simplement divisé en deux, sans assurance que les copies initiales des gènes soient réparties équitablement (Maurer-Alcalá and Nowacki, 2019).

6.2.2 Les réarrangements programmés du génome

Il a été montré que les MIC de *P. tetraurelia* contiennent environ 100 mégabases (100 Mb, soit 100 million de paires de bases), mais le MAC ne contient que 72 Mb (Aury *et al.*, 2006). Cette différence s'explique par les réarrangements que rencontre le génome issu du MIC lors du développement du MAC. En effet, lors de la génération du MAC, le génome doit passer de deux copies (dans le MIC) à plus de 800, et une partie du génome du micronoyau doit être éliminée, de plusieurs façons (Figure 34).

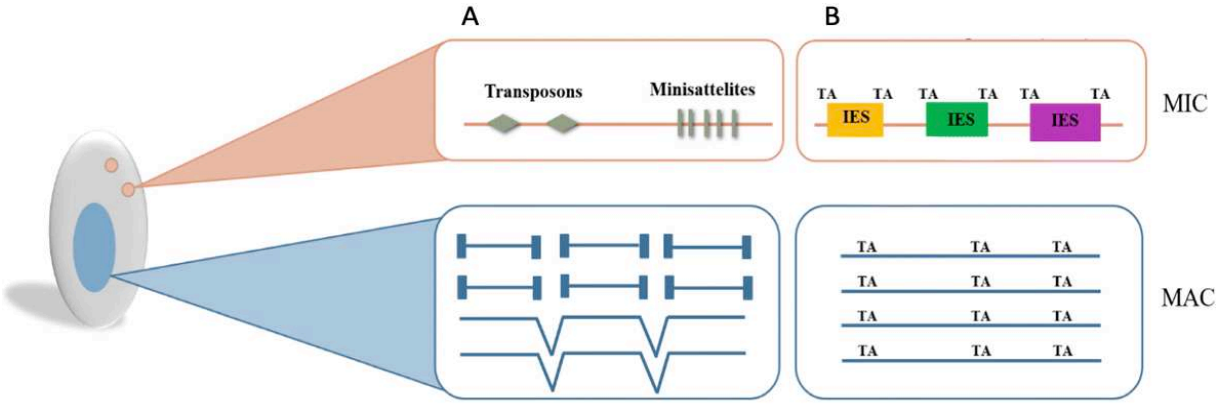


Figure 34: Les deux types de réarrangements du génome chez P. tetraurelia. En orange : les séquences présentes dans le MIC ; en bleu : les séquences équivalentes présentes dans le MAC. A : Les réarrangements imprécis. Ceux-ci visent à éliminer les séquences minisatellites et les transposons. B : les réarrangements précis. Ceux-ci concernent l'élimination des IES, qui se fait de façon précise. D'après Rzeszutek I et al. Cell Mol Life Sci. 2020.

6.2.2.1 Les réarrangements imprécis

Un des types de réarrangements du génome chez la paramécie est le réarrangement dit « imprécis » (Rzeszutek *et al.*, 2020). Ce type de réarrangement se concentre sur les régions intergéniques et les extrémités des chromosomes, pour l'élimination de transposons et de séquences répétées. Ces réarrangements sont dits imprécis car les bornes des régions à éliminer sont variables, et la réparation peut soit être faite par ligation soit par télomérisation, ce qui coupe le chromosome concerné. L'élimination de ces séquences est réalisée par PiggyMac (ou Pgm), qui est une transposase « domestiquée » par la paramécie à partir de la transposase piggybac (Baudry *et al.*, 2009).

6.2.2.2 Les réarrangements précis : l'élimination des IES

6.2.2.2.1 Les IES

Au sein du génome des MIC, certaines séquences, appelées IES (pour *Internal Eliminated Sequences* ou séquences internes éliminées) doivent être éliminées précisément lors du développement du MAC. Il y en a 45 000 par génome, elles représentent 3 Mb, et on en retrouve une tous les 1,6 kb environ (Arnaiz *et al.*, 2012). Ces IES peuvent être retrouvées partout dans le génome et 47% des gènes sont interrompus par au moins une IES. Par conséquent, leur élimination est essentielle pour la survie de la cellule.

Il a été montré que la taille des IES est variable, mais suit une périodicité de 10 pb (soit environ 1 tour d'hélice) ; de plus, aucune ne fait moins de 26 pb, et il en existe peu qui font entre 32 et 45 pb (Arnaiz *et al.*, 2012). Une explication possible de ces paramètres serait la contrainte topologique lors de l'excision des IES (Abello, 2019). Un élément essentiel concernant les IES est la séquence située à leurs extrémités, qui est toujours la même : un dinucléotide TA, entouré d'autres nucléotides peu conservés. Par ailleurs, les IES sont généralement riches en nucléotides A et T (70-100%) et leurs extrémités sont reconnaissables car ce sont des zones de baisse drastique du taux de nucléotides G et C (Maurer-Alcalá *et al.*, 2018).

L'origine des IES pourrait se trouver dans des invasions du génome des paramécies par des transposons Tc1/mariner (Dubois *et al.*, 2012). En effet, les séquences TA situées en bordures des IES sont aussi la séquence consensus pour ce type de transposon ; et il a été montré que certaines des plus grandes IES portent des gènes homologues de ceux de Tc1/mariner (Dubois *et al.*, 2012). À la suite des invasions du génome, les séquences des transposons auraient fini par dégénérer et être raccourcies, jusqu'à la taille minimale de 26 pb retrouvée aujourd'hui. Les IES les plus récentes seraient donc les plus longues, et les plus anciennes seraient les plus courtes.

Avec ces caractéristiques communes avec les transposons de la famille Tc1/mariner, il semblerait logique que le mécanisme d'excision des IES soit semblable à celui de cette famille, mais ce n'est pas le cas: alors que les transposases Tc1/mariner laissent les deux dinucléotides TA sur le chromosome après l'excision, le mécanisme d'excision des IES ne laisse qu'un seul site TA (Figure 35).

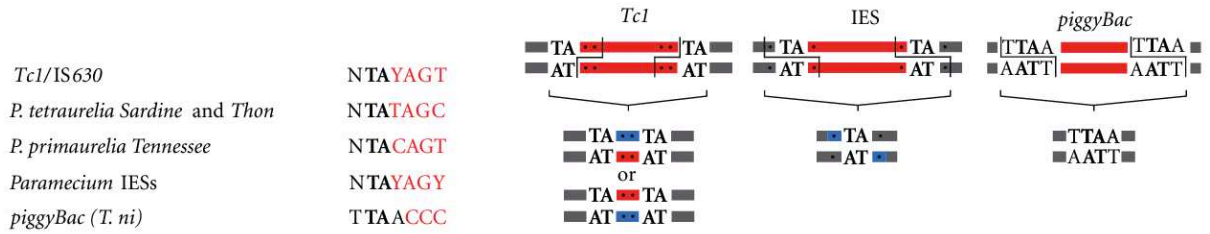


Figure 35 : Comparaison des IES et des mécanismes d'excision des transposons Tc1, piggybac et des IES de Paramecium. À gauche : séquences consensus aux extrémités des transposons ou IES indiqués à gauche. À droite : comparaison des mécanismes d'excision des transposons Tc1 et piggyBac avec les IES de Paramecium. D'après Dubois E et al. Int J Evol Biol. 2012.

Dès le début des années 2000, il a été montré que l'élimination des IES passait par l'introduction de CDB à leurs extrémités (Gratias and Bétermier, 2003), laissant 4 bases sortantes en 5', avec le TA au milieu. La base en 5' est ensuite éliminée, et la micro-homologie entre les deux TA permet de former une synapse, qui facilite ensuite la réparation. L'IES éliminée subit exactement la même chose, ce qui entraine sa circularisation si sa taille le permet (sinon, l'IES s'assemble avec d'autres IES courtes, ce qui permet sa circularisation) (Bétermier *et al.*, 2000).

6.2.2.2.2 La reconnaissance des IES

Le processus permettant aux IES d'être reconnues commence dès le début de la reproduction, lors de la prophase de méiose des MIC (Figure 36). À cette étape, le génome des MIC est transcrit bidirectionnellement par l'ARN polymérase II, ce qui forme de longs ARN qui sont ensuite découpés par des ribonucléases Dicer-like (Dcl2 et 3 chez *Paramecium*) pour former des ARN scans ou scnRNA (Abello, 2019). Ces derniers font 25 nt chez *Paramecium*, et sont ensuite transportés dans le cytoplasme lors de la fragmentation du MAC et pris en charge par les protéines Ptiwi1/9. Ces protéines les transportent à leur tour vers le MAC parental, et les scnRNA scannent son génome, en s'associant aux protéines Nowa1 et 2 (Nowacki *et al.*, 2005), avec l'aide de Ptmb220. Les ARN correspondant à des séquences du MAC (n'étant donc pas des IES) se fixent à des ARN non codants transcrits au sein du MAC et sont stabilisés par Ptiwi09. La protéine Gtsf1a pour rôle de reconnaitre ces complexes et entrainer l'ubiquitinylation de Ptiwi09, provoquant sa dégradation ainsi que celle des ARN associés. (Charmant *et al.*, 2023). Ainsi, seuls les scnRNA sans ADN correspondant dans le MAC, c'est-à-dire ceux correspondant aux séquences absentes du MAC (les IES) restent. Ils sont ensuite transportés vers le MAC en formation, et permettent de reconnaitre les IES et de guider leur élimination. Les IES éliminées sont circularisées (après concaténation si elles sont courtes) et transcrites puis découpées par Dcl5 en iesRNA, qui après association avec les protéines Ptiwi10/11 permettent eux aussi de

guider l'élimination des IES dans le génome en pleine réplication (c'est donc un mécanisme autocatalytique) (Allen *et al.*, 2017; Rzeszutek *et al.*, 2020) (Figure 37).

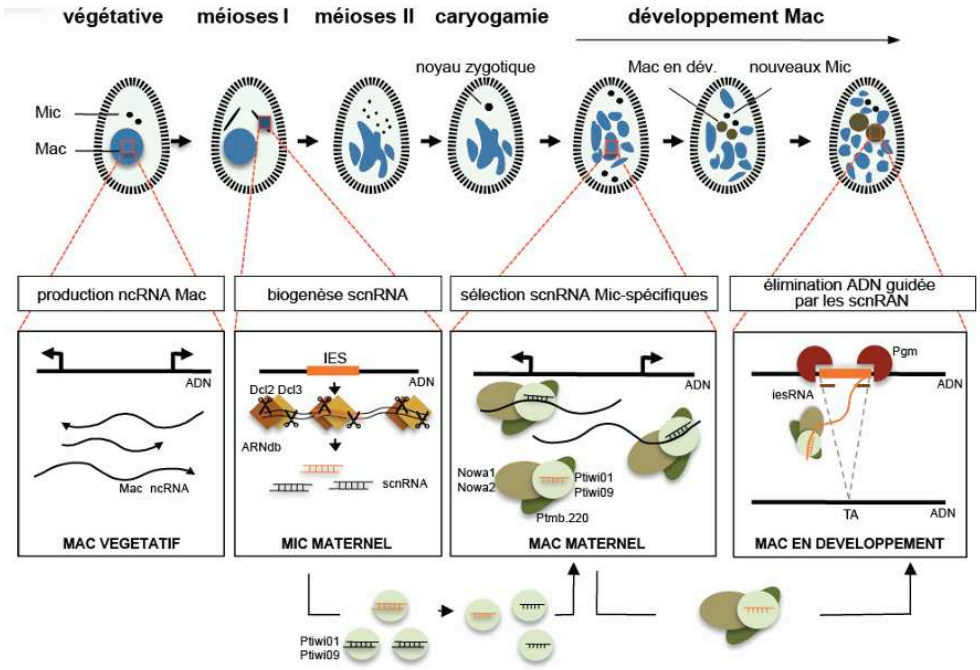


Figure 36 : La reconnaissance des IES est aidée par des ARN non codants. Lors de la méiose des MIC, leur génome est transcrit de façon bidirectionnelle et découpé en ARN scans (scnARN) par des ribonucléases. Ces scnARN sont ensuite transportés dans le MAC maternel par les protéines Ptiwi et pris en charge par les protéines Nowa. Les ARN pouvant s'hybrider avec l'ADN du MAC sont retenus, et ceux sans correspondance sont ensuite transportés vers le MAC en cours de développement, où ils se fixent à l'ADN, de façon à cibler les IES pour leur élimination. D'après Bétermier et al. Microbiol Spectr. 2014

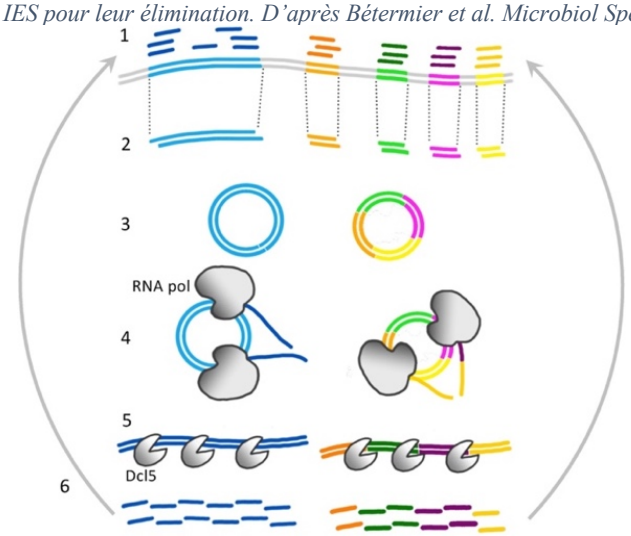


Figure 37 :Les IES éliminées et circularisées sont transcrites et permettent l'élimination d'autres IES. La reconnaissance des IES par les scnARN permet leur élimination (1 et 2). Ces IES sont ensuite circularisées (3) et transcrites en iesARN (4). Ces iesARN sont ensuite découpés par les protéines Dicer like (5 et 6), et peuvent alors être utilisés pour reconnaitre les IES et guider leur élimination (6 et 1). D'après Allen SE et al. Cell. 2017

Cependant, la façon dont ces scnRNA et iesRNA permettent le clivage des IES est encore mal comprise. Il semble que pendant les réarrangements du génome du MAC, une ARN polymérase transcrit le génome, grâce au facteur de transcription TFIIS4. Au même moment, les

scnRNA et iesRNA, pris en charge par les protéines Ptiwi, peuvent s'hybrider aux ARN transcrits. Un complexe protéique appelé PRC2 est alors recruté, et sa sous unité Ezl1 triméthyle les lysines 9 et 27 des histones H3 environnants. Une histone chaperone appelée Spt16 permet ensuite la reconnaissance de ces marque d'hétérochromatine, et entraine le recrutement de Pgm, qui peut alors reconnaitre les extrémités des IES et les éliminer (Drews *et al.*, 2022) (Figure 38).

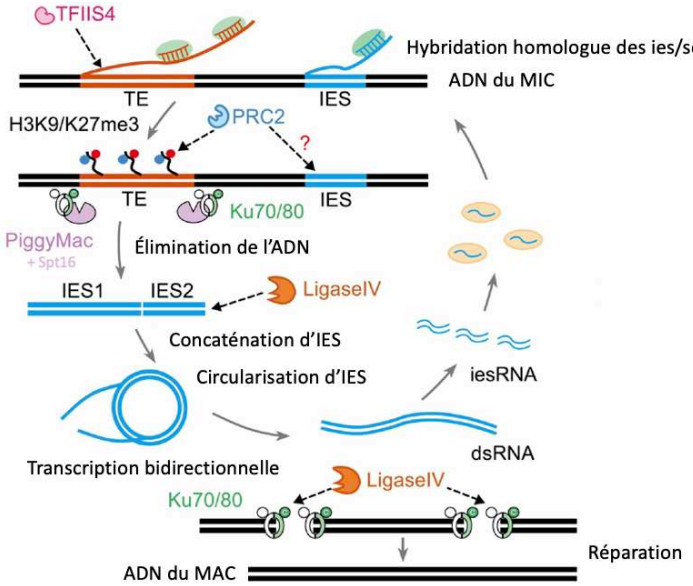


Figure 38 : Ciblage et élimination des IES par PiggyMac. Les ARN (scn et ies) permettant la reconnaissance des IES se fixent aux transcrits formés par une ARN polymérase et TFIIS4. Leur fixation entraine le recrutement de PRC2, qui introduit des modifications d'histones (H3K9me3 et H3K27me3). PiggyMac, accompagné de Ku70/80, peut enfin reconnaitre ces séquences grâce à Spt16. De là, l'élimination des IES est possible, puis leur transcription bidirectionnelle et la formation d'autres iesRNA ; en parallèle, les CDB introduites par PiggyMac peuvent être réparées. D'après Gao Y et al. Trends Genet. 2023.

6.2.2.2.3 L'élimination des IES

À la suite du séquençage et de l'annotation du génome de *Paramecium*, il a été découvert que c'est la protéine PiggyMac qui est impliquée dans l'élimination des IES. Celle-ci est une version domestiquée de la transposase Piggybac (Baudry *et al.*, 2009). Une transposase est une enzyme codée par un transposon, qui permet son introduction dans un génome (Aziz *et al.*, 2010).

6.2.2.2.3.1 Le transposon (et la transposase) Piggybac

Un transposon est un élément génétique mobile c'est-à-dire un morceau d'ADN qui peut se déplacer dans un génome ou entre des génomes. Ces éléments sont à l'origine de nombreuses fonctions cellulaires, de réarrangements et de mutations présents chez les eucaryotes, existent sous de nombreuses formes et sont présents à des endroits spécifiques du génome. Leur

expression est finement régulée, car ils peuvent modifier des voies de signalisation des cellules, entrainer des réponses immunitaires, et affecter les cellules somatiques et germinales (Bourque *et al.*, 2018).

PiggyBac est connu pour s'intégrer sur des sites TTAA, et ne laisser aucune trace lors de son excision du génome. Ce transposon peut s'introduire chez de nombreux hôtes, de la levure aux mammifères en passant par les insectes, chez qui a il a été découvert dans les années 1980 (Yusa, 2015). Ce transposon est d'ailleurs utilisé comme outil en génétique pour ces raisons (Woodard and Wilson, 2015).

Son mécanisme de transposition (Figure 39) commence par une hydrolyse par la transposase aux extrémités du transposon, ce qui libère des extrémités 3'OH (Q. Chen *et al.*, 2020). Ces extrémités réalisent ensuite une transestérification sur l'autre brin, 4 nucléotides plus loin, ce qui permet la formation de structures en épingle à cheveux aux extrémités du transposon. PiggyBac réalise ensuite une autre hydrolyse, cette fois aux extrémités du transposon libre, pour libérer l'extrémité 3'OH et une extrémité débordante TTAA. L'extrémité 3'OH réalise une nouvelle transestérification sur le site TTAA ciblé, ce qui permet l'introduction du transposon, avec une dernière étape de rétablissement d'une liaison phosphodiester.

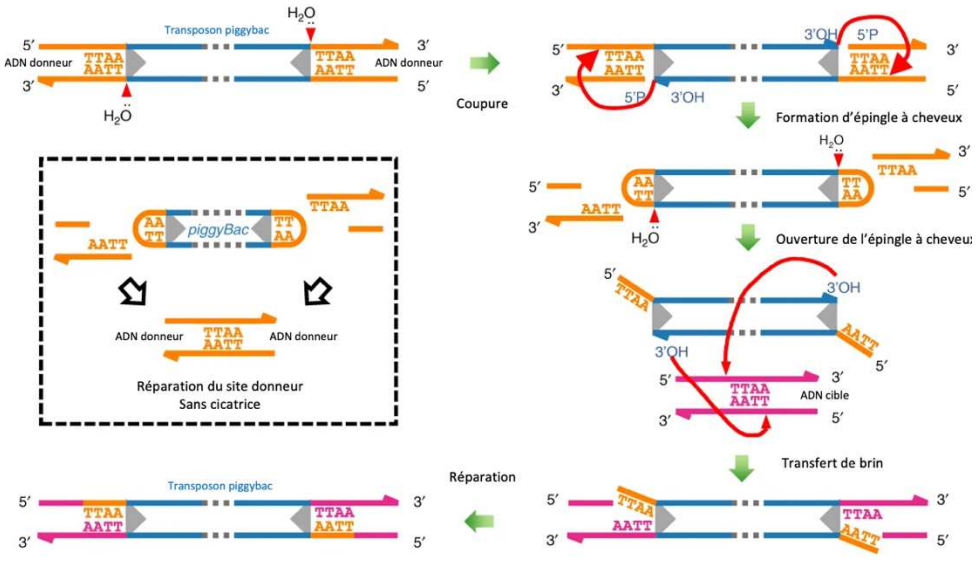


Figure 39 : Mécanisme d'excision du transposon piggybac. Après l'introduction d'une cassure simple brin de chaque côté du transposon par la transposase, l'extrémité 3'OH libérée se lie à l'extrémité 5'P de la séquence TTAA, ce qui forme une structure en épingle à cheveux, et l'ADN du transposon est libre. Le site donneur est quant à lui réparé, sans traces. Cette épingle est ensuite ouverte par la transposase, et l'extrémité 3'OH libérée attaque l'ADN cible, de chaque côté du site de reconnaissance TTAA. L'ADN est inséré dans l'ADN cible, et les cassures de chaque côté sont réparées. D'après Chen Q et al. Nat Commun. 2020

La transposase piggybac est composée de plusieurs domaines (Figure 40) : un domaine N-terminal désordonné, un premier domaine de dimérisation et de fixation à l'ADN (*Dimerization and DNA Binding Domain* ou DDBD), un domaine catalytique interrompu par une insertion et portant les trois aspartates catalytiques, puis un autre domaine DDBD et enfin un domaine C-terminal riche en cystéines (Q. Chen *et al.*, 2020).

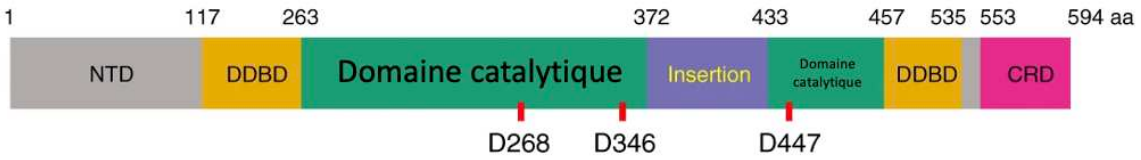


Figure 40: PiggyBac. PiggyBac est composée d'un domaine N-terminal, de deux domaines de dimérisation (en jaune), d'un domaine catalytique interrompu par une insertion, et d'un domaine C-terminal riche en cystéines. D'après Chen Q et al. Nat Commun. 2020

6.2.2.2.3.2 PiggyMac

Des travaux ont montré que le génome de *Paramecium* porte un gène présentant une forte homologie avec la transposase PiggyBac, qui a donc été nommé PiggyMac (Baudry *et al.*, 2009). Cette protéine présente entre autres choses les trois résidus catalytiques de piggybac, ainsi qu'un domaine riche en cystéines situé en C-terminal, absent chez piggybac.

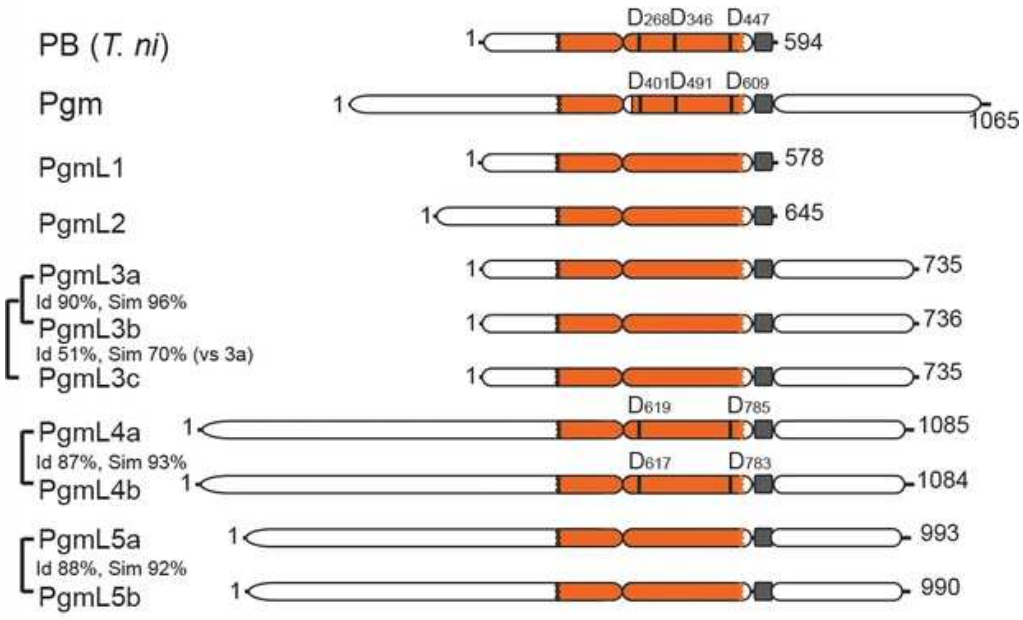


Figure 41 : PiggyBac, PiggyMac et les Pgm-like. PiggyMac (Pgm) présente une forte homologie avec PiggyBac (PB), et présente trois aspartates catalytiques. Les protéines Pgm-like sont elles aussi des homologues proches, mais ont des tailles variables et ne possèdent pas tous les aspartates catalytiques, ce qui les prive d'activité enzymatique. D'après Bischerour J et al. Elife. 2018

La présence en une seule copie de ce gène par génome haploïde, ainsi que l'absence de séquences répétées et inversées (signature des éléments transposables), a permis de déduire que ce gène est une évolution domestiquée du gène de Piggybac. Pgm présente cependant une extension C-terminale avec une structure en doigt de zinc, équivalente au CRD chez piggybac (Guérineau *et al.*, 2021). D'après ces travaux, le rôle de Pgm est assez large dans les réarrangements programmés du génome chez *Paramecium* : cette enzyme serait impliquée dans plusieurs types de réarrangements (imprécis, élimination des IES).

En 2018 (Bischerour *et al.*, 2018), une étude a montré que le génome de *Paramecium* permettait aussi l'expression de protéines proches de Pgm, appelées PgmL (Pgm-*like*, Figure 41), qui permettent malgré leur absence d'activité enzymatique de favoriser et stabiliser la localisation nucléaire de Pgm, pour permettre une élimination précise des IES. Il a été proposé que les PgmL forment un complexe avec Pgm, et que c'est cet ensemble qui permet de cibler correctement les extrémités des IES : s'il manque des partenaires, le complexe n'est pas (ou incorrectement) actif (Figure 42). L'ensemble de ces protéines sont exprimées pendant l'élimination des IES, et sont localisées dans le génome des MAC en formation.

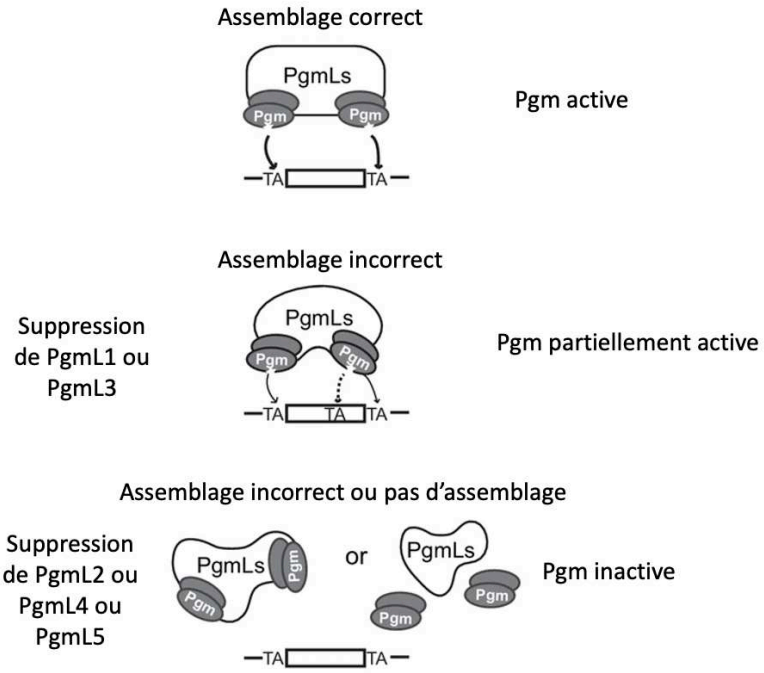


Figure 42 : Le complexe de Pgm et des Pgm like nécessite tous les PgmL. Avec tous les partenaires, le complexe Pgm-PgmL se forme correctement et a une activité normale. La suppression de PgmL1 ou 3 ne permet pas une activité correcte, et la suppression de PgmL2, 4 ou 5 supprime totalement l'activité de Pgm. D'après Bischerour J et al. Elife. 2018.

Comme nous l'avons vu, PiggyMac est une transposase dérivée de la transposase PiggyBac. Et comme cette dernière, elle a pour substrat les séquences TA aux extrémités des IES et laisse après coupure un site TA sur le chromosome (Elick *et al.*, 1996). Leurs mécanismes catalytiques semblent donc être très semblables. Il a de plus été montré que l'élimination des IES par Pgm nécessite une interaction entre les deux extrémités TA des IES (Gratias *et al.*, 2008).

Enfin, il a été montré qu'un partenaire essentiel à la coupure par Pgm n'était autre que Ku70/80, le premier acteur du NHEJ (Marmignon *et al.*, 2014) (Figure 43).

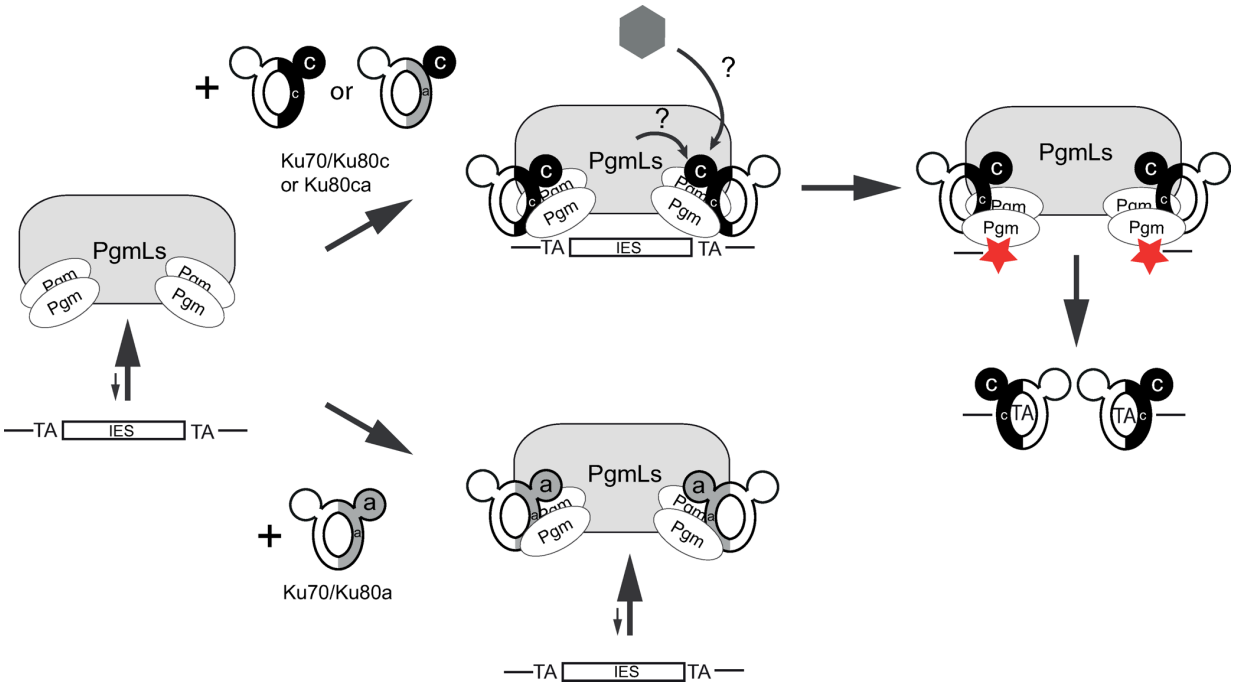


Figure 43 : L'activité de Pgm nécessite son interaction avec Ku70/80c. Pour être actif, Pgm doit se lier à Ku70/80c (et possiblement d'autres partenaires), ce qui permet l'introduction de CDB et le chargement de Ku70/80 sur les extrémités d'ADN. Cependant, bien qu'une interaction soit possible entre le complexe Pgm et le dimère Ku70/80a, celle-ci ne favorise pas l'activité de Pgm. D'après Abello A et al. PLoS Genet. 2020

6.2.2.2.3.3 La réparation des CDB introduites par PiggyMac

À la suite de l'introduction des CDB par Pgm, il est indispensable de les réparer, dans le génome du MAC mais aussi au sein des IES, qui doivent être circularisées pour permettre la formation des iesRNA.

Comme le suggère l'implication de Ku70/80 dans le clivage des extrémités des IES par PiggyMac, il semble que le système utilisé pour réparer ces CDB repose sur le NHEJ (Figure 44). En effet, il a été montré que le clivage par Pgm ne peut pas se faire sans Ku70/80, en particulier Ku70a et Ku80c, qui sont surexprimés lors des réarrangements du génome

(Abello *et al.*, 2020; Marmignon *et al.*, 2014). Cela a amené à la conclusion que la réparation des CDB induites par Pgm est couplée à leur introduction : dès que le dommage est introduit, sa réparation commence. D'autres partenaires connus du NHEJ ont été identifiés parmi les protéines pouvant faire partie du système de réparation de ces CDB : Lig4p/XRCC4p (Kapusta *et al.*, 2011), DNA-PKcs (Betermier and Duharcourt, 2014). Dès lors, il semble logique de penser que le système de réparation de ces CDB programmées repose sur le NHEJ.

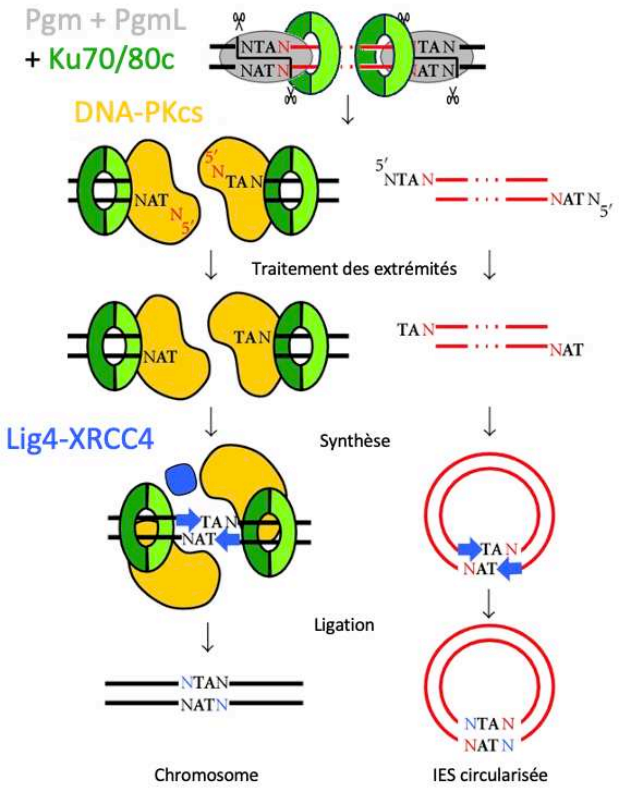


Figure 44 : Modèle d'introduction des CDB par Pgm et de leur réparation impliquant des protéines du NHEJ. PiggyMac (Pgm) et ses homologues PgmL introduisent des CDB aux extrémités des IES, avec le dimère Ku70/80c. Ce dimère est ensuite chargé sur les extrémités de l'ADN endommagé, et recrute d'autres partenaires du NHEJ classique, comme DNA-PKcs et Lig4p-XRCC4p, ce qui permet la réparation des CDB introduites. D'après Dubois E et al. Int J Evol Biol. 2012

Ce système de réparation a cependant une particularité de taille en comparaison avec le NHEJ classique : il ne fait pas d'erreurs.

7 Problématique de la thèse

Le système NHEJ impliqué dans la réparation des CDB programmées chez *P. tetraurelia* ne fait pas d'erreurs. En effet, il est estimé que pendant ces réarrangements du génome, il y a environ 10^6 CDB à réparer au sein de chaque MAC, et cette réparation semble extrêmement efficace et fidèle (Bétermier *et al.*, 2000). Le couplage entre l'introduction des CDB et leur réparation est évidemment un avantage, puisque l'ADN lésé n'a pas à attendre pour que le système de réparation se mette en marche. Cependant, ce système de réparation est particulièrement efficace, puisque la ligation finale se fait au nucléotide près, sans introduire de mutations, de façon reproductible, rapide, et ce sur plusieurs millions d'événements par cellule (Abello, 2019). C'est donc un très bon exemple de NHEJ ne faisant pas d'erreurs (Bétermier *et al.*, 2014).

Cependant, même si le système de réparation est recruté avant l'introduction des cassures, et que la ligation est très efficace, il reste une étape très importante : la complétion des brins d'ADN lésés. En effet, lorsque Pgm introduit une CDB, le nucléotide situé en 5' du site TA est retiré, ce qui laisse un espace à remplir. Chez les eucaryotes, ce type de dommage est pris en charge par les ADN polymérases de la famille X (λ ou μ).

Récemment, des travaux (non encore publiés) ont montré l'existence de 4 gènes chez *P. tetraurelia* codant pour des protéines montrant une homologie de séquence avec l'ADN polymérase λ (38% d'identité). Ces 4 protéines, appelées ADN polymérases X a, b, c et d sont très semblables les unes aux autres, et probablement issues de deux duplications du génome de *P. tetraurelia* (Figure 45).

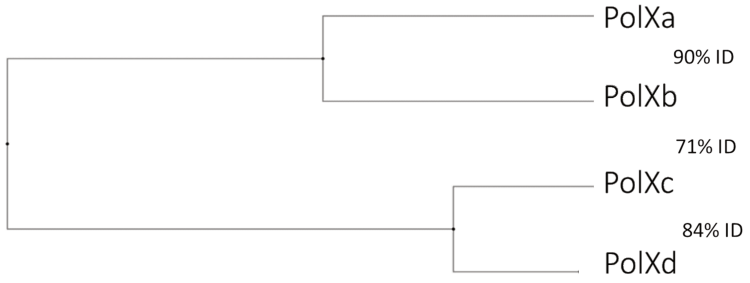


Figure 45 : Arbre phylogénétique liant les 4 PolX de P. tetraurelia. L'identité de séquence entre les PolXa et b, ainsi que PolXc et d, et entre les sous-groupes PolXab et PolXcd est indiquée.

En plus de leurs légères différences de séquence, ces 4 ADN polymérases X présentent aussi des niveaux d'expression différents au cours du cycle de vie de *P. tetraurelia* : l'ADN polymérase Xa est surexprimée dès la fragmentation du MAC parental et surtout pendant le

développement des MAC descendants ; l'ADN polymérase Xb est légèrement surexprimée de la méiose des MIC au développement des MAC, et les ADN polymérases Xc et d sont exprimées de façon constitutive (Figure 46).

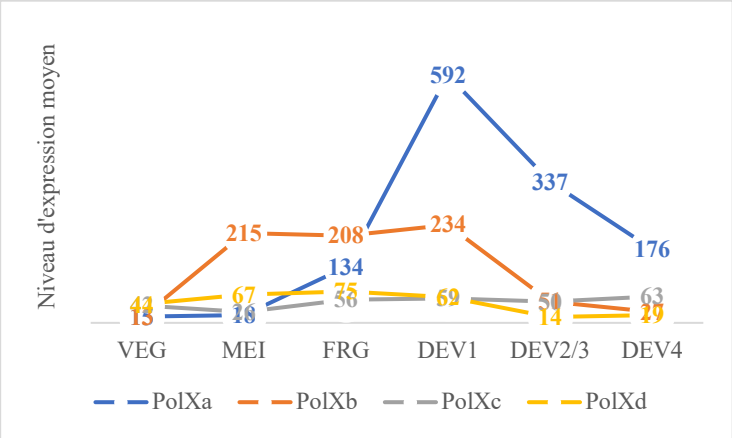


Figure 46 : Niveau d'expression moyen des 4 ADN polymérases X de P. tetraurelia au cours du cycle de vie de la cellule, mesuré par RNA-seq. VEG : cycle végétatif ; MEI : méiose des MIC, FRG : fragmentation du MAC, DEV1/2/3/4 : développement du MAC. D'après Arnaiz O et al. BMC Genomics. 2017.

L'hypothèse de départ de mes travaux de thèse est la suivante : les ADN polymérases X de *Paramecium* sont centrales dans la fidélité du système de réparation des cassures double brins introduites par PiggyMac. Par conséquent, au cours de mes travaux, j'ai cherché à répondre à la question suivante : **quelle est la fidélité des ADN polymérases X impliquées dans la réparation des cassures double brins programmées chez *Paramecium tetraurelia* et comment expliquer cette fidélité en termes moléculaires ?**

Les chapitres suivants seront consacrés à l'étude de cette hypothèse : je commencerai par situer les ADN polymérases X de Paramécie dans cette famille d'enzymes (Chapitre 1), en actualisant la classification des ADN polymérases de la famille X, ce qui permettra d'émettre deux hypothèses sur la grande fidélité de ces enzymes. Ensuite, après avoir détaillé l'expression, la production et la purification de certaines constructions de ces polymérases, je présenterai une étude de leur activité (Chapitre 2). Après avoir montré la grande fidélité de ces ADN polymérases, j'étudierai les deux mécanismes pouvant en être à l'origine.

Un chapitre additionnel, basé sur un article publié pendant ma thèse, portera sur un projet secondaire : la reclassification des ADN polymérases de la famille A, ainsi que la caractérisation fonctionnelle de deux ADN polymérases issues de nouveaux sous-groupes de cette famille.

Chapitre 1

La classification des ADN polymérases de la famille X

1 Introduction

Les cellules eucaryotes contiennent au moins 15 ADN polymérases différentes (Weill and Reynaud, 2008). Pour faciliter leur étude, celles-ci sont classées selon les similitudes entre leurs séquences et leurs structures, parmi plusieurs familles (Bebenek and Kunkel, 2004; Filée *et al.*, 2002; Raia *et al.*, 2019) :

- A : ADN polymérases de réplication et de réparation (ex : Pol γ eucaryote, T7 ADN polymérase, Pol I d'*Escherichia coli*) portant des activités 5'-3' et 3'-5' exonucléases, avec un repliement de type « *Klenow-fold* » (Czernecki *et al.*, 2023) ;
- B : ADN polymérases réplicatives et de réparation (ex : Pol α, δ et ε, PolB des archées) avec une activité 3'-5' exonucléase et une structure de type « *Klenow-fold* » ;
- C : ADN polymérases réplicatives bactériennes (ex : Pol III) avec une activité 5'-3' exonucléase et une structure de type « *Pol β-like* » ;
- D : ADN polymérases d'archées (ex : PolD d'archées comme *Pyrococcus abyssii* (Madru *et al.*, 2020)) avec une activité 3'-5' exonucléase, et une structure de type « *two-beta-barrel* » ;
- Y : ADN polymérases de translésion (tolérant les lésions de l'ADN lors de la synthèse), avec une faible fidélité (ex : Pol η eucaryote) (Yang, 2014), avec une structure de type « *Klenow-fold* » ;
- Primpol (Guilliam *et al.*, 2015) : ADN polymérases de réplication, impliquées dans l'initiation de la réplication (y compris dans des génomes viraux) (X. Chen *et al.*, 2020), mais aussi dans certains mécanismes de réparation, avec une structure de type « *Archaeo-Eukaryotic-Primase* » (AEP) ;
- Transcriptases inverses : ADN polymérases de rétrovirus utilisant de l'ARN comme matrice pour synthétiser de l'ADN ;
- X : le sujet principal de ces travaux.

Les ADN polymérases de la famille X ont un repliement de type « *Pol β-like* », et sont spécialisées dans la réparation de certains dommages de l'ADN. En comparaison avec d'autres ADN polymérases, elles sont généralement considérées comme peu fidèles (Yamtich and Sweasy, 2010). Elles sont actives en présence d'ions Mg^{2+} mais tolèrent aussi le Mn^{2+} (qui

modifie leur activité), comme d'autres ADN polymérases impliquées dans des systèmes de réparation (Vashishtha *et al.*, 2016; Wang and Konigsberg, 2022). Leur rôle est de remplir des vides (*gaps*) allant de 1 à plusieurs nucléotides lors de la réparation de certains dommages de l'ADN (Yamtich and Sweasy, 2010).

Jusqu'à aujourd'hui, les 4 ADN polymérases X humaines (β, λ, μ et *Terminal deoxynucleotidyl transferase* ou Tdt) ont été les plus étudiées. En effet, leur implication dans la réparation des dommages de l'ADN en fait des cibles de choix pour la compréhension et le traitement de certaines maladies humaines (Wallace *et al.*, 2012; Zhao *et al.*, 2020). Des homologues de ces ADN polymérases X sont retrouvés chez tous les eucaryotes (Uchiyama *et al.*, 2009) - même si dans certains cas comme chez *Saccharomyces cerevisiae*, seul un représentant est présent - : plantes, champignons et animaux ; et même chez certaines bactéries (Lecointe *et al.*, 2004).

De façon générale, les ADN polymérases X partagent des caractéristiques structurales communes (Ghosh and Raghavan, 2021) : ce sont de petites protéines (d'une masse de 30 à 70 kDa), qui portent un domaine de 8 kDa et un domaine polymérase composé de sous-domaines « pouce », « paume » et « doigts ». Le domaine de 8-kDa et les doigts du domaine polymérase portent chacun un motif « hélice-épingle à cheveux-hélice », qui servent à interagir avec l'ADN, respectivement en aval des dommages et avec le brin amorce. Le site actif du domaine polymérase est formé d'un motif DxD (x étant un acide aminé variable) et d'un autre aspartate dans un motif RxDx(Φ/+[1]), tous situés dans le domaine « paume ».

Dans ce premier chapitre, je détaille mes travaux réalisés par des méthodes de bio-informatique. La première et plus grande partie concerne la classification des ADN polymérases X sur la base d'analyses de séquences, que j'ai mise à jour en utilisant une approche plus large que celle utilisée jusqu'ici. L'objectif de la seconde partie de ce chapitre a été d'étudier les ADN polymérases X de *Paramecium tetraurelia* en essayant de déterminer leur place dans la

[1] Φ représente un acide aminé hydrophobe (fréquemment une leucine ou isoleucine) ; + représente un acide aminé chargé positivement (arginine ou lysine)

classification des PolX, afin de dégager des hypothèses pouvant expliquer leur grande fidélité au sein du système de réparation des CDB programmées chez *P. tetraurelia*.

1.1 Les ADN polymérases X des métazoaires (animaux multicellulaires)

1.1.1 L'ADN polymérase β

L'ADN polymérase β a été découverte en 1974 (Matsukage *et al.*, 1974), et a longuement été étudiée depuis, entre autres par Samuel Wilson et son groupe (Beard and Wilson, 2014; Hanawalt, 2020; Van Houten, 2020; Whitaker and Freudenthal, 2020). Elle est impliquée dans le système de réparation BER (Krokan and Bjørås, 2013). Au sein de ce système de réparation, l'ADN polymérase β a pour rôle de synthétiser la base manquante après la suppression de la base endommagée par une ADN glycosylase et le retrait du site abasique. Ce système nécessite une grande fidélité, ce qui fait de l'ADN polymérase β la plus fidèle des ADN polymérases X de métazoaires. Cette fidélité s'explique de plusieurs façons.

Premièrement, comme l'ADN polymérase λ et à l'inverse de l'ADN polymérase μ et surtout de la Tdt, l'ADN polymérase β est une ADN polymérase ADN dépendante. Cela signifie qu'elle prend ses instructions d'un brin d'ADN matrice (ou instructeur ou *template*), et qu'elle synthétise de l'ADN en ajoutant des dNTPs sur un brin amorce. Elle ne peut pas ajouter de NTPs sur un brin d'ADN (on parle de discrimination contre les NTPs). Cette discrimination repose sur la présence dans son site catalytique d'un *steric gate* ou filtre stérique (composé des résidus YFTGS). Ces résidus, via leurs chaînes latérales et leur squelette peptidique, entrent en collision avec le groupement 2'OH des ribonucléotides entrant dans le site actif, ce qui empêche le bon placement de la base du nucléotide face à la base du brin d'ADN matrice : ainsi, le *steric gate* interdit l'incorporation de NTPs, malgré leur plus grande concentration dans les cellules (Brown *et al.*, 2010).

De plus, il a été montré que la grande fidélité de l'ADN polymérase β humaine repose sur un mécanisme unique chez les ADN polymérases X appelé *induced fit* (Beard *et al.*, 2014). La polymérase doit passer par des points de contrôle avant de former un complexe enzyme-substrat actif. Ces points de contrôle sont les suivants : l'ADN doit se placer correctement dans le site actif, et le bon dNTP doit se placer dans le site actif en respectant la géométrie lui

permettant de s'hybrider au nucléotide instructeur (*template*). Si tout est correctement placé, l'ADN polymérase passe de son état ouvert (inactif) à un état fermé (actif) (Moscato *et al.*, 2016). Si ces conditions ne sont pas respectées, elle ne peut pas se fermer, donc la catalyse n'a pas lieu et le dNTP ressort. Ce mécanisme d'alternance « ouvert-fermé » est unique parmi les ADN polymérases X (Beard *et al.*, 2014; Garcia-Diaz *et al.*, 2005b; Moon *et al.*, 2014), mais assez commun chez les ADN polymérases en général (Johnson, 2008; Tsai and Johnson, 2006). Il est détaillé plus loin, à partir de la page 123.

L'ADN polymérase β a aussi une activité désoxyribose phosphate (dRP) lyase, portée par son domaine de 8 kDa. Dans certains cas, en effet, lors de l'excision de la base endommagée et du site abasique par les enzymes du système BER, l'extrémité 5' de l'ADN en aval de la cassure peut porter un groupement dRP. Or, pour que la polymérase puisse ajouter la base manquante, elle a besoin de reconnaitre un groupement phosphate grâce à une poche de son domaine 8-kDa. Elle porte donc dans cette poche une activité dRP lyase, qui lui permet d'éliminer le désoxyribose, afin que l'ADN soit dans les bonnes conditions pour être reconnu et réparé (Figure 47) (Belousova and Lavrik, 2015; Prasad *et al.*, 1998).

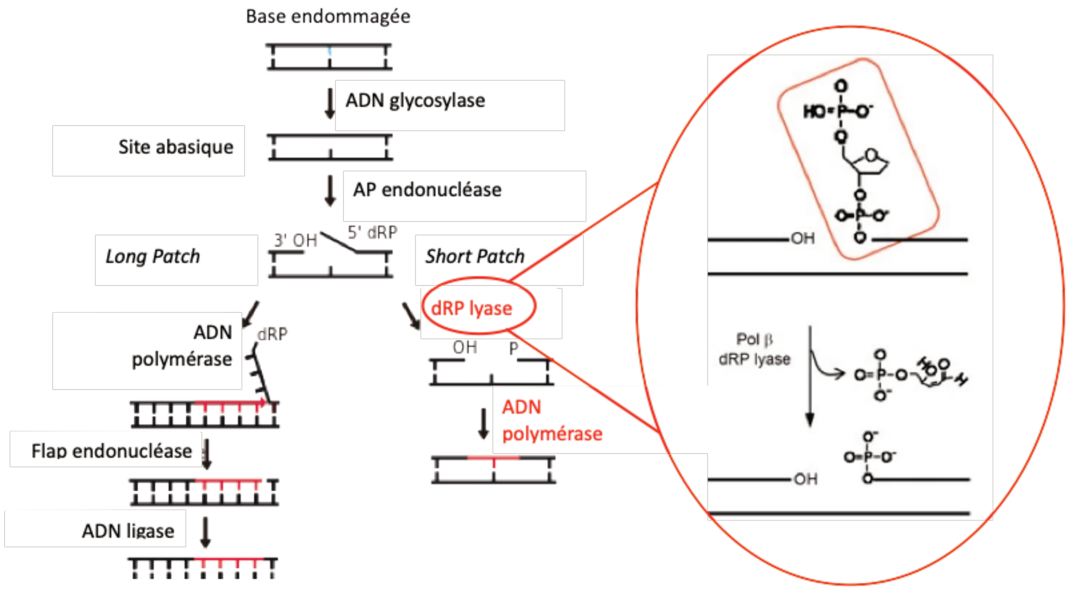


Figure 47 : L'activité dRP lyase s'intègre au sein du système de réparation par excision de base.
Dans certains cas, lors de l'excision de la base endommagée et du site abasique par les enzymes du système BER (ADN glycosylase et AP endonucléase), l'extrémité 5' de l'ADN en aval de la cassure peut porter un groupement dRP. Or, pour que l'ADN polymérase puisse ajouter la base manquante, elle doit reconnaitre un groupement 5' Phosphate. Elle porte donc une activité dRP lyase, qui lui permet d'éliminer le désoxyribose.

Enfin, une séquence de 3 acides aminés appelée SD2 située dans son domaine « pouce » (*Sequence Determinant region 2*) permet de différencier les quatre ADN polymérases X (Romain *et al.*, 2009) humaines : dans le cas de l'ADN polymérase β, le SD2 est un motif NEY (le glutamate et la tyrosine sont d'ailleurs impliqués dans le mécanisme *induced-fit*).

1.1.2 Les ADN polymérases X impliquées dans le NHEJ

Les ADN polymérases λ, μ et la Tdt sont quant à elles impliquées dans le NHEJ (Capp *et al.*, 2006; Lee *et al.*, 2004), où leur rôle est d'ajouter des nucléotides à l'ADN endommagé avant la ligation des deux brins lésés par le complexe Ligase IV / XRCC4.

L'implication de ces polymérases dans le NHEJ dépend de leur domaine N-terminal BRCT (*BReast cancer Carboxy Terminal associated*), qui leur permet de former des complexes avec Ku70/80 (DeRose *et al.*, 2007; Ramsden and Asagoshi, 2012).

Ces trois ADN polymérases, bien que proches, ont cependant des caractéristiques structurales et fonctionnelles différentes.

1.1.2.1 L'ADN polymérase λ

L'ADN polymérase λ est la plus similaire à l'ADN polymérase β (34% d'identité de séquence). Elles partagent en effet de nombreuses caractéristiques, comme l'activité dRP lyase (García-Díaz *et al.*, 2001), la reconnaissance des groupements 5' phosphates par le domaine 8-kDa, et le *steric gate* (qui fait de l'ADN polymérase λ une ADN polymérase ADN dépendante). Elles ont également un domaine polymérase très similaire d'un point de vue structural. Cependant, elles se différencient par plusieurs points, dont la présence d'un domaine BRCT et l'absence d'*induced-fit mechanism* chez l'ADN polymérase λ: cette dernière reste en conformation fermée tout au long du cycle catalytique (Garcia-Diaz *et al.*, 2005b). Le motif SD2 de l'ADN polymérase λ est formé par les résidus SEH.

De plus, l'ADN polymérase λ porte à son extrémité C-terminale une boucle appelée boucle 3, spécifique des ADN polymérases λ, qui pourrait être liée à un mécanisme de fidélité (Jamsen *et al.*, 2022). Il a été observé que cette boucle réalise un mouvement lors de la catalyse : en absence d'un nucléotide correct, elle n'est pas au contact de l'ADN et peut même être assez

flexible ; en présence d'un nucléotide correct, elle est au contact de l'ADN instructeur et stabilise sa position au sein du site actif.

L'ADN polymérase λ est active au sein du NHEJ grâce à son domaine BRCT, mais aussi au sein du BER, comme l'ADN polymérase β (Lee *et al.*, 2004), et même dans la réparation par synthèse translésionnelle (TLS) (Blanca *et al.*, 2004; Yoon *et al.*, 2021). Elle intervient dans le NHEJ lorsqu'il y a une microhomologie de 2 à 4 paires de bases entre les deux brins d'ADN séparés par la cassure (Pryor *et al.*, 2015; Ramsden *et al.*, 2022). Contrairement à la Tdt (et à l'ADN polymérase μ), elle a toujours besoin d'un brin instructeur pour synthétiser de l'ADN (Nick McElhinny *et al.*, 2005).

1.1.2.2 L'ADN polymérase μ

L'ADN polymérase μ a été découverte en même temps que l'ADN polymérase λ en 2000 (Aoufouchi *et al.*, 2000), d'après des analyses de séquences qui ont révélé l'existence de deux nouvelles ADN polymérases de la famille X, l'une proche de l'ADN polymérase β (l'ADN polymérase λ), et l'autre proche de la Tdt (l'ADN polymérase μ). L'ADN polymérase μ est construite globalement comme l'ADN polymérase λ, avec le domaine polymérase classique des ADN polymérases X et un domaine BRCT N-terminal. Cependant, l'ADN polymérase μ est assez différente de l'ADN polymérase λ: son domaine 8-kDa ne porte pas d'activité dRP lyase, mais lui permet d'interagir avec des substrats ADN portant un groupement 5' phosphate ; et elle n'a pas de *steric gate* (la séquence équivalente est GWTGS), ce qui lui permet d'incorporer des ribonucléotides (Ghosh and Raghavan, 2021) : elle est donc considérée comme une ADN/ARN polymérase ADN dépendante. Cette capacité est un avantage, étant donnée la concentration cellulaire en NTPs, qui est bien supérieure à celle des dNTPs (Ferraro *et al.*, 2010; Traut, 1994) ; et il a été observé que l'incorporation de NTPs par l'ADN polymérase μ facilite la ligation par le complexe XRCC4-Ligase 4 (Nick McElhinny and Ramsden, 2003; Pryor *et al.*, 2018). De plus, l'ADN polymérase μ porte comme la Tdt une boucle de 17 à 21 résidus appelée boucle 1, variable en longueur et en séquence, qui lui confère une activité terminal transférase (la possibilité d'ajouter des nucléotides sans brin instructeur) en présence d'ions Mn^{2+} (Domínguez *et al.*, 2000; Juárez *et al.*, 2006) , et qui lui permet aussi de prendre ses instructions depuis un brin instructeur situé en aval de la coupure (sans micro-homologie). Il a aussi été montré que

cette boucle 1 sert à la polymérase à vérifier que le nucléotide présent dans le site actif est correct, en stabilisant l'ADN en aval de la cassure pour permettre au nucléotide d'interagir avec lui, après quoi l'ADN en amont de la cassure entre lui aussi dans le site actif, la boucle 1 est déplacée, et la catalyse peut avoir lieu (Loc'h *et al.*, 2019). De plus, une courte séquence située juste après cette boucle 1, appelée SD1 (*Substrate Specificity Sequence Determinant* 1), sert à maintenir la micro-homologie en place ; et permet de différencier l'ADN polymérase μ (DHFQKCF) de la Tdt (DAFERSF) (Gouge *et al.*, 2015). Enfin, l'ADN polymérase μ a pour motif SD2 la séquence NSH.

Elle est spécialisée au sein du NHEJ avec les substrats présentant des petites micro-homologies (0-1 pb) (Ramsden *et al.*, 2022).

1.1.2.3 La Terminal déoxynucléotidyltransférase (Tdt)

Dès le début des années 1960, la Tdt a été l'une des premières ADN polymérases humaines à être étudiées (Bollum, 1960). La Tdt est très similaire à l'ADN polymérase μ (44% d'identité de séquence) : elle porte elle aussi un domaine BRCT, une boucle 1 (à l'origine de sa spécificité de substrat (Romain *et al.*, 2009)), un motif SD1 de séquence DAFERSF, n'a pas d'activité dRP lyase et pas de *steric gate*, mais n'a pas non plus de capacité à fixer les groupements 5' phosphate des substrats ADN. La séquence de son motif SD2 est DNH.

Cependant, la Tdt présente une activité particulière, qui est liée à sa spécialisation dans la recombinaison V(D)J : elle présente naturellement une activité nucléotidyltransférase (Kato *et al.*, 1967). Elle n'a donc pas besoin d'un brin d'ADN matrice pour synthétiser de l'ADN, et peut donc en synthétiser de façon aléatoire. Cela lui permet de réaliser des « N-additions » de 2 à 20 nucléotides aux jonctions V-D et D-J des gènes des chaines lourdes des immunoglobulines et des récepteurs des lymphocytes T, ce qui contribue à la diversité du répertoire immunitaire.

Il a été montré que la Tdt peut assembler une synapse avec une micro-homologie de 1 pb après une CDB (Gouge *et al.*, 2015) et avoir une activité matrice-dépendante en *trans* (Loc'h *et al.*, 2016), c'est-à-dire qu'elle utilise l'information du brin template en aval de la cassure pour ajouter des nucléotides sur le brin amorce en amont, comme les ADN polymérase μ ou λ.

1.2 Les ADN polymérases X des *Fungi* (champignons et levures)

Chez les *Fungi*, les eucaryotes les plus proches des métazoaires, trois sous-groupes de polymérases de la famille X sont identifiés à ce jour :

- l'ADN polymérase λ (chez les champignons)
- l'ADN polymérase μ (chez les champignons)
- l'ADN polymérase IV (chez les levures)

1.2.1 Les ADN polymérases λ et μ de champignons

En 2007, Sakamoto *et al*, lors de travaux sur la méiose chez le coprin cendré (*Coprinus cinereus*), ont identifié de nombreuses protéines impliquées dans la réplication du génome lors de la méiose, dans la réparation de l'ADN et dans la recombinaison. Parmi celles-ci, en cherchant des homologues des ADN polymérases X humaines, ils ont identifié des séquences proches de l'ADN polymérase λ (28% d'identité) et de l'ADN polymérase μ (24% d'identité), qu'ils ont ensuite produites et étudiées (Sakamoto *et al.*, 2007).

Ces deux séquences portent chacune un domaine BRCT et un domaine polymérase proche des ADN polymérases X connues. Ils ont pu montrer que ces deux protéines avaient bien une activité ADN polymérase en présence de Mn^{2+}, mais pas d'activité terminal transférase. Ils ont de plus montré que ces deux polymérases pourraient avoir une activité d'extension d'amorce lors de la formation de boucles D lors de la méiose, ce qui les rapproche d'un autre système de réparation : la recombinaison homologue.

1.2.2 Les ADN polymérases IV de levures

Chez la levure *Saccharomyces cerevisiae*, seule une ADN polymérase de la famille X a été identifiée, en 1993 (Prasad *et al.*, 1993) : l'ADN polymérase IV (Bebenek *et al.*, 2005). Elle porte plusieurs activités : dRP lyase, *gap-filling*, ADN/ARN polymérase, a un taux d'erreurs élevé et n'a pas besoin de reconnaitre un 5' phosphate en aval des cassures de l'ADN. Ces différentes caractéristiques semblent empruntées aux 4 ADN polymérases X des métazoaires. Cette ADN polymérase porte un domaine BRCT N-terminal et interagit avec des partenaires du NHEJ *in vivo*, ce qui lui permet de remplir des trous entre des brins d'ADN présentant une micro-

homologie (Wilson and Lieber, 1999). Elle semble de plus être active au sein de la voie de réparation BER (McInnis *et al.*, 2002).

1.3 Les ADN polymérases X des *Viridiplantae*

Chez les plantes vertes, seul un gène codant pour une ADN polymérase X a été identifié, en 2004 (Uchiyama *et al.*, 2004). Dans ces travaux portant sur *Oryza sativa* (le riz), une séquence codant pour ce qui a été catégorisé comme une ADN polymérase λ a été identifiée. Cette séquence porte un domaine BRCT et un domaine polymérase proche de ceux connus chez les ADN polymérases X. Un gène équivalent a été identifié chez *Arabidopsis thaliana*, un organisme modèle classique en biologie végétale. Ces deux séquences partagent environ 60% d'identité, et seulement 29% avec l'ADN polymérase λ humaine. Après avoir produit et purifié ces deux enzymes, ils ont montré qu'elles ont une activité ADN polymérase plus efficace avec des ions Mn^{2+} que Mg^{2+}, ce qui est caractéristique des ADN polymérases impliquées dans la réparation (Wang and Konigsberg, 2022), et qu'elles ont aussi une faible activité terminal transférase, et une activité dRP lyase.

Afin de confirmer ces résultats et de les élargir aux plantes en général, cette même séquence a été recherchée et trouvée chez l'organisme modèle *Chlamydomonas reinhardtii* (Uchiyama *et al.*, 2009): il semble donc que les plantes en général n'aient qu'une ADN polymérase X, considérée comme proche d'une ADN polymérase λ.

1.4 Les ADN polymérases X bactériennes

1.4.1 Les ADN polymérase X bactériennes canoniques

En 2004, des travaux portant sur la bactérie *Deinococcus radiodurans*, une bactérie particulièrement résistante aux agents endommageant l'ADN, ont montré qu'elle portait dans son génome une séquence codant pour une protéine proche des ADN polymérase X connues (Lecointe *et al.*, 2004) : 24 % d'identité avec l'ADN polymérase λ humaine, 25% avec l' ADN polymérase β humaine, et 26% avec l'ADN polymérase IV de *S. cerevisiae*. Des protéines proches ont ensuite été trouvées chez d'autres bactéries, comme *Bacillus subtilis* ou *Thermus thermophilus*, et même chez certaines archées (Aravind and Koonin, 1999, 1998). Ces travaux ont montré que ces protéines ne possèdent pas de domaine BRCT, mais qu'elles présentent un

steric gate équivalent à ceux des ADN polymérases λ ou β, et qu'elles portent en C-terminal un domaine PHP (Polymérase / Histidinol Phosphatase).

Ils ont montré que cette protéine avait bien une activité ADN polymérase, amplifiée en présence d'ions Mn^{2+} (des ions présents en grandes quantités dans le cytoplasme de *D. radiodurans* (Daly *et al.*, 2004)), et indispensable à la réparation des CDB. Ces ADN polymérases portent une activité 3'-5' exonucléase modulée par la structure de l'ADN, elle aussi importante pour la réparation des CDB (Blasius *et al.*, 2006), ainsi qu'une activité AP-endonucléase (pour supprimer les sites abasiques), une activité 3'-phosphodiestérase et 3'-phosphatase. Ces activités sont portées par le domaine PHP (Baños *et al.*, 2010, 2008; Nakane *et al.*, 2012b, 2009), en particulier par un ensemble de résidus basiques (Rodríguez *et al.*, 2019).

1.4.2 Les ADN polymérase X bactériennes non canoniques

Plus récemment, en 2022, des travaux de bio-informatique plus exhaustifs sur les ADN polymérase X procaryotes ont montré que de nombreuses bactéries portent ces ADN polymérase X, mais qu'une part non négligeable (27,5%) de ces polymérases semblent « non-canoniques », car elles ont perdu plusieurs caractéristiques du site actif d'ADN polymérase (absence d'un ou plusieurs résidus catalytiques, pas de *steric gate*, mutations dans la poche de fixation des dNTPs), et qu'elles n'ont plus d'activité ADN polymérase (Prostova *et al.*, 2022). Leur domaine « paume » est fréquemment tronqué. Ces protéines semblent en revanche avoir conservé un domaine PHP intègre et pleinement fonctionnel (Prostova *et al.*, 2022).

De façon surprenante, ces ADN polymérase X non canoniques sont retrouvées principalement chez des bactéries des groupes *Thermus* et *Deinococcus*, ce qui suggère une origine commune, mais aussi plus rarement chez d'autres groupes, suggérant des transferts horizontaux (ce qui est aussi le cas pour certaines ADN polymérase X bactériennes canoniques (Bienstock *et al.*, 2014)). Il existe cependant d'autres groupes de bactéries et même des archées qui présentent ces ADN polymérase X non canoniques.

2 Révision de la classification des ADN polymérases de la famille X

Jusqu'ici, peu d'analyses phylogénétiques des ADN polymérases de la famille X, incluant les ADN polymérase X ne provenant pas des métazoaires ont été réalisées. On peut cependant

citer une étude de l'équipe de Samuel Wilson (Bienstock *et al.*, 2014), qui a réalisé une analyse phylogénétique incluant des ADN polymérase X bactériennes et provenant des Fungi. Ils ont notamment proposé une hypothèse sur l'origine et l'évolution des ADN polymérase X, allant des ADN polymérase X bactériennes les plus ancestrales (chez *Bacillus subtilis*) aux ADN polymérase X de métazoaires. Cependant, le plus souvent, lorsque de nouvelles séquences ont été étudiées, elles n'ont pas été intégrées dans le paysage global de cette famille, mais comparées à quelques ADN polymérase X connues par des alignements de séquences. Par exemple, lors de la découverte des ADN polymérase X de plantes, elles ont été rapidement catégorisées comme des ADN polymérases λ, or comme nous le verrons, elles présentent des différences majeures avec les ADN polymérases λ de métazoaires. Pour d'autres familles de polymérases, l'étude phylogénétique s'est jusqu'ici limitée à l'utilisation d'arbres phylogénétiques ; or ces méthodes utilisent une « horloge moléculaire » (Van Der Wal and Ho, 2019) fixée *a priori*, qui suppose que l'évolution de telles séquences ne se fait que verticalement (d'un organisme parent vers sa descendance). Or, et c'est par exemple le cas pour les ADN polymérases X bactériennes, on sait que des transferts horizontaux ont eu lieu au cours de l'évolution. De tels transferts (et de nombreux autres facteurs (Drake, 1999)) sont évidemment un problème pour ces méthodes d'analyse.

Durant mon doctorat, j'ai eu l'occasion de découvrir un outil de classification de séquences de protéines appelé CLANS (Frickey and Lupas, 2004), basé sur l'algorithme de Fruchterman-Reingold (Fruchterman and Reingold, 1991), et permettant d'étudier des séquences par *clustering*. Le but de cette méthode d'analyse n'est pas de donner les relations phylogénétiques entre les séquences à la façon d'un arbre, mais plutôt de regrouper (dans un espace tridimensionnel) les séquences proches d'après leur similarité, sans *a priori*. Ce programme fonctionne en deux grandes étapes : un calcul et une visualisation. Il commence par calculer les similarités entre chaque paire de séquence donnée (par exemple : avec 7000 séquences en entrée, il calcule la similarité de chaque séquence avec les 6999 autres) en utilisant le programme BLAST (Altschul *et al.*, 1990), et à partir des scores de similarité, il dérive des pseudo-forces. Ces forces sont à la base du fonctionnement de l'algorithme de Fruchterman-Reingold pour révéler des clusters éventuels : le programme projette les séquences (sous forme de points) dans un espace tridimensionnel de façon aléatoire, puis utilise les pseudo-forces calculées précédemment pour attirer les séquences proches les unes des autres, jusqu'à atteindre

une convergence lorsque les séquences proches sont regroupées de façon stable en clusters individuels. CLANS permet donc, à partir d'un ensemble de séquences, de rassembler les séquences proches, et cela indépendamment de toute hypothèse phylogénétique. Cette méthode a été utilisée récemment pour étudier les polymérases de la famille B et de la superfamille AEP, ce qui a permis de découvrir de nouvelles sous-familles (Kazlauskas *et al.*, 2020, 2018). Elle a également été utilisée par Dariusz Czernecki et moi-même au sein du laboratoire pour étudier les ADN polymérases de la famille A au début de mon doctorat et révéler l'existence de sous familles nouvelles et jusqu'ici non étudiées (Czernecki *et al.*, 2023).

2.1 Matériel et méthodes

2.1.1 Obtention et clustering des séquences d'ADN polymérases X

A partir de la séquence de la première ADN polymérase X de *P. tetraurelia* étudiée (appelée PolXa / séquence ParameciumDB ID : PTET.51.1.P0210235), une recherche de séquences par PSI-BLAST (Altschul *et al.*, 1997) a été lancée sur le serveur du NCBI (Wheeler *et al.*, 2003) le 25 Janvier 2023. Les séquences ont été triées selon leur *query coverage*, c'est-à-dire la proportion de chaque séquence qui couvre la séquence donnée en entrée, et les 6500 premières séquences ont été extraites (pour la dernière séquence sélectionnée : *query coverage* = 52% ; e-value = 2.10^{-37} ; % identité = 32,02%). Cette base de données incluait des ADN polymérase X de métazoaires, de plantes, de champignons, mais pas d'ADN polymérases IV de levures, ni d'ADN polymérase X bactériennes. Afin d'avoir une base de données plus exhaustive permettant d'étudier les similitudes entre des représentants de toutes les ADN polymérase X connues à ce jour, 250 séquences de PolIV, 250 séquences d'ADN polymérase X bactériennes canoniques, et 250 non canoniques ont donc été recherchées (par PSI-BLAST, dans les mêmes conditions, avec en entrée l'ADN polymérase X canonique de *Thermus thermophilus* [NCBI ID : WP_096410530.1], l'ADN polymérase X non canonique de *Deinococcus radiodurans* [NCBI ID : WP_010887112.1] et l'ADN polymérase IV de *Saccharomyces cerevisiae* [NCBI ID : AJP37443.1]) et ajoutées. Le fichier fasta contenant les 7250 séquences a été traité sur l'outil CLANS intégré dans le serveur MPI Bioinformatics Toolkit (Gabler *et al.*, 2020; Zimmermann *et al.*, 2018), et le fichier obtenu a permis de lancer une simulation dans la version Java de CLANS. Les paramètres par défaut ont été utilisés, en particulier la non-utilisation d'un seuil de p-value, car d'autres simulations faites en utilisant des seuils de p-value à 10^{-10} et 10^{-20} se sont

avérées moins claires, et ne permettaient pas toujours de retrouver des indices de la classification connue des ADN polymérase X de métazoaires, or celle-ci devait servir de témoin pour valider les résultats obtenus. Le clustering a été lancé, et les clusters formés ont convergé après quelques centaines d'itérations. Le clustering a été poursuivi jusqu'à 20 000 itérations et répété, sans changements. Des simulations ont été réalisées sans les séquences bactériennes et de levures ajoutées, et la distribution des 6500 séquences restantes est restée équivalente. La recherche de clusters automatique par l'outil Java n'a pas permis de détecter les clusters pourtant bien présents visuellement, donc la sélection des clusters a dû être réalisée manuellement : chacun des nuages de points discernables à l'œil nu a été sélectionné indépendamment.

2.1.2 Analyses des séquences, phylogénie et analyses structurales

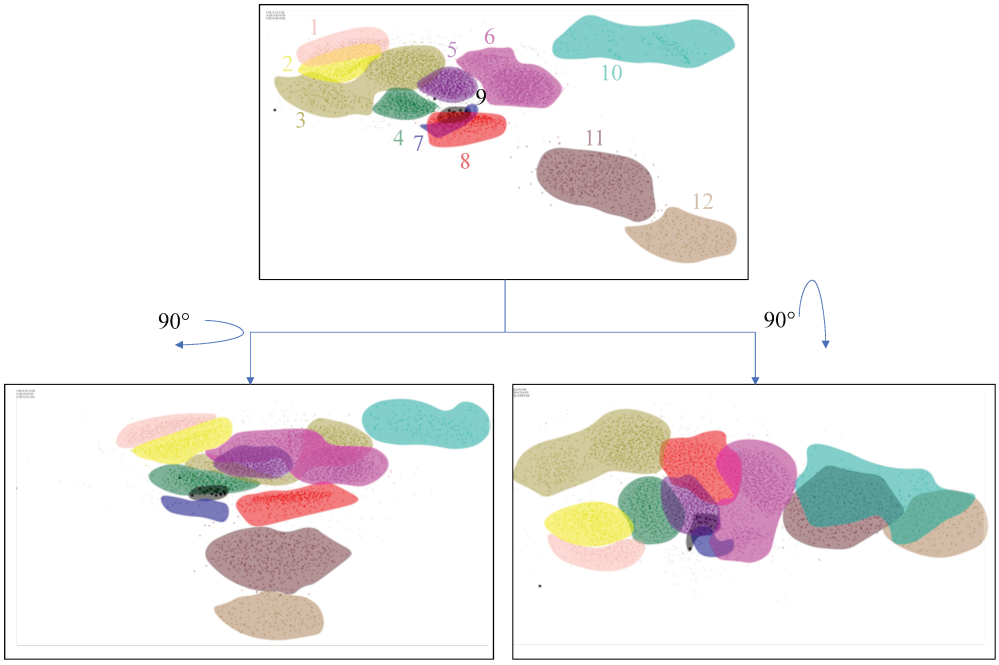
Pour chaque cluster, l'ensemble des séquences a été sélectionné et extrait. Tout d'abord, l'origine monophylétique de chaque cluster a été testée : des séquences représentatives de chaque cluster ont été obtenues avec l'outil HHFilter (Remmert *et al.*, 2011) du MPI Bioinformatics Toolkit (90% d'identité maximum au sein de chaque cluster) ; puis ces séquences ont été traitées avec l'outil IQ-Tree (Trifinopoulos *et al.*, 2016) (paramètres : Ultrafast Bootstrap, 1000 alignements, 1000 itérations) pour obtenir des arbres phylogénétiques, qui ont ensuite été visualisés grâce à iTOL (Letunic and Bork, 2021), ce qui a permis de vérifier, en se référant à la base de données du NCBI, que les séquences les moins similaires au sein de chaque cluster (donc les plus éloignées sur chaque arbre) provenaient bien des mêmes clades, supposant que l'ensemble des séquences appartiennent aussi à ce clade. Plusieurs séquences ont également été sélectionnées au hasard dans chaque groupe afin de vérifier leur organisme d'origine. Pour comparer les longueurs des séquences de chaque cluster, le nombre de résidus dans chaque séquence a été extrait, et pour chaque cluster la moyenne des tailles des séquences a été calculée, ainsi que l'écart type. Ces données ont ensuite été représentées sous forme d'histogrammes grâce au logiciel GraphPad Prism 9. Les séquences de chaque cluster ont ensuite été alignées avec MAFFT (Katoh *et al.*, 2019; Kuraku *et al.*, 2013), en utilisant les paramètres par défaut. Chaque fichier d'alignement a ensuite été traité avec WebLogo3 (Crooks *et al.*, 2004; Schneider and Stephens, 1990), avec les paramètres suivants : unité : kcal/mol ; pas de barres d'erreurs, schéma de couleurs : chimie. Cela a permis d'obtenir des motifs consensus caractéristiques des séquences de chaque cluster. Une séquence représentative de chaque cluster a ensuite été extraite, et l'ensemble des séquences représentatives choisies ont été alignées avec PSI-Coffee (Di

Tommaso *et al.*, 2011). Plusieurs versions de ces alignements ont été réalisées, avec toutes les séquences ou non : séquences entières, domaines N-terminaux, domaines C-terminaux. Enfin, pour chacune de ces séquences, des structures expérimentales connues ont été recherchées dans la Protein Data Bank (PDB) (Berman *et al.*, 2000). Lorsque de telles structures n'existaient pas, une version de l'outil de prédiction de structures AlphaFold2 (Jumper *et al.*, 2021) accessible depuis ChimeraX (Goddard *et al.*, 2018; Pettersen *et al.*, 2021) (version 1.6.dev202301310903, ColabFold 1.3.0) a été utilisé pour obtenir des modèles 3D des protéines concernées. La différence entre ColabFold et AlphaFold2 est l'algorithme utilisé pour obtenir des alignements de séquences : AlphaFold2 utilise HMMER et ColabFold utilise MMseqs2. Enfin, pour visualiser et comparer ces structures, Chimera X a été utilisé. Les comparaisons des structures ont été réalisées séparément sur les domaines BRCT, 8-kDa et polymérase. Pour chacun de ces domaines, et pour chaque ADN polymérase X concernée, les prédictions AlphaFold2 choisies ont été comparées à 10 autres modèles proposés par AlphaFold2, et à 15 modèles proposés par Rosetta2 (un autre outil de prédiction de structures utilisant le moteur HHblits pour aligner les séquences) - tous générés grâce à l'aide du Dr Pierre Legrand - en les superposant sur ChimeraX pour obtenir une mesure du RMSD (*Root-Mean-Square Deviation*). Cette mesure indique ici la différence entre les différentes structures, en donnant la distance moyenne entre chaque carbone α homologue entre les modèles prédits et le modèle choisi comme référence.

2.2 Résultats

2.2.1 Les ADN polymérase X se divisent en 12 groupes monophylétiques

À partir de la séquence d'une des ADN polymérase X de *Paramecium tetraurelia*, j'ai pu obtenir 6500 séquences d'ADN polymérase X par une recherche dans la base de données du NCBI via un PSI-BLAST. J'ai ajouté à ces séquences 250 séquences de PolIV de levures, 250 séquences d'ADN polymérase X bactériennes canoniques, et 250 séquences non canoniques. J'ai pu ensuite réaliser un clustering de ces 7250 séquences grâce à l'outil CLANS. Après 20000 itérations, des clusters de séquences se sont formés de façon stable, formant 12 clusters distincts.



N°	Nom	Phylum	Séquence représentative		Code **PDB** / AlphaFoldDB		
			Organisme	Identifiant NCBI	Domaine catalytique	Domaine BRCT	Domaine PHP
1	Tdt	Metazoa	*Mus musculus*	CAA48634.2	**4qz8**	**2coe**	NA
2	Polμ	Metazoa	*Homo sapiens*	NP_037416.1	**6ak8**	**2dun**	NA
3	Polμ like	Fungi	*Aspergillus sp. HF37*	RMJ22786.1			NA
4	PolX	Viridiplantae	*Arabidopsis thaliana*	NP_172522.2	**AF-Q9FNY4-F1**		NA
5	Polλ	Metazoa	*Homo sapiens*	AAH68529.1	**1rzt**	**2jw5**	NA
6	Polλ like	Fungi	*Aspergillus niger*	GKZ72300.1			NA
7	Polβ	Virus	*Mimivirus reunion*	QTF49230.1			NA
8	Polβ	Metazoa	*Homo sapiens*	AAA60133.1	**5tb8**	NA	
9	PolX	Sar	*Paramecium tetraurelia*	XP_001439407.1			NA
10	PolIV	Fungi	*Saccharomyces cerevisiae*	AJP37443.1	AF-P25615-F1		NA
11	PolX canoniques	Bacteria	*Thermus thermophilus*	WP_096410530.1	**3au2**	NA	**3au2**
12	PolX non canoniques	Bacteria	*Deinococcus radiodurans*	WP_010887112.1	**2w9m**	NA	**2w9m**

Figure 48 : Clustering des séquences d'ADN polymérase X. Un ensemble de 7250 séquences de PolXa été obtenu d'après la base de données du NCBI en janvier 2023. Une distribution 3D en clusters a été générée par CLANS en utilisant les score de similarités entre chaque paire de séquence comme valeurs d'attraction (en haut). Les 12 clusters obtenus sont présentés et colorés ici, et les noms qui leur ont été donnés sont indiqués dans le tableau en bas. Ce même tableau indique pour chaque cluster une séquence représentative choisie : les identifiants NCBI de ces protéines sont indiqués, ainsi que le code PDB / AlphaFoldDB(Varadi et al., 2022) des structures connues pour ces protéines.

Afin de comprendre cette subdivision, j'ai commencé par rechercher des origines monophylétiques au sein de chaque cluster. Pour cela, pour chaque cluster contenant initialement plus de 200 séquences, j'ai extrait des séquences représentatives (avec un seuil maximum de 90% d'identité) à partir desquelles j'ai pu obtenir des arbres phylogénétiques. J'ai ensuite déterminé l'origine des séquences les plus éloignées l'une de l'autre (et des séquences choisies au hasard) de chaque arbre, en recherchant leurs identifiants dans la base de données du NCBI. Pour tous les clusters, les séquences testées provenaient du même phylum (metazoa, fungi, viridiplantae, virus, bacteria, sar (qui contient les sous-groupes *Stramenopiles*, *Alveolata* et

Rhizaria)), et j'ai donc considéré que toutes les séquences de ces clusters provenaient également de ces mêmes phyla.

Tableau 2 : Nombre de séquences sélectionnées dans chaque cluster, et nombre de séquences initial de chaque cluster

Cluster	Nombre de séquences représentatives	Nombre de séquences initial
1	114	338
2	149	451
3	346	1045
4	144	512
5	176	925
6	830	1463
7	14	14
8	90	503
9	18	18
10	146	146
11	338	526
12	188	188

2.2.2 La distribution des séquences connues correspond à celle attendue

Le principal contrôle utilisé pour valider ces résultats était la comparaison avec les connaissances actuelles sur les ADN polymérases X connues et leurs relations.

Tout d'abord, j'ai donc recherché les ADN polymérases X connues, qui m'ont servi de références. J'ai pu associer chaque cluster avec une séquence de référence. J'ai donc pu associer la grande majorité des clusters à des protéines connues : les clusters 1, 2, 4, 5, 8, 9, 10, 11 et 12 ont ainsi pu être catégorisés facilement. Après les avoir annotés, j'ai donc pu chercher si les relations entre ces clusters correspondaient à celles attendues. En étudiant la distribution, il semble clair que le cluster 5 (ADN polymérases λ de métazoaires) est proche du cluster 8 (ADN polymérases β de métazoaires), ce qui est une relation attendue compte tenu des connaissances actuelles sur les ADN polymérases X. Le cluster 4 (ADN polymérases X de *Viridiplantae*) semble lui aussi proche du cluster 5. Les clusters 1 et 2 sont très proches, voire entremêlés, ce qui est attendu pour des séquences d'ADN polymérase μ et de Tdt de métazoaires. Enfin, les clusters de ADN polymérase X bactériennes sont à part, ce qui n'est pas très surprenant compte tenu de leur distance phylogénétique avec les autres séquences, presque toutes eucaryotes. Elles sont cependant plus proches du cluster 8, ce qui est logique puisque ce sont les seules ADN

polymérase X sans domaine BRCT. On remarque de plus que les deux clusters de séquences bactériennes semblent former un continuum, avec les séquences canoniques plus proches des ADN polymérases X eucaryotes. Enfin, le cluster 10 correspondant aux ADN polymérases IV de levures semble lui tout à fait à part, alors que dans la littérature, ces enzymes ont longtemps été décrites comme proches des ADN polymérases X de métazoaires.

2.2.3 La prédiction des structures des ADN polymérases X représentatives n'ayant pas de structure connue dans la PDB

Pour les clusters 3, 4, 6, 7, 9 et 10, aucune structure expérimentale n'était présente dans la PDB au moment de ces travaux. J'ai cependant pu profiter de l'essor d'AlphaFold2, qui dès 2021 a commencé à se montrer particulièrement efficace pour prédire des structures de protéines. Pour chaque séquence, j'ai d'abord utilisé l'outil AlphaFold2 accessible depuis ChimeraX pour obtenir un modèle de référence (ou bien, pour les clusters 4 et 10, j'ai pu trouver des prédictions existantes sur la base de données AlphaFoldDB).

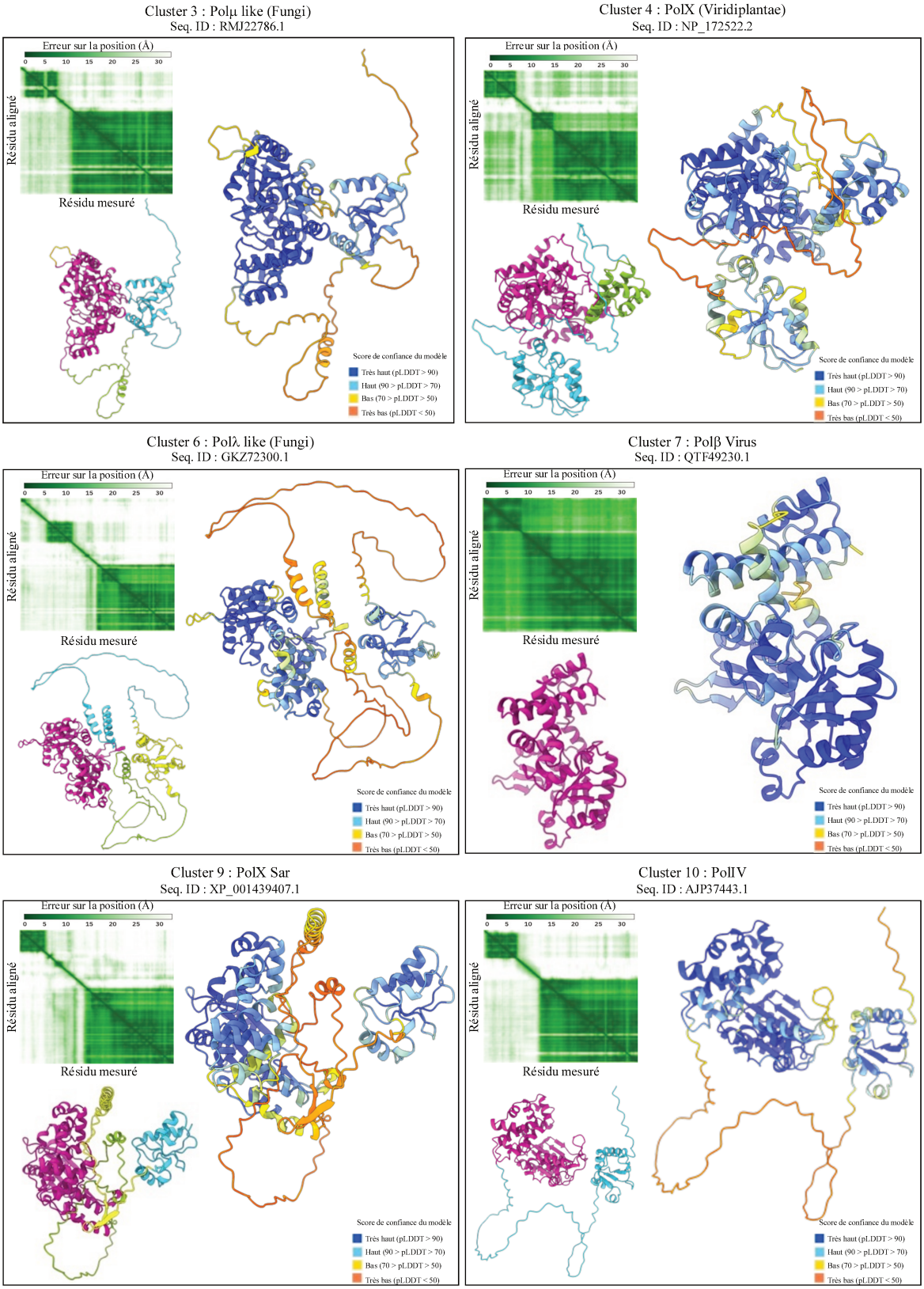


Figure 49 : Résumé des informations fournies par AlphaFold2 pour la prédiction des structures des séquences représentatives n'ayant pas de structures connues. Pour chaque cluster, une séquence a été choisie (voir Figure1). Pour les séquences n'ayant pas de structures connues, AlphaFold2 a été utilisé pour prédire la structure. Les représentants des clusters sont représentés ici dans des panneaux séparés. Pour chaque, le graphique PAE (Predicted Aligned Error) est indiqué. Pour chaque résidu (mesuré), l'intensité de la couleur verte représente l'erreur sur sa position relative à un autre résidu (aligné). Cette information permet de séparer des domaines, représentés par des couleurs différentes sur le modèle en bas à gauche. Le modèle de droite représente quant à lui le score de confiance du modèle par résidu : pour chaque résidu, la couleur indique la confiance du modèle dans sa localisation.

La figure 49 présente les modèles de référence des prédictions des structures des séquences choisies, et des données indiquant la confiance des modèles. Les éléments en vert représentent l'erreur sur la position entre tous les résidus de la protéine (*Predicted Aligned Error*, PAE). Plus l'intersection entre les coordonnées de deux résidus est foncée, plus la PAE est faible, donc leur proximité est certaine. Les résidus d'un même domaine sont donc regroupés dans des carrés vert foncé, séparés des autres domaines. A partir de ces informations, ChimeraX a pu séparer les différents domaines pour chaque protéine : ces domaines sont dans des couleurs différentes dans la structure présentée en bas à gauche de chaque panneau. On remarque que à l'exception des séquences de virus, toutes les séquences ont une extension N-terminale présentant au moins un domaine, en plus du cœur polymérase, toujours indiqué en pourpre. Enfin, la structure présentée à droite est la même, mais colorée en fonction du score pLDDT de chaque résidu de la protéine, qui reflète le niveau de confiance du modèle. Ce score, compris entre 0 (orange) et 100 (bleu), est un témoin de la précision de la prédiction de chaque résidu au sein de chaque structure secondaire : plus il est élevé, plus on considère que la prédiction est de qualité (Mariani *et al.*, 2013). Pour tous les modèles prédits ici, on remarque que les domaines détectés par l'analyse précédente ont un score pLDDT élevé, on peut donc être confiant dans les modèles proposés, au moins en ce qui concerne les parties structurées. Cependant, les séquences liant les domaines ont généralement un score pLDDT assez faible, et le PAE semble indiquer une assez grande incertitude quant à la position de ces résidus vis-à-vis des autres résidus de chaque protéine : on ne peut donc pas avoir une grande confiance dans la prédiction de ces parties, et dans les positions de chaque domaine vis-à-vis des autres. D'après ces informations, on peut donc globalement avoir confiance en ces prédictions, tant qu'on s'en tient aux domaines structurés.

Avec l'aide du Dr Pierre Legrand, j'ai ensuite pu obtenir pour chaque séquence 10 prédictions AlphaFold2 et 15 prédictions RosettaFold, différentes des modèles de références présentés ci-dessus. En plus des données fournies par AlphaFold2 indiquant la crédibilité des modèles de références, j'ai superposé les domaines structurés (BRCT, polymérase) de chaque modèle de référence avec ses homologues prédits par AlphaFold2 et RosettaFold. Pour chaque comparaison [Référence VS Modèle], j'ai ainsi obtenu une mesure de RMSD, indiquant la distance moyenne entre les carbones α homologues des deux structures comparées. J'ai ensuite

moyenné les mesures entre les modèles de références et leurs homologues AlphaFold2 et RosettaFold.

Tableau 3 : Statistiques des alignements pour chaque cluster entre le modèle de référence et 10 prédictions AlphaFold2 ou 15 prédictions RosettaFold. Les modèles de référence AlphaFold2 ont été comparés par superposition à 10 modèles AlphaFold2 obtenus indépendamment et 15 modèles RosettaFold. Pour chaque cluster, le RMSD moyen et le nombre de carbones α homologues concernés par ce RMSD sont indiqués. Pour les modèles présentant plusieurs domaines, les statistiques des différents domaines sont indiquées séparément.

	Cluster 3 : Polμ like				**Cluster 4 : PolX Viridiplantae**				**Cluster 6 : PolX like**				**Cluster 7 : Polβ Virus**		**Cluster 9 : PolX Sar**				**Cluster 10 : PolIV**			
	RMJ22786.1				**NP_172522.2**				**GKZ72300.1**				**QTF49230.1**		**XP_001439407.1**				**AJP37443.1**			
	AlphaFold2		RosettaFold		AlphaFold2		RosettaFold		AlphaFold2		RosettaFold		AlphaFold2	RosettaFold	AlphaFold2		RosettaFold		AlphaFold2		RosettaFold	
	BRCT	Pol	BRCT	Pol	BRCT	Pol	BRCT	Pol	BRCT	Pol	BRCT	Pol	Pol	Pol	BRCT	Pol	BRCT	Pol	BRCT	Pol	BRCT	Pol
RMSD moyen (Å)	0,396	0,304	0,809	0,876	0,265	0,554	0,725	1,083	0,241	0,624	0,883	1,264	0,717	1,121	0,364	0,627	0,808	1,113	0,370	0,524	0,839	1,248
Nombre de paires d'atomes comparées	100.0	356.0	83.7	315.1	97.5	253.7	94.9	250.8	92.0	311.0	88.0	189.7	260.2	177.4	100.4	323.4	85.8	239.6	113.7	307.0	103.5	240.6

Comme l'indique le tableau 3, pour les domaines structurés des modèles obtenus, on n'observe aucun RMSD supérieur à 1,264 angströms (Å). Cela suppose que toutes les prédictions obtenues sont proches, et ce peu importe leur provenance (AlphaFold2 ou RosettaFold). Cette information, en plus des données fournies par AlphaFold2 pour les modèles de référence, indique que ces modèles sont assez crédibles, puisque deux algorithmes de prédiction structurale différents prédisent des repliements très proches.

2.2.4 Analyse des séquences des différents clusters

Après avoir sélectionné un ensemble de séquences représentatives pour chaque cluster, j'ai pu analyser l'ensemble des séquences dans chaque cluster. Dans un premier temps, j'ai voulu connaitre la longueur moyenne des séquences de chaque cluster, et la variabilité au sein des clusters.

N° du cluster	1	2	3	4	5	6	7	8	9	10	11	12
Moyenne (nombre d'acides aminés)	496,43	493,4	653,29	536,68	563,95	781,78	350,14	343,14	621,44	576,99	569	566,46
Écart-type	29,2	34,33	134,64	86,39	61,52	150,17	17,86	51,97	83,72	83,03	38	18,28
Nombre de séquences	338	451	1045	512	925	1463	14	503	18	146	526	188

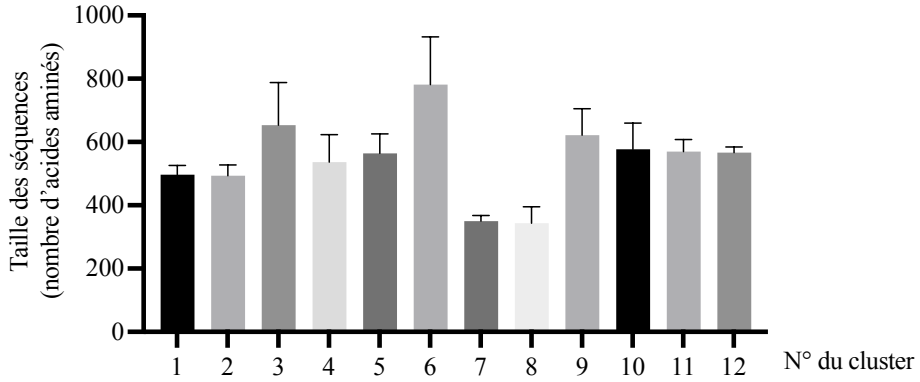


Figure 50 : Tailles moyennes des séquences des 12 clusters. En haut, le tableau indique les tailles moyennes des séquences des 12 clusters (en nombre d'acides aminés), ainsi que l'écart type et le nombre de séquences de chaque cluster. En bas, les mêmes informations sont représentées sous forme d'histogramme.

Concernant la longueur des séquences, quatre clusters semblent se démarquer : les clusters 3, 6, 7 et 8. Concernant le cluster 7, cela n'est pas surprenant, les ADN polymérases β de métazoaires n'ayant pas de domaine BRCT, elles sont plus courtes que les autres. On peut donc supposer que les ADN polymérases X de virus (cluster 8) se rapprochent des ADN polymérases β. Concernant les clusters 3 et 6, cela est plus surprenant, car ces séquences sont annotées comme des ADN polymérases μ et λ, et devraient donc être plus proches des clusters 2 et 5 respectivement. On note cependant que c'est aussi dans ces deux clusters que la variabilité est la plus grande : l'écart-type est de 134,64 résidus pour le cluster 3, et de 150,17 résidus pour le cluster 6, alors que pour les autres séquences, cet écart type est inférieur à 90 résidus. Cela suppose que parmi les très nombreuses séquences présentes dans ces clusters, certaines sont plus longues de façon aberrante. Les autres clusters semblent tous présenter des séquences avec des tailles relativement similaires, autour de 500-550 résidus.

Afin d'essayer de caractériser les séquences présentes dans ces clusters, j'ai ensuite réalisé avec PSI-Coffee des alignements des séquences représentatives choisies, que j'ai complété avec des alignements structuraux à partir des structures et des modèles connus, ou à partir de prédictions AlphaFold2. Les premiers alignements globaux ont indiqué que les séquences des clusters 7, 11 et 12 ne portent pas de domaine BRCT, tout comme les ADN

polymérases β de métazoaires du cluster 8 ; et que seules les séquences bactériennes (clusters 11 et 12) possédaient une extension C-terminale, correspondant au domaine PHP décrit en introduction.

2.2.4.1 Les domaines BRCT

Dans un premier temps, j'ai donc réalisé un alignement des séquences des séquences situées en N-terminal des domaines catalytiques des séquences représentatives des clusters 1, 2, 3, 4, 5, 6, 9 et 10.

Comme le montre la figure 51, ces domaines N-terminaux présentent une assez grande variabilité de séquence. En particulier, certains d'entre eux présentent de longues insertions dans des boucles qui chez les ADN polymérases X connues sont peu structurées et qui font la liaison entre les structures secondaires importantes. On peut cependant noter que certains patchs de résidus hydrophobes sont très bien conservés, parmi toutes les séquences alignées ici. Cette observation est cohérente avec les connaissances actuelles sur les domaines BRCT (Leung and Glover, 2011).

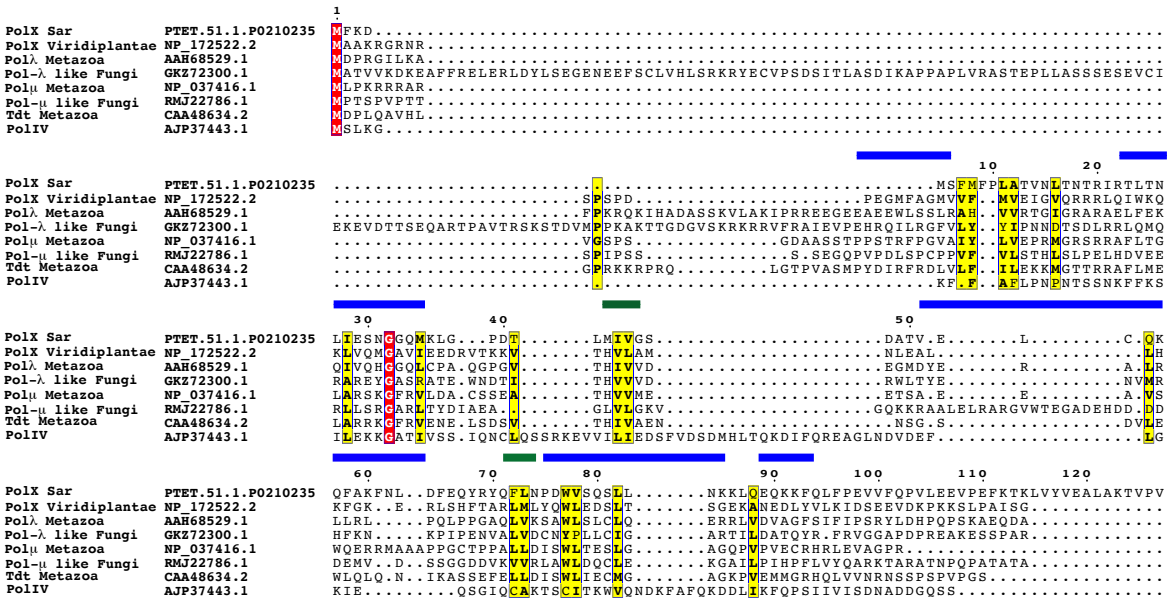


Figure 51 : Alignement des séquences des domaines N-terminaux des protéines représentatives des clusters 1, 2, 3 ,4, 5, 6, 9 et 10, obtenu par PSI-Coffee. Les structures secondaires du domaine BRCT de l'ADN polymérase λ humaine sont indiquées (en bleu : hélices α ; en vert : brins β). Les séquences linker entre ces domaines et les domaines catalytiques ne sont pas montrées.

Cependant, les séquences des domaines BRCT étant variables, le meilleur moyen de déterminer si les ADN polymérases X étudiées ici ont bien un domaine BRCT est de comparer

leurs structures. J'ai donc réalisé un alignement des structures de ces ADN polymérases X, en utilisant des structures expérimentales ou des prédictions AlphaFold2.

Figure 52 : Alignement structural des domaines N-terminaux des protéines représentatives des clusters 1, 2, 3,4, 5, 6, 9 et 10, réalisé sur ChimeraX. Le repliement global de tous ces domaines est très similaire, sauf en ce qui concerne les régions faisant le lien entre les structures secondaires.

Comme le montre la figure 52, le repliement global de ces extensions N-terminales est très proche pour toutes les structures comparées. Celles-ci incluant des ADN polymérase X connues pour porter un domaine BRCT (ADN polymérase λ, μ, Tdt, ADN polymérase IV), on peut affirmer que toutes ces polymérases ont bien un domaine BRCT en N-terminal. On retrouve les structures secondaires caractéristiques des domaines BRCT : un feuillet β central (4 brins) entouré d'une hélice α d'un côté, et 2 de l'autre. Comme évoqué plus haut avec les alignements de séquences, on observe cependant une certaine variabilité au niveau des boucles liant les structures secondaires (avec en particulier une grande insertion chez l'ADN polymérase λ-like de *Aspergillus* ; chez cette même polymérase, le domaine BRCT est d'ailleurs précédé d'une longue séquence, non structurée d'après AlphaFold2). Toutes ces informations sont cohérentes avec les connaissances actuelles sur les domaines BRCT (Leung and Glover, 2011). La présence de domaines BRCT chez les polymérase moins connues (ADN polymérases λ de Fungi, ADN polymérases μ de Fungi, ADN polymérases X de Viridiplantae, de Sar) suppose que ces ADN polymérases X sont impliquées dans le NHEJ, ou du moins qu'elles peuvent interagir avec des protéines impliquées dans le NHEJ, comme les ADN polymérases X connues.

2.2.4.2 Les domaines catalytiques

La partie la plus approfondie des comparaisons entre ces ADN polymérases X se focalise sur leurs domaines catalytiques. Comme pour les domaines BRCT, j'ai commencé par aligner

les séquences de ces domaines (figure 53), mais j'ai aussi utilisé leurs structures, afin de rechercher des caractéristiques précises.

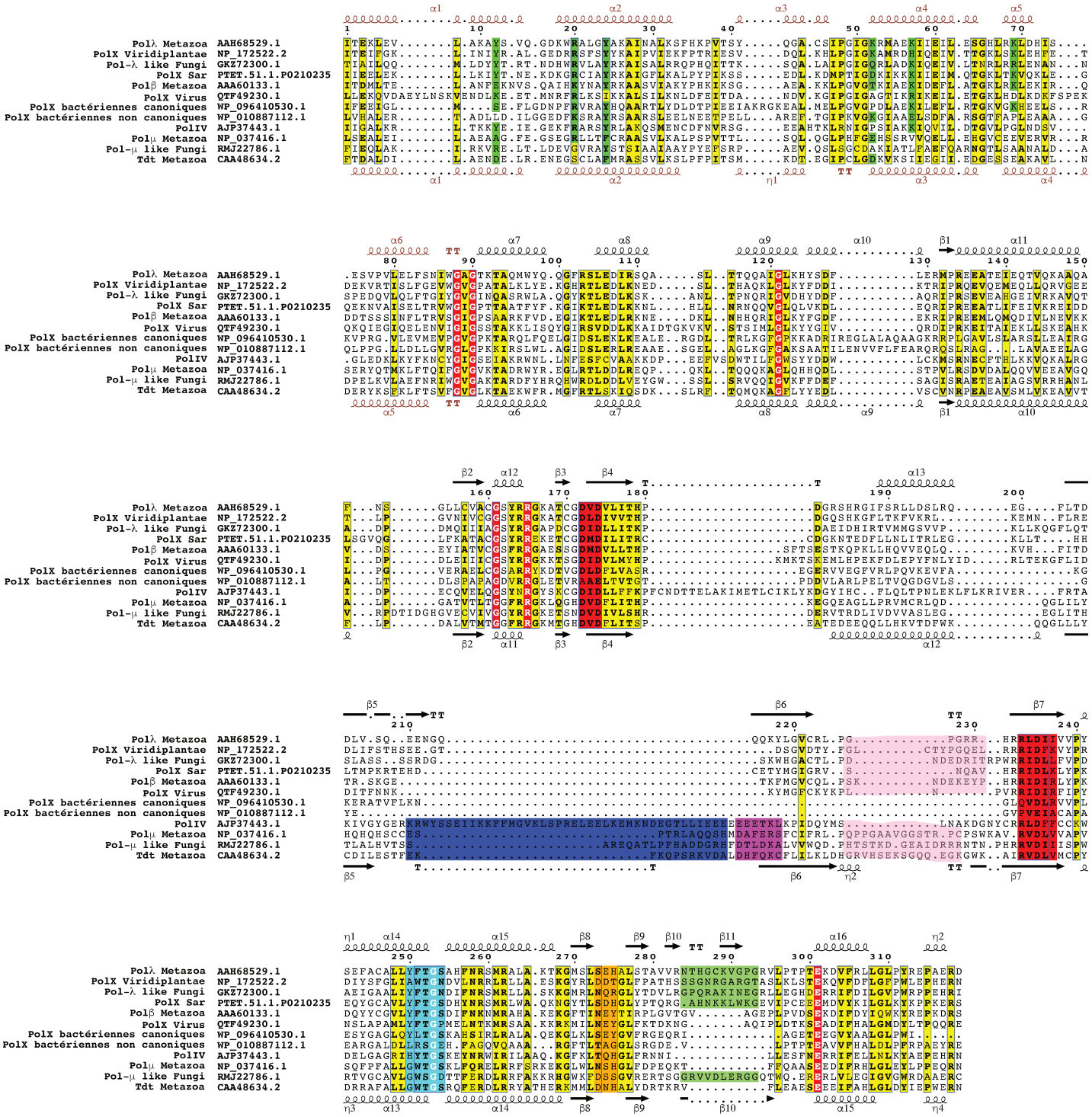


*Figure 53 : Alignement des séquences des domaine polymérase des séquences représentatives des 12 clusters, obtenu par PSI-Coffee. Les éléments caractéristiques des PolX connues sont indiqués en couleurs : résidus impliqués dans la reconnaissance des groupements 5'P (et activité dRP lyase) en **vert** ; motifs catalytiques en **rouge** ; boucle 1 en **bleu**, motif SD1 en **rose** ; boucle 2 en rose pâle ; steric gate en **cyan**, motif SD2 en **orange** et boucle 3 en **vert clair** . Les structures secondaires de la Polλ humaine (PDB 7M43) et de la Tdt de souris sont indiquées, respectivement au-dessus et en dessous de l'alignement. Les structures secondaires en **marron** au domaine de 8 kDa. Le reste correspond au domaine catalytique. Pour annoter les résidus impliqués dans la reconnaissance des groupements 5'P, la Polβ humaine a été utilisée comme référence. Pour les boucles 1 et 2 et le motif SD1, la Tdt de souris a été utilisée. Pour la boucle 3, c'est la Polλ humaine qui a été utilisée comme référence.*

Chez les ADN polymérases X connues, ce domaine catalytique porte plus qu'un cœur catalytique avec une activité ADN polymérase. Mon objectif a donc été de trouver plusieurs particularités, pouvant être spécifiques de certaines activités ou de certaines polymérases :

- Résidus impliqués dans la reconnaissance du 5'P et l'activité dRP lyase
- Le motif DxD catalytique
- La boucle 1
- Le motif SD1
- La boucle 2
- Le motif RxDx(Φ/+) catalytique
- Le *steric gate*
- Le motif SD2
- La boucle 3

Dans cette partie, je me suis attardé sur les différentes séquences représentatives étudiées en précisant la présence ou l'absence de chacune de ces caractéristiques, puis j'ai étendu ces observations à l'ensemble des séquences des clusters associés pour esquisser une description de chacun des groupes d'ADN polymérases X (**je recommande donc de se référer à la figure 53 pour simplifier la lecture**).

2.2.4.2.1 Cluster 8 : Les ADN polymérases β de Metazoa (séquence représentative : AAA60133.1)

Les séquences du cluster 8, en particulier celle de l'ADN polymérase β humaine, sont bien connues, puisqu'elles sont étudiées depuis plus de 40 ans. La séquence choisie ici pour détailler les spécificités de ce cluster est justement celle de l'ADN polymérase β humaine.

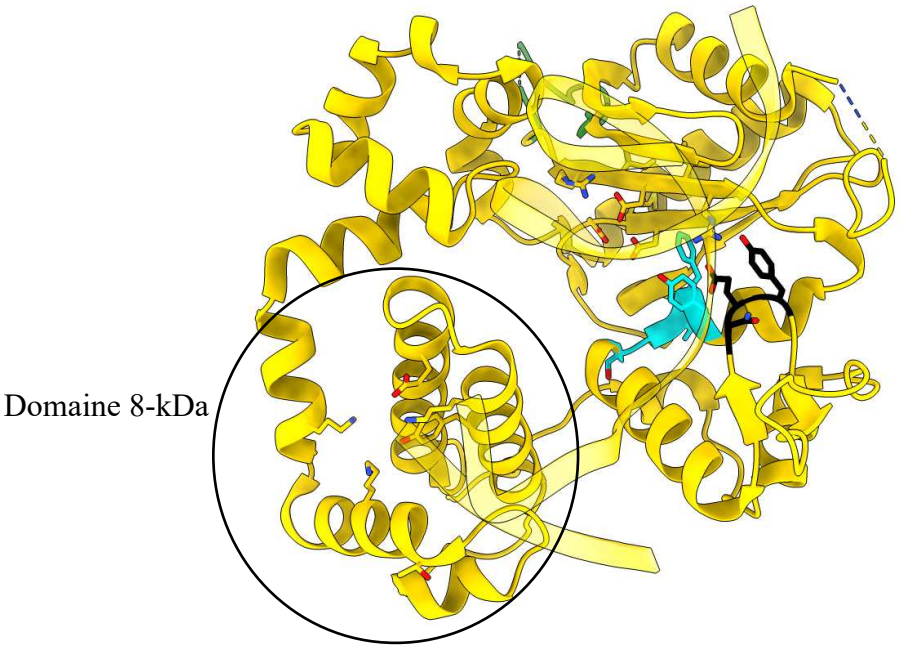


*Figure 54 : Structure de l'ADN polymérase β humaine (PDB : 5tb8). Entouré : domaine 8 kDa, pour lequel les résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase. Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; la boucle 2 est représentée en vert ; le steric gate est représenté en cyan ; le motif SD2 est représenté en **noir**.*

Celle-ci ne présente pas de domaine BRCT, et a donc en N-terminal un domaine 8-kDa, qui porte tous les résidus impliqués dans la reconnaissance des groupements 5'P et dans l'activité dRP lyase, principalement des résidus chargés, surtout des tyrosines et arginines. Pour l'étude des autres clusters, ce sont ces résidus qui ont été recherchés dans le domaine 8-kDa. Le motif catalytique de l'ADN polymérase β est un DMD associé à un motif RIDIR. Ce second motif porte non seulement un résidu aspartate catalytique, mais aussi une arginine (en position 5) impliquée dans le mécanisme *induced-fit* de l'ADN polymérase β. Ce mécanisme ne repose cependant pas que sur ce résidu, mais aussi sur les résidus du motif SD2 (NEY). L'ADN polymérase β porte un *steric gate* YFTGS, bloquant l'entrée aux NTPs dans le site actif. Enfin, elle ne possède pas de boucle 1 ou 3 ni de motif SD1 mais a une courte boucle 2.

Pour confirmer si ces observations étaient généralisables aux séquences de ce cluster, j'ai généré une visualisation « Logo » des séquences de ce cluster. Dans ce type de visualisation, plus un acide aminé est fréquent à sa position parmi les séquences données, plus il prend de place : si un résidu est grand, c'est donc qu'il est très conservé.

Figure 55 : Motifs d'intérêt des séquences du cluster 8. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre)

D'après la figure 55, l'ensemble des observations présentées ci-dessus peuvent être généralisées aux ADN polymérases β de métazoaires présentes dans ce cluster.

2.2.4.2.2 Cluster 1 : Les Tdt de métazoaires (séquence représentative : CAA48634.2)

Le cluster 1 représente lui aussi un groupe d'ADN polymérases X bien caractérisées, puisque la Tdt humaine est étudiée depuis les années 1960. Ces ADN polymérases X ne portent pas d'activité dRP lyase, ni de capacité à reconnaitre les groupements 5'P, contrairement aux ADN polymérases λ et β. La comparaison avec les autres séquences montre l'absence des résidus impliqués dans ces activités.

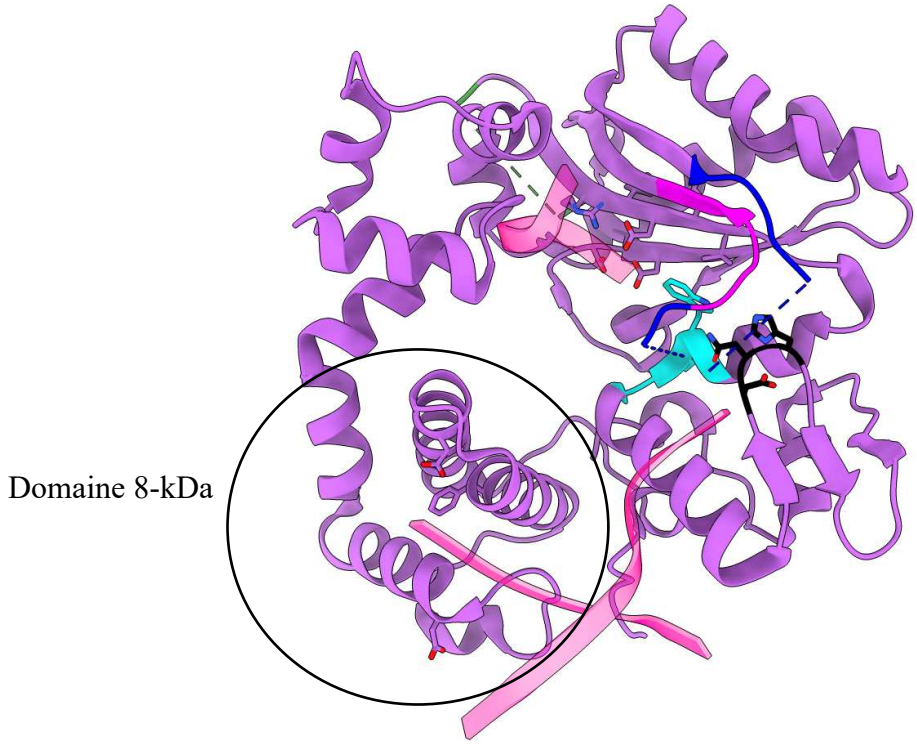


*Figure 56 : Structure de la Tdt humaine (PDB : 4qz8). Entouré : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase. Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; la boucle 1 est représentée en **bleu** ; le motif SD1 en magenta ; la boucle 2 est représentée en vert ; l'équivalent du steric gate est représenté en cyan ; le motif SD2 est représenté en **noir**.*

Le premier motif catalytique, ici un DVD, partage l'activité polymérase avec un dernier aspartate catalytique situé dans un motif RVDLV. On note ensuite la présence d'une boucle 1. Cette boucle est connue pour permettre à la polymérase d'assembler une synapse avec un brin template plus court que le brin primer. Cette boucle 1 est suivie du motif SD1, formé des résidus DHFQKC, impliqué lui aussi dans les mêmes rôles. La Tdt porte aussi une boucle 2, dont le rôle n'est pas déterminé à ce jour. En aval, juste après le second motif catalytique, la Tdt porte un motif GWTGS, ne présentant pas les résidus YF formant le *steric gate* des ADN polymérases λ et β : la Tdt peut donc incorporer des NTPs (Boulé *et al.*, 2001). Le motif SD2, qui permet de discriminer les ADN polymérases X, est pour la Tdt un DNH. Enfin, la Tdt ne porte pas de boucle 3.

Figure 57 : Motifs d'intérêt des séquences du cluster 1. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre)

D'après la figure 57, il semble que la majorité des observations ci-dessus est généralisable aux Tdt de métazoaires. Il semble que la boucle 1 soit assez peu variable, contrairement à la boucle 2. Le rôle de la première étant principalement structural, cela n'est pas surprenant : il a été montré que la conservation des résidus de la boucle 1 de la Tdt est ce qui permet de garder sa conformation tout au long du cycle catalytique (Loc'h *et al.*, 2019) ; concernant la boucle 2 cependant, son rôle n'étant pas déterminé, il est difficile de conclure.

2.2.4.2.3 Cluster 2 : Les ADN polymérases μ de métazoaires (séquence représentative : NP_037416.1)

Cette fois encore, ce cluster regroupe des ADN polymérases bien caractérisées, puisque l'ADN polymérase μ humaine est connue et étudiée depuis 2000.

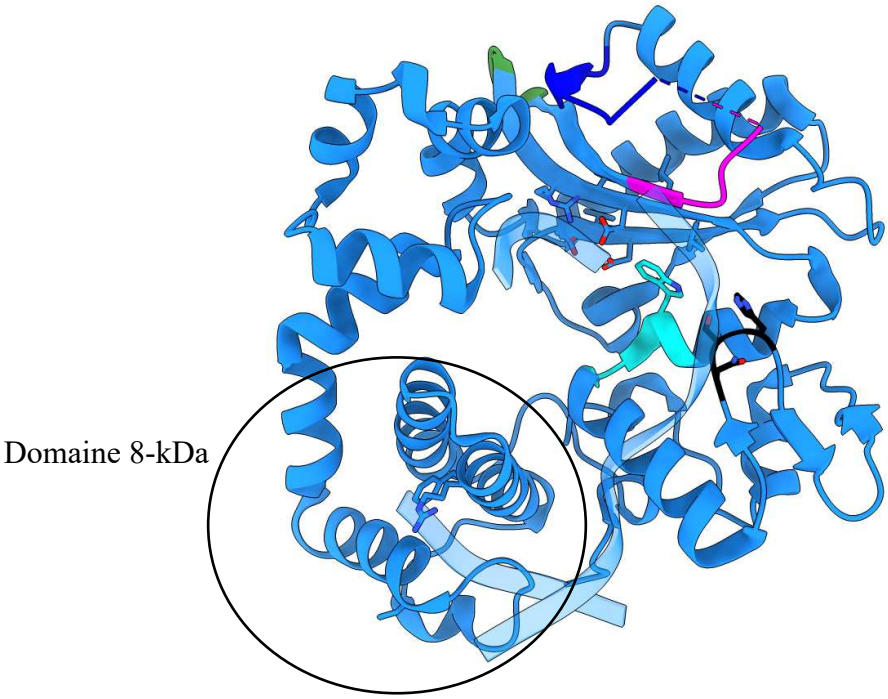


*Figure 58 : Structure de l'ADN polymérase*μ *humaine (PDB : 6ak8). Entouré : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase. Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; la boucle 1 est représentée en bleu ; le motif SD1 en magenta ; la boucle 2 est représentée en vert ; l'équivalent du steric gate est représenté en cyan ; le motif SD2 est représenté en* ***noir****.*

Bien qu'elle ne porte pas d'activité dRP lyase, l'ADN polymérase μ a une poche lui permettant de reconnaitre les groupements 5'P dans son domaine 8-kDa, mais ne porte pas d'équivalents de tous les résidus de l'ADN polymérase β (6 résidus chez l'ADN polymérase β, 3 chez l'ADN polymérase μ). On retrouve cependant bien les motifs catalytiques DVD et RVDLV, et la boucle 1 (absente de la structure présentée car trop flexible dans la structure déposée dans la PDB). Elle est directement suivie du motif SD1, formé des résidus DAFERS. L'ADN polymérase μ semble également porter une boucle 2, non visible ici, comme la boucle 1. Le motif du *steric gate* montre là encore une absence des résidus bloquant le 2' OH du sucre des ribonucléotides (GWTGS, comme pour la Tdt). Le motif SD2 de la Polμ est un NSH, et cette polymérase ne présente pas non plus de boucle 3.

Figure 59 : Motifs d'intérêt des séquences du cluster 2. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre)

Comme pour la Tdt, la plupart des observations faites ci-dessus semblent généralisable, d'après la figure 59. Cependant, la boucle 1 semble beaucoup plus variable que chez la Tdt, et la boucle 2 encore plus (elle semble même pouvoir varier en longueur, comme on peut le voir aux quelques résidus rares présents dans la figure 58).

2.2.4.2.4 Cluster 5 : Les ADN polymérases λ de métazoaires (séquence représentative : AAH68529.1)

Comme les clusters 1, 2 et 8, ce cinquième cluster contient des séquences bien caractérisées, comme l'ADN polymérase λ humaine. Celle-ci présente dans son domaine de 8 kDa tous les éléments lui conférant une activité dRP lyase. Ses motifs catalytiques sont formés des résidus DVD et RVDII, elle ne porte pas de boucle 1 ni de SD1 mais porte une courte boucle 2 et a un *steric gate* (YFTGS) lui permettant de bloquer l'incorporation de NTPs. Son motif SD2 est formé des résidus SEH, et elle a une boucle 3.

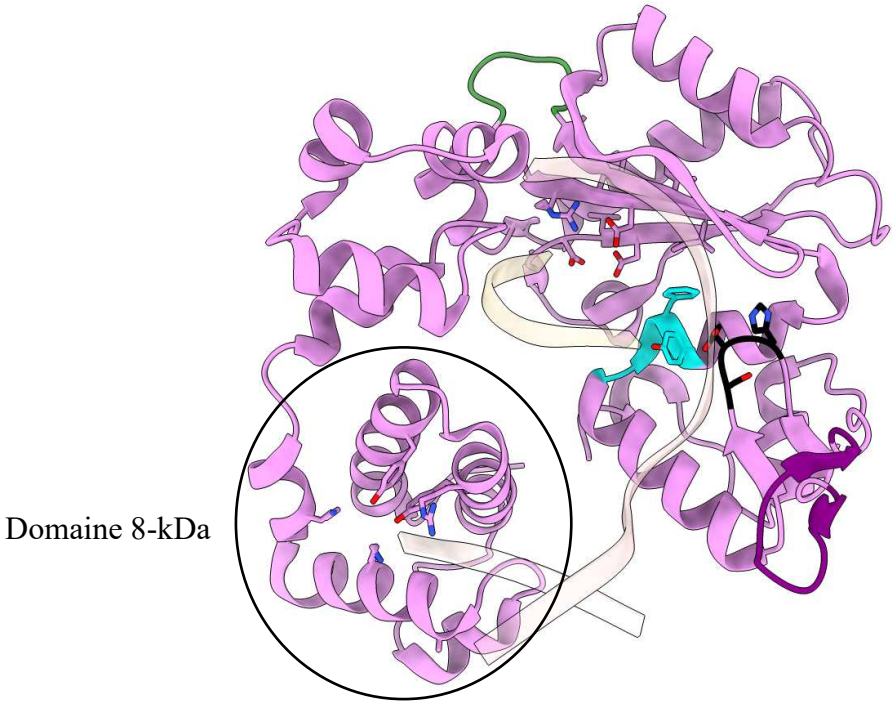


*Figure 60 : Structure de l'ADN polymérase λ humaine (PDB : 1rzt). Entouré : domaine 8 kDa, pour lequel les résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase. Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; la boucle 2 est représentée en vert ; l'équivalent du steric gate est représenté en cyan ; le motif SD2 est représenté en **noir** ; la boucle 3 est représentée en pourpre.*

Comme le montre la figure 61, toutes ces observations sont généralisables aux ADN polymérases λ de métazoaires en général.

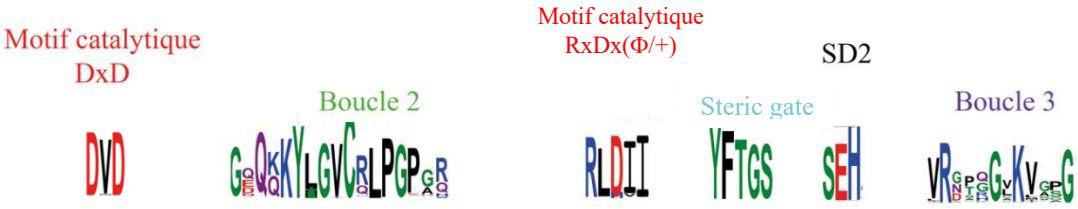


Figure 61 : Motifs d'intérêt des séquences du cluster 5. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre)

2.2.4.2.5 Cluster 3 : les ADN polymérases μ-like de Fungi (séquence représentative : RMJ22786.1)

Ce troisième cluster est le premier à présenter des ADN polymérases qui ont été peu caractérisées par le passé. L'analyse de la séquence choisie s'est donc faite sans *a priori* sur l'activité de ces polymérases, et je vais donc relever ici les particularités de cette ADN polymérase X et en déduire les éventuels impacts sur son activité.

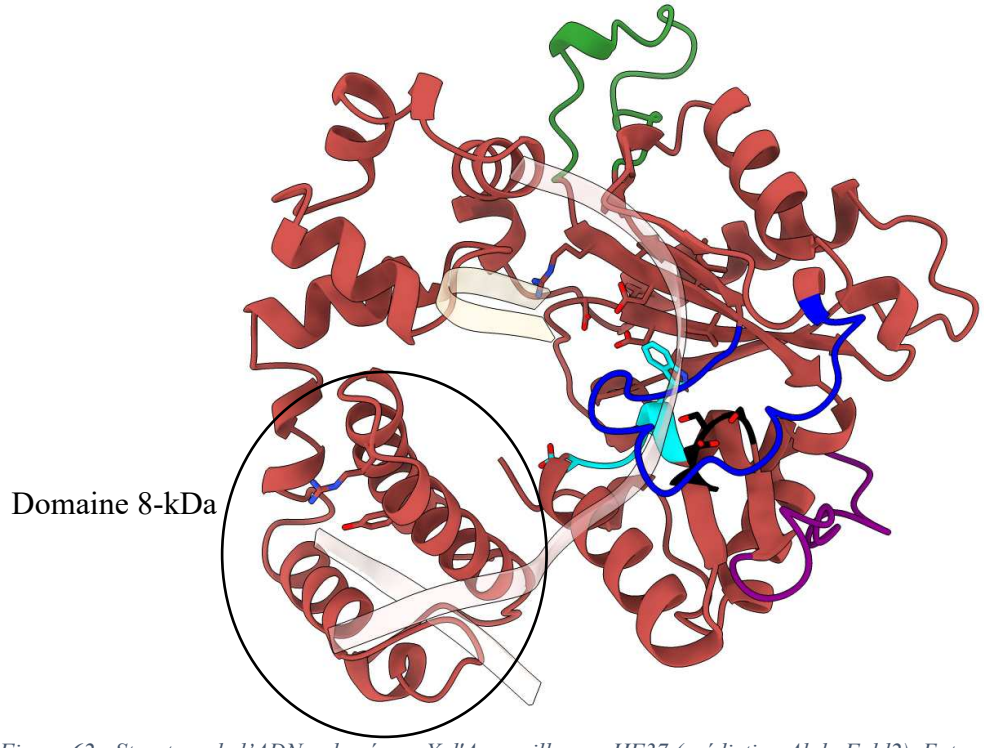


*Figure 62 : Structure de l'ADN polymérase X d'Aspergillus sp. HF37 (prédiction AlphaFold2). Entouré : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (l'ADN est celui de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; la boucle 1 est représentée en bleu ; le motif SD1 en magenta ; la boucle 2 est représentée en vert ; l'équivalent du steric gate est représenté en cyan ; le motif SD2 est représenté en **noir**; la boucle 3 est représentée en pourpre.*

Que ce soit *via* sa structure ou *via* sa séquence, on remarque que cette ADN polymérase a des caractéristiques globalement proches des versions métazoaires de la Tdt ou de l'ADN polymérase μ. Tout d'abord, elle semble porter dans son domaine 8 kDa des vestiges de résidus impliqués dans la reconnaissance du 5'P, comme l'ADN polymérase μ. Son motif catalytique est un DVD, associé à un RVDII, à mi-chemin entre les ADN polymérases μ et λ de métazoaires. Elle porte bien une boucle 1, un motif SD1 (DTLDKA), et une boucle 2. Elle n'a pas de *steric gate*, mais la séquence du motif associée est un GWSGS, proche du GWTGS de l'ADN

polymérase μ métazoaire. Son motif SD2 la différencie de toutes les ADN polymérases X connues (DSS), et elle semble porter une boucle 3, comme l'ADN polymérase λ métazoaire.

Figure 63 : Motifs d'intérêt des séquences du cluster 5. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre)

Comme le montre la figure 63, ces observations sont généralisables à l'ensemble des séquences de ce cluster, même si la boucle 1 et surtout la boucle 2 semblent très variables en longueur (de 5 à 20 résidus).

L'ensemble de ces caractéristiques en font une ADN polymérase X proche de l'ADN polymérase μ, donc son activité est probablement proche de cette dernière : grâce à son domaine BRCT elle est probablement impliquée dans le NHEJ chez les *Fungi* chez qui elle est présente, elle peut incorporer des NTPs et présente probablement une faible fidélité (voire une activité terminal transférase en présence d'ions Mn^{2+}). Cependant, les différences, en termes de motif SD2 et de présence d'une boucle 3, ne permettent pas d'affirmer qu'il s'agit d'un cluster regroupant des ADN polymérases μ. C'est pourquoi ces séquences sont ici nommées ADN polymérases μ-like.

Cependant, ce cluster pourrait probablement être subdivisé dans de futures études plus poussées sur ces polymérases encore peu étudiées.

2.2.4.2.6 Cluster 4 : Les ADN polymérases X de Viridiplantae (séquence représentative : NP_172522.2)

Comme indiqué plus haut, ce cluster est le seul à contenir des ADN polymérases X provenant de plantes, dont les ADN polymérases X nommées jusqu'ici ADN polymérases λ (Uchiyama *et al.*, 2004). Cependant, la séquence choisie ici montre des différences avec les ADN polymérases λ de métazoaires.

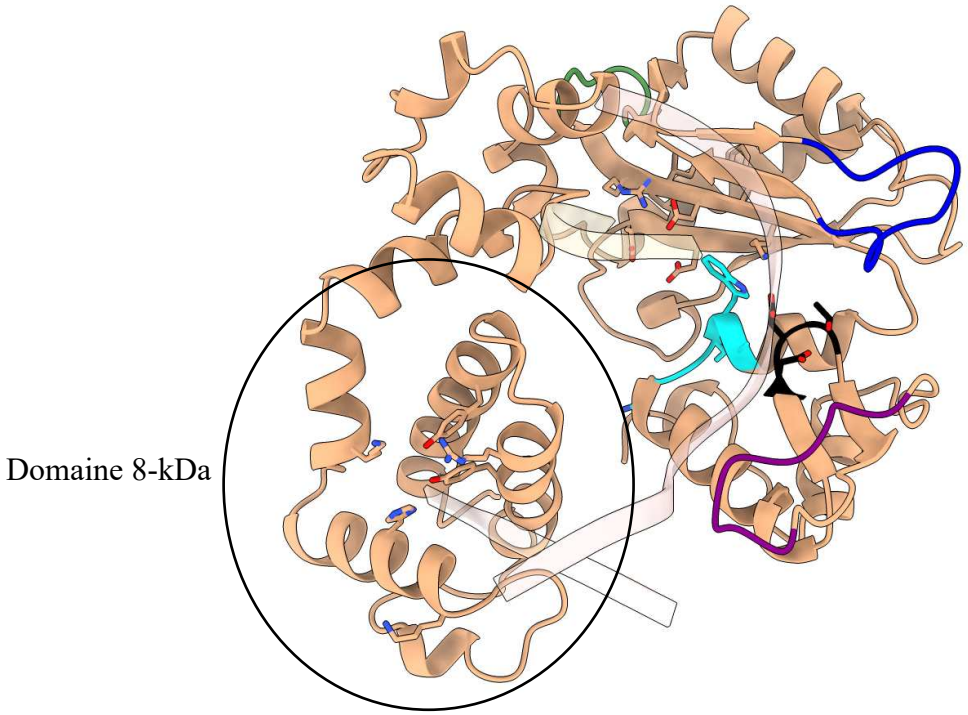


Figure 64 : Structure de l'ADN polymérase X d'Arabidopsis thaliana (prédiction AlphaFold2). Entouré : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (l'ADN est celui de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; la boucle 1 est représentée en bleu ; le motif SD1 en magenta ; la boucle 2 est représentée en vert ; l'équivalent du steric gate est représenté en cyan ; le motif SD2 est représenté en ***noir****; la boucle 3 est représentée en pourpre.*

Cette ADN polymérase semble avoir presque tous les résidus nécessaires à une activité dRP lyase, ainsi que tous les résidus catalytiques impliqués dans l'activité polymérase (DLD et RIDFK). Elle n'a pas de boucle 1, ni de motif SD1 ou de boucle 2, ce qui la différencie clairement des Tdt et ADN polymérases μ. Là où cette polymérase se différencie le plus des ADN polymérases X connues, c'est au niveau du *steric gate*. Ici la séquence est un motif AWTGN, ce qui diffère des YFTGS et GWTGS très conservés retrouvés habituellement. Cependant, cela semble plus proche de la séquence GWTGS retrouvée chez la Tdt et l'ADN polymérase μ. On peut donc supposer que cette enzyme peut incorporer des NTPs, ce qui la différencie clairement des ADN polymérases λ. Enfin, cette ADN polymérase a un motif SD2 formé des résidus DDT, et semble avoir une boucle 3, spécifique des séquences proches de l'ADN polymérase λ. De façon intéressante, cette prédiction structurale est aussi la seule pour laquelle la localisation du domaine BRCT semble proche de celle du domaine polymérase (figure 48). D'après les connaissances sur les ADN polymérases X et la flexibilité du linker liant ces deux domaines,

cette prédiction est probablement fausse, mais il serait intéressant d'étudier cette possible interaction

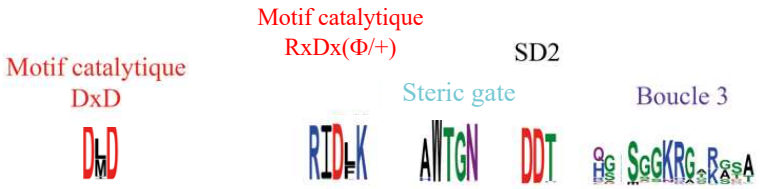


Figure 65 : Motifs d'intérêt des séquences du cluster 4. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre)

Toutes ces observations semblent généralisables aux ADN polymérases X présentes dans le cluster 4 (voir figure 65), et semblent indiquer que ces ADN polymérases X ne sont pas des ADN polymérases λ, contrairement à ce qui était proposé jusqu'ici. En effet, l'absence de *steric gate* est incompatible avec une description d'ADN polymérase λ, et le motif SD2 différencie aussi ces ADN polymérases X des ADN polymérases λ connues jusqu'ici. Dans ces travaux, ce cluster sera donc considéré comme un cluster regroupant les ADN polymérases X de plantes, en absence d'appellation fixe pour ces polymérases à mi-chemin entre les ADN polymérases λ et μ.

2.2.4.2.7 Cluster 6 : Les ADN polymérases λ-like de Fungi (séquence représentative : GKZ72300.1)

Le cluster 6 représente l'autre versant des ADN polymérases X de *Fungi* présentes dans ce clustering. Là encore, ces ADN polymérases X ont été peu étudiées, ce qui permet une analyse de leurs séquences sans *a priori*.

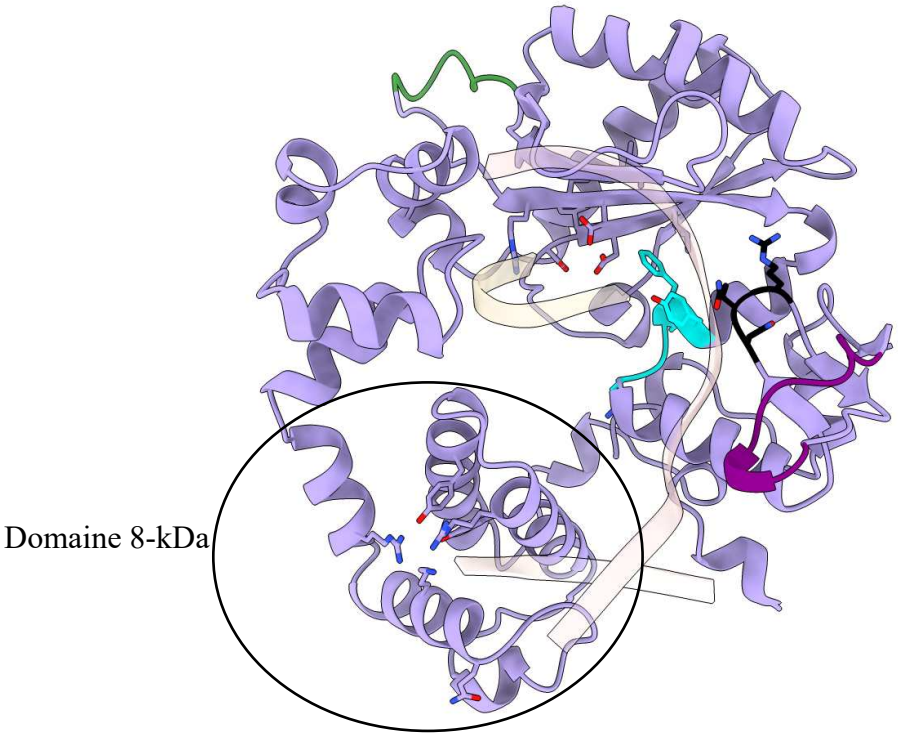


*Figure 66 : Structure de la PolX d'Aspergillus niger (prédiction AlphaFold2). Entouré: domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (l'ADN est celui de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; la boucle 1 est représentée en bleu ; le motif SD1 en magenta ; la boucle 2 est représentée en vert ; l'équivalent du steric gate est représenté en cyan ; le motif SD2 est représenté en **noir**; la boucle 3 est représentée en pourpre.*

Tout d'abord, la séquence étudiée ici semble clairement contenir les résidus essentiels à l'activité dRP lyase. Ensuite, cette polymérase a des motifs catalytiques formés des acides aminés DID et RIDLL. Elle n'a pas de boucle 1 ni de motif SD1, mais a une boucle 2 et une *steric gate* YFTGN, qui empêche probablement l'incorporation de NTPs. Son motif SD2 est de séquence NQR, et elle porte une boucle 3. Mis à part le motif SD2, qui la différencie des ADN polymérases λ de métazoaire, cette polymérase partage de nombreuses caractéristiques avec les ADN polymérases λ connues.

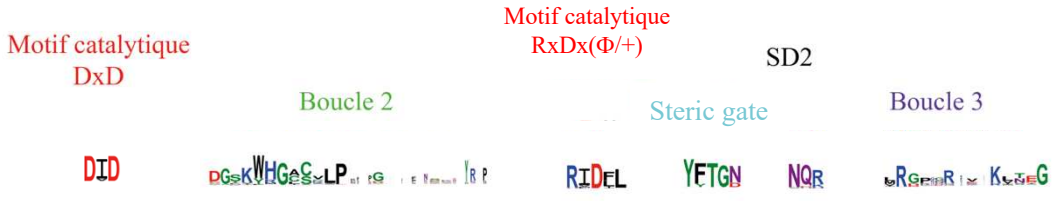


Figure 67 : Motifs d'intérêt des séquences du cluster 6. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre).

Ces observations semblent généralisables aux séquences de ce cluster d'après la figure 67, bien que les boucles 2 et 3 semblent de tailles et de séquences variables. La forte similarité avec les caractéristiques des ADN polymérases λ de métazoaires (*steric gate* et boucle 3) en font donc un groupe appelé ici ADN polymérases λ-like.

Tout comme pour le cluster 3, ce cluster pourrait cependant être divisé en deux dans de futures études visant à mieux caractériser les ADN polymérases X chez les Fungi.

2.2.4.2.8 Cluster 7 : Les ADN polymérases X de virus (séquence représentative : QTF49230.1)

Le septième cluster obtenu à partir des données montrées ici semble être une nouveauté, du fait de son origine virale (principalement des *Megavirus* ou « virus géants »). D'abord, du point de vue de sa séquence, la polymérase choisie ici ressemble beaucoup à une ADN polymérase β. En effet, elle ne porte pas de domaine BRCT ou d'extension N-terminale, a presque tous les résidus nécessaires à une activité dRP lyase, n'a pas de boucles 1, 2 ou 3 ou de SD1, mais elle a bien un *steric gate* (YFTGP), et son motif SD2 est le même que pour les ADN polymérases β de métazoaires : NEY. Ses motifs catalytiques sont également très proches des ADN polymérases β métazoaires : DID et RIDIR.

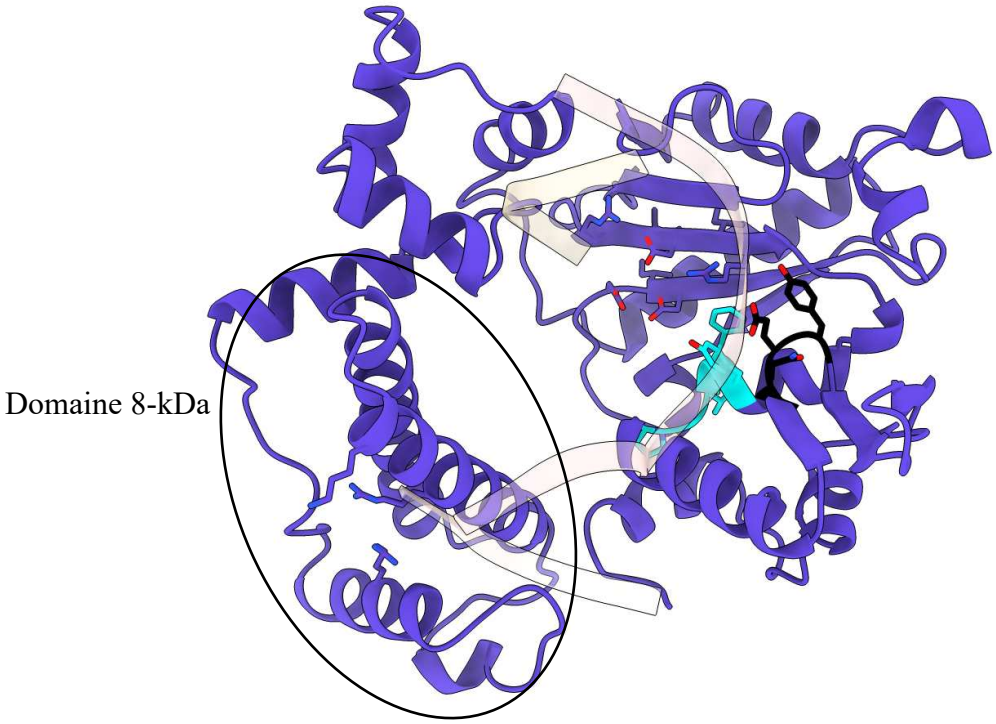


Figure 68 : Structure de l'ADN polymérase X de Mimivirus reunion (prédiction AlphaFold2). Entouré : domaine 8 kDa, pour lequel les résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (l'ADN est celui de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; le steric gate est représenté en cyan ; le motif SD2 est représenté en ***noir****.*

Ces observations sont généralisables aux autres séquences de ce cluster (même s'il est petit : 14 séquences seulement), donc nous parlerons ici d'ADN polymérase β de virus.

Figure 69 : Motifs d'intérêt des séquences du cluster 7. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre).

Plus précisément, les séquences de ce cluster appartiennent à des virus géants, des virus infectant des amibes (eucaryotes unicellulaires). Ces virus sont connus pour avoir un grand génome (jusqu'à 2,5 Mb) contenant des gènes codants pour des protéines du métabolisme ou de la réparation du génome (Raoult *et al.*, 2004; Yoshida *et al.*, 2011). Il semble donc que parmi ces protéines, on puisse trouver des ADN polymérases X, en particulier des ADN polymérases β.

2.2.4.2.9 Cluster 9 : Les ADN polymérases X de Sar (séquence représentative : XP_001439407.1)

Le cluster 9 est celui qui présente entre autres les ADN polymérases X de *Paramecium tetraurelia* qui sont au centre des travaux de cette thèse, mais aussi d'autres séquences issues du clade Sar qui regroupe les Stramenopiles, Alveolata et Rhizaria, comme *Tetrahymena thermophila*. Ces ADN polymérases X semblent à mi-chemin entre les ADN polymérases λ et β de métazoaires.

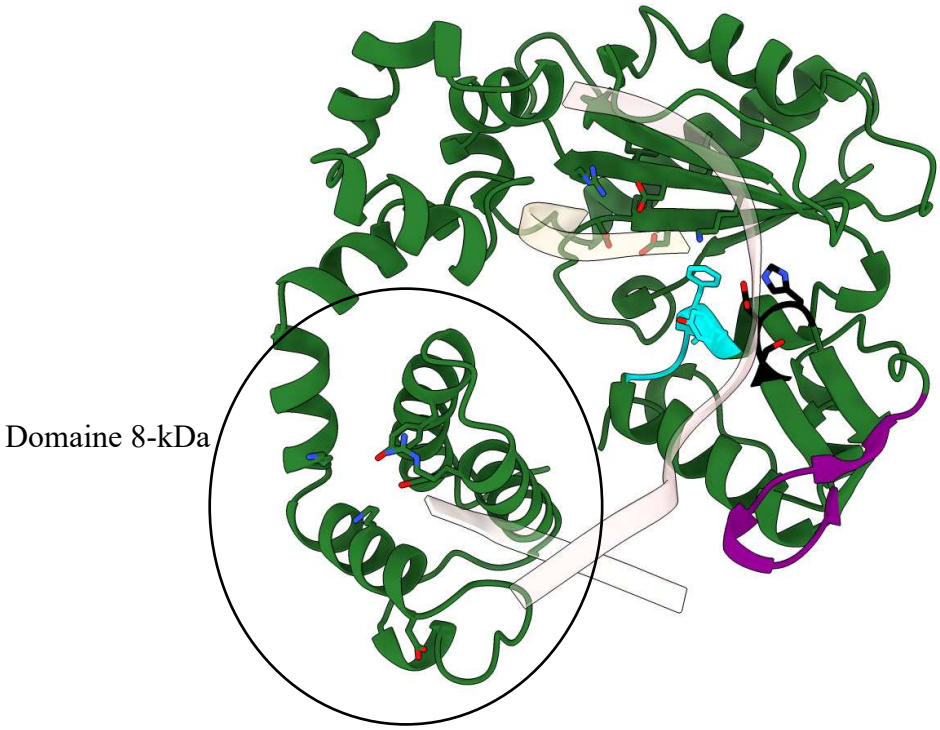


*Figure 70 : Structure de l'ADN polymérase X a de Paramecium tetraurelia (prédiction AlphaFold2). Entouré : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (l'ADN est celui de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; le steric gate est représenté en cyan ; le motif SD2 est représenté en **noir**.*

Tout d'abord, cette ADN polymérase X semble avoir dans son domaine 8-kDa tous les résidus impliqués dans la reconnaissance des groupements 5'P et à une activité dRP lyase. Elle ne présente pas de boucles 1 et 2 ni de motif SD1. Ses motifs catalytiques sont formés des résidus DMD et RIDLK. Elle semble avoir un *steric gate* (YFTGS) similaire aux ADN polymérases λ et β de métazoaires, et son motif SD2 est un SDH. De plus, elle porte une boucle 3. Enfin, comme

pour les autres clusters, ces observations peuvent être généralisées à l'ensemble des 18 séquences de ce cluster.

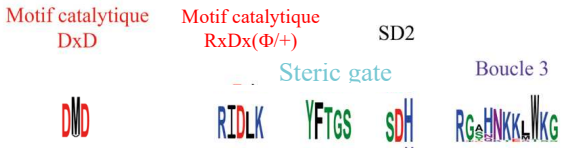


Figure 71 : Motifs d'intérêt des séquences du cluster 9. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre).

2.2.4.2.10 Cluster 10 : Les ADN polymérases IV (séquence représentative : AJP37443.1)

L'analyse phylogénétique a montré que le cluster 10 contient presque exclusivement des séquences de levures proches de *Saccharomyces cerevisiae*, annotées comme des ADN polymérases IV. Celles-ci ont souvent été considérées comme proches des ADN polymérases X métazoaires, mais il semble qu'il y ait plusieurs différences, surtout visibles en comparant les structures de ces polymérases. Tout d'abord, il semble que l'ADN polymérase IV étudiée ici ait dans son domaine 8-kDa certains des résidus permettant de reconnaitre un 5'P. Ses motifs catalytiques sont un DID et un RLDFF.

Domaine 8-kDa

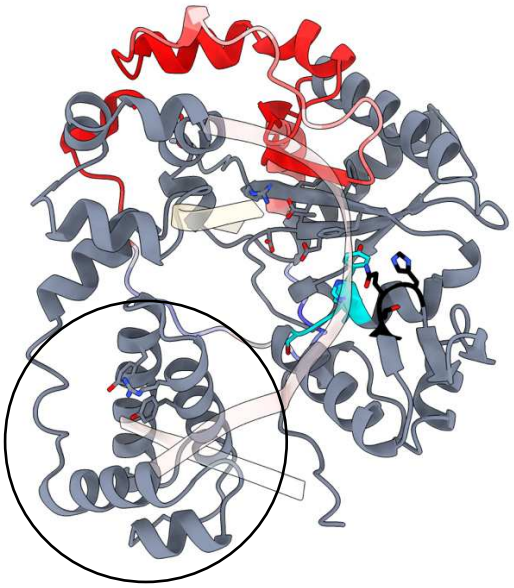


*Figure 72 : Structure de l'ADN polymérase IV de Saccharomyces cerevisiae (prédiction AlphaFold2). Entouré : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (l'ADN est celui de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; le steric gate est représenté en cyan ; le motif SD2 est représenté en **noir**. Les résidus correspondant à l'insertion « boucle 1 -SD1 – boucle 2) sont représentés selon leur score pLDDT obtenu lors de la prédiction de la structure par AlphaFold2 : en rouge les régions avec un score élevé ; en bleu les régions avec un score faible et possiblement mal prédites.*

Une particularité de cette ADN polymérases IV, en comparaison avec les autres ADN polymérases X, semble se situer au niveau de la boucle 1, du SD1 et de la boucle 2. En effet, cette zone, composée de résidus chargés positivement en N-terminal puis négativement en C-terminal (avec une suite de 7 glutamates), est particulièrement étendue chez cette ADN polymérase, bien plus que chez la Tdt ou l'ADN polymérase μ. On pourrait croire à des boucles 1 et 2 de très grande taille, mais la prédiction structurale d'AlphaFold2 indique plutôt que cet ensemble forme une seule grande structure qui ne ressemble pas aux boucles 1 et 2 des ADN polymérases μ et Tdt, ce qui est unique chez les ADN polymérases X. Cette structure est représentée dans la figure 72 selon le niveau de confiance du modèle (score pLDDT *(Jumper et al.*, 2021)) en rouge (score haut, très confiant) et bleu (score bas, peu confiant). La prédiction globale semble être de bonne qualité. On peut donc supposer que la structure proposée par ce modèle est assez fiable. Vu sa position dans la structure, il est possible que cette insertion puisse interagir avec l'ADN lorsqu'il se fixe dans le domaine polymérase.

Ce type de structure est unique chez les ADN polymérases X, et son rôle n'est pas connu à ce jour. Cependant, cela est cohérent avec la localisation du cluster 10 vis-à-vis des autres clusters d'ADN polymérases X : ce cluster est assez éloigné des autres, donc les séquences qui le forment doivent être assez différentes des autres ADN polymérases X, qui n'ont pas cette structure supplémentaire.

Enfin, l'ADN polymérase IV ne semble pas porter le *steric gate* des ADN polymérases λ et β, mais un motif HYTGS (assez éloigné aussi du GWTGS des ADN polymérases μ et Tdt), ce qui est cohérent avec son activité ADN/ARN polymérase (Bebenek *et al.*, 2005; McInnis *et al.*, 2002; Prasad *et al.*, 1993; Wilson and Lieber, 1999). Elle porte de plus un motif SD2 spécifique : TQH.

D'après la figure 73, ces observations peuvent être généralisées à l'ensemble du cluster 10, donc aux ADN polymérases IV en général, y compris l'insertion remplaçant les boucles 1 et 2 et le SD1.

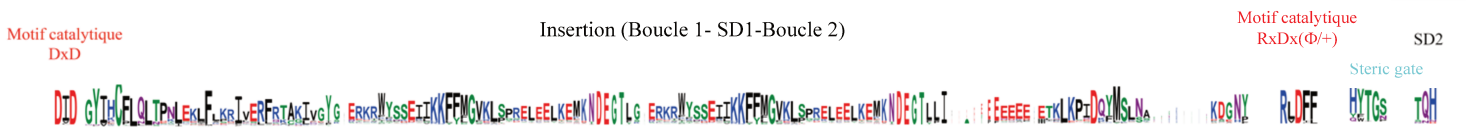


Figure 73 : Motifs d'intérêt des séquences du cluster 10. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre).

2.2.4.2.11 Clusters 11 et 12 : Les ADN polymérases X bactériennes

Les deux derniers clusters forment un *continuum* à part des clusters de séquences eucaryotes et virales. En effet, il semble que les ADN polymérases X bactériennes, probablement en raison de leur domaine PHP C-terminal, se classent à part. Cependant, comme indiqué en introduction, la littérature fait état de 2 catégories d' ADN polymérases X bactériennes : canoniques et non canoniques. Ici, ces ADN polymérases X sont représentées respectivement par des séquences de *Thermus thermophilus* et de *Deinococcus radiodurans*.

2.2.4.2.11.1 Les domaines PHP

La principale caractéristique communes aux séquences de ces deux clusters est une extension C-terminale, caractérisée dans la littérature comme un domaine PHP (Rodríguez *et al.*, 2019). Les domaines PHP de ces deux ADN polymérases X partagent en effet la grande majorité des résidus catalytiques, ce qui leur confère leurs activités nucléases, très conservées tout comme le repliement de ce domaine.

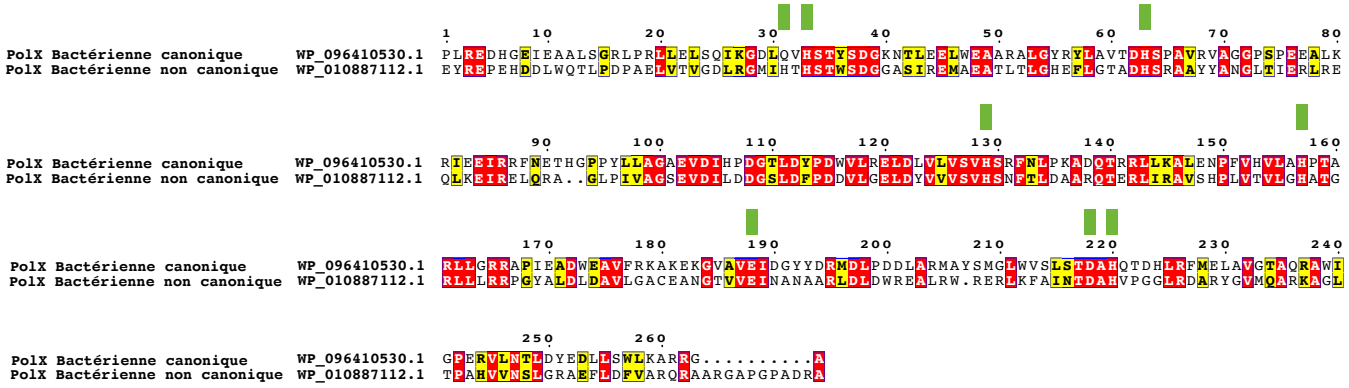


Figure 74 : Alignement des séquences des domaines C-terminaux des protéines représentatives des clusters 11 et 12, obtenu par PSI-Coffee. Les résidus catalytiques sont indiqués par des marqueurs verts.

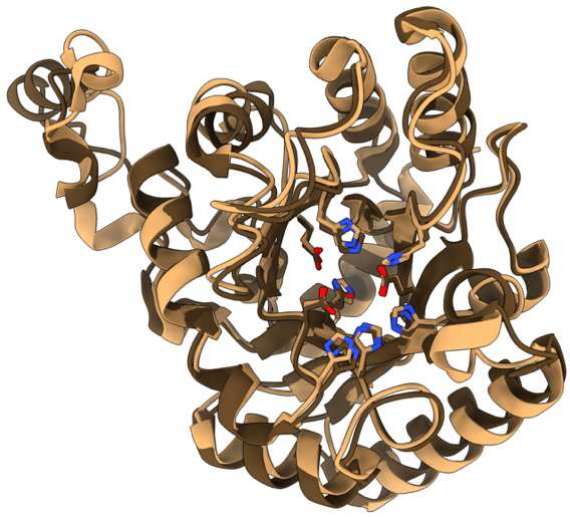


Figure 75 : Alignement structural des domaines C-terminaux des protéines représentatives des clusters 11 et 12, réalisé sur ChimeraX. Les résidus impliqués dans l'activité nucléase sont indiqués en bâtonnets.

Ces deux clusters semblent donc avoir en commun le domaine PHP, très conservé, ce qui est en accord avec la littérature.

2.2.4.2.11.2 Cluster 11 : Les ADN polymérases X bactériennes canoniques (séquence représentative : WP_096410530.1)

Ce cluster est le plus proche des séquences eucaryotes parmi les ADN polymérases X bactériennes, en particulier vis-à-vis du cluster des ADN polymérases β de métazoaires. En effet, le domaine catalytique de ces ADN polymérases X partage de nombreuses caractéristiques des ADN polymérases β, comme la majorité des résidus impliqués dans l'activité dRP lyase, la présence d'un résidu basique en position 5 du second motif catalytique (QVDLR), une partie des résidus du *steric gate* (YLTGS) et un motif SD2 de séquence SEY. Son premier motif catalytique est un DLD, et cette séquence ne porte pas de boucle 2.

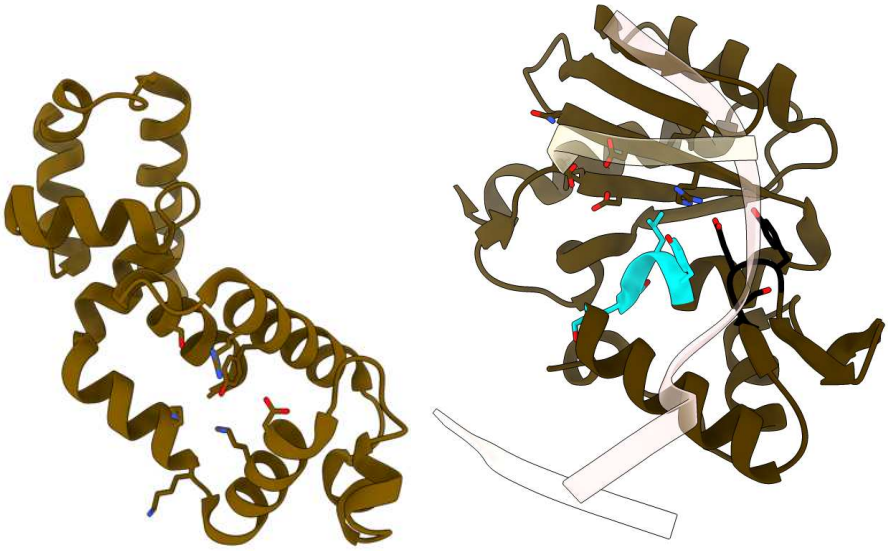


*Figure 76 : Structure de l'ADN polymérase X canonique de Thermus thermophilus (code PDB : 3au2). À gauche : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (avec l'ADN de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs catalytiques DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; le steric gate est représenté en cyan ; le motif SD2 est représenté en **noir**.*

Ces observations sont généralisables à toutes les séquences de ce cluster, qui sont donc des ADN polymérases X canoniques puisqu'elles portent tous les motifs leur permettant d'avoir une activité d'ADN polymérase. Il existe cependant une certaine variabilité sur deux points : le *steric gate*, qui semble avoir en majorité un motif YFTGS et non YLTGS comme dans l'exemple choisi ci-dessus ; et le motif SD2 qui semble pouvoir varier entre SEY et NEY (ce dernier étant caractéristique des ADN polymérases β).

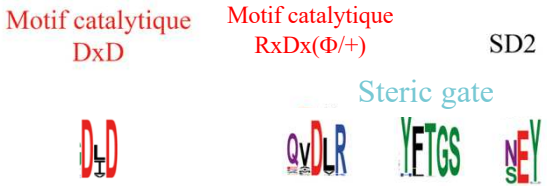


Figure 77 : Motifs d'intérêt des séquences du cluster 11. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre).

2.2.4.2.11.3 Cluster 12 : les ADN polymérases X bactériennes non canoniques (séquence représentative : WP_010887112.1)

Ce dernier cluster regroupe lui aussi des séquences bactériennes, mais est plus éloigné des ADN polymérases X eucaryotes. Cependant, les séquences de ce cluster présentent au niveau du domaine catalytique de nombreuses différences avec toutes les séquences présentées ci-dessus, y compris les bactériennes. Tout d'abord, le domaine 8-kDa ne présente pas tous les résidus impliqués dans la reconnaissance du 5'P, et le domaine polymérase ne présente pas les caractéristiques conférant une activité polymérase. En effet, cette polymérase n'a pas les résidus catalytiques habituels : les motifs catalytiques sont ici composés des résidus AAE et PVEIA. Elle ne présente pas non plus de *steric gate* (LLRSG), son motif SD2 est formé des résidus TAG, et elle ne porte pas de boucle 1 ou 2 ni de motif SD1. Comme indiqué dans la littérature, cette dégénérescence semble indiquer une perte de l'activité ADN polymérase (Prostova *et al.*, 2022).

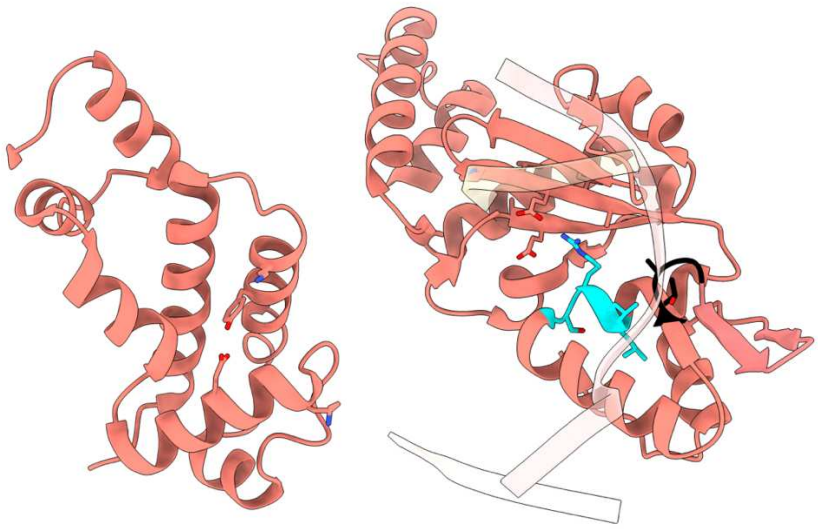


Figure 78 : Structure de l'ADN polymérase X non canonique de Deinococcus radiodurans (code PDB : 2w9m). À gauche : domaine 8 kDa, pour lequel les équivalents des résidus impliqués dans l'activité dRP lyase et la reconnaissance des groupements 5'P sont indiqués en batônnets. À droite : domaine polymérase (avec l'ADN de la structure de l'ADN polymérase λ 1rzt). Les différents points d'intérêt sont indiqués comme suit : les résidus des motifs DxD et RxDx(Φ/+) sont indiqués en bâtonnets, de la couleur du modèle ; l'équivalent du steric gate est représenté en cyan ; le motif SD2 est représenté en ***noir****.*

Ces observations semblent généralisables aux autres séquences de ce cluster, mais contrairement aux autres clusters, c'est ici parce que tous les motifs d'intérêt dans le domaine polymérase présentent une très haute variabilité : aucun résidu n'est conservé, même dans le motif SD2. Ces ADN polymérases X ne présentent donc que les caractéristiques attendues pour

une activité nucléase, au sein de leur domaine PHP, alors que leur domaine polymérase semble dégénéré.

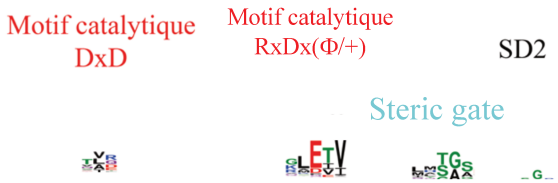


Figure 79 : Motifs d'intérêt des séquences du cluster 12. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre).

2.3 Discussion

À travers l'analyse de tous ces clusters, nous avons pu voir que la famille des ADN polymérases X ne se limite pas aux ADN polymérases λ, μ, β et Tdt de métazoaires, desquelles les autres séquences découleraient. En effet, jusqu'ici les ADN polymérases X humaines ont servi de référence pour étudier celles des autres organismes, comme les *Fungi* ou les *Viridiplantae*. Or, pour ces exemples précis, nous avons pu voir que classifier les ADN polymérases X de ces organismes de cette façon peut se montrer insuffisant, puisqu'elles ont parfois des caractéristiques les différenciant clairement des enzymes des métazoaires.

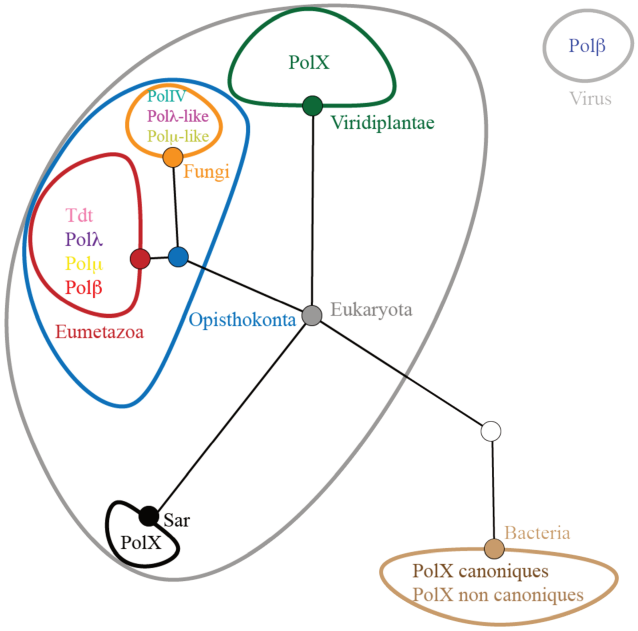


Figure 80 : Vue d'ensemble des groupes d'ADN polymérases X étudiées ici, séparées selon leur origine au sein du vivant

Pour classifier toutes les ADN polymérases X, la méthode utilisée ici a été de rechercher des motifs dans leurs séquences (motifs catalytiques, boucles 1, 2 et 3, *steric gate*, motifs SD1 et 2, résidus liés à une activité dRP lyase), ainsi que des domaines particuliers (BRCT et PHP). Cela a permis de définir un ensemble de motifs caractéristiques de chaque groupe. En premier lieu, la présence d'un domaine BRCT permet d'exclure une appartenance aux groupes « β like » (ADN polymérases β métazoaires, ADN polymérases β virales et ADN polymérases X bactériennes), et suppose un lien avec le système NHEJ. Concernant le premier motif catalytique (DxD), il est toujours présent sauf chez les ADN polymérases X bactériennes non canoniques, qui n'ont pas d'activité ADN polymérase (Prostova *et al.*, 2022). Le second motif catalytique (RxDx(Φ/+)) permet de séparer plusieurs classes d'ADN polymérases X. Les ADN polymérases X bactériennes n'ont pas d'arginine en position 1 de ce motif. Le résidu catalytique, en position 3, est quant à lui conservé puisqu'il s'agit d'un aspartate (ou d'un glutamate conservé chez les ADN polymérases X bactériennes non canoniques). Le résidu qui permet le mieux de distinguer les différentes classes, d'un point de vue phylogénétique et fonctionnel, est le dernier : chez toutes les ADN polymérases β, il s'agit d'une arginine (ADN polymérases β métazoaires, ADN polymérases β virales et ADN polymérases X bactériennes). La grande majorité des classes restantes portent un résidu hydrophobe à cette position, à l'exception des ADN polymérases X de plantes et de Sar, qui portent une lysine. Les seules ADN polymérases X à avoir une boucle 1 et un motif SD1 sont les ADN polymérases μ et Tdt de métazoaires, et les ADN polymérases μ-like de *Fungi*, ce qui est cohérent avec les informations connues sur l'ADN polymérase μ et la Tdt chez les eucaryotes. La boucle 2 est présente dans davantage de groupes, puisqu'en plus des 3 précédents, elle est aussi présente chez les ADN polymérases λ et β de métazoaires et les ADN polymérases λ like de Fungi. Les ADN polymérases IV de levures portent un ensemble « boucle 1- boucle 2 » qui semble fortement se différencier des ADN polymérases X citées ci-dessus, et dont on ne connait pas encore la fonction. Un des motifs permettant le mieux de distinguer les ADN polymérases X est le *steric gate*. En effet, il permet de faire la distinction entre les ADN polymérases « β/λ like » (n'incorporant pas de ribonucléotides) et « μ/Tdt like » (qui peuvent incorporer des NTPs). Les ADN polymérases X bactériennes canoniques, ADN polymérases X de virus, de Sar, les ADN polymérases λ like de *Fungi* ainsi que les ADN polymérases β et λ de métazoaires ont un *steric gate* de type « YFTGS », et ne peuvent donc pas incorporer de NTPs. Les ADN polymérases X de plantes, les ADN polymérases μ like de Fungi, les ADN polymérases

μ et Tdt de métazoaires ont quant à elles un motif de type « A/G-WTG », et peuvent incorporer des ribonucléotides. Là encore, les ADN polymérases IV font figure d'exception, avec un motif HYTGS, qui ne bloque pas l'incorporation des ribonucléotides.

Le motif SD2 est lui aussi un bon indice d'appartenance à une classe pour les ADN polymérases X, puisque chaque motif est spécifique de son groupe. Toutes les ADN polymérases β et « β like » ont un motif NEY (ou SEY), et la présence d'un résidu chargé négativement en position 2 indique une appartenance au groupe « β/λ like » : les ADN polymérases X de plantes ont un motif DDT, celles de Sar ont un motif SDH, et les ADN polymérases λ de métazoaires ont un motif SEH (il existe cependant l'exception des ADN polymérases λ like de Fungi, qui ont un motif NQR). Tous les autres groupes d'ADN polymérases X ont des motifs SD2 variables.

La présence d'une boucle 3 est quant à elle spécifique des ADN polymérases X « λ like », et est donc présente chez les ADN polymérases X de plantes et de ciliés, ainsi que chez les ADN polymérases λ de métazoaires et les ADN polymérases λ like de Fungi. Enfin, seules les ADN polymérases X bactériennes ont un domaine PHP C-terminal, très conservé, et qui leur confère leurs activités nucléases. Elles peuvent cependant être subdivisées, entre polymérases canoniques et non canoniques ; ces dernières ayant apparemment perdu toutes les caractéristiques des ADN polymérases X (motifs catalytiques, *steric gate*, …).

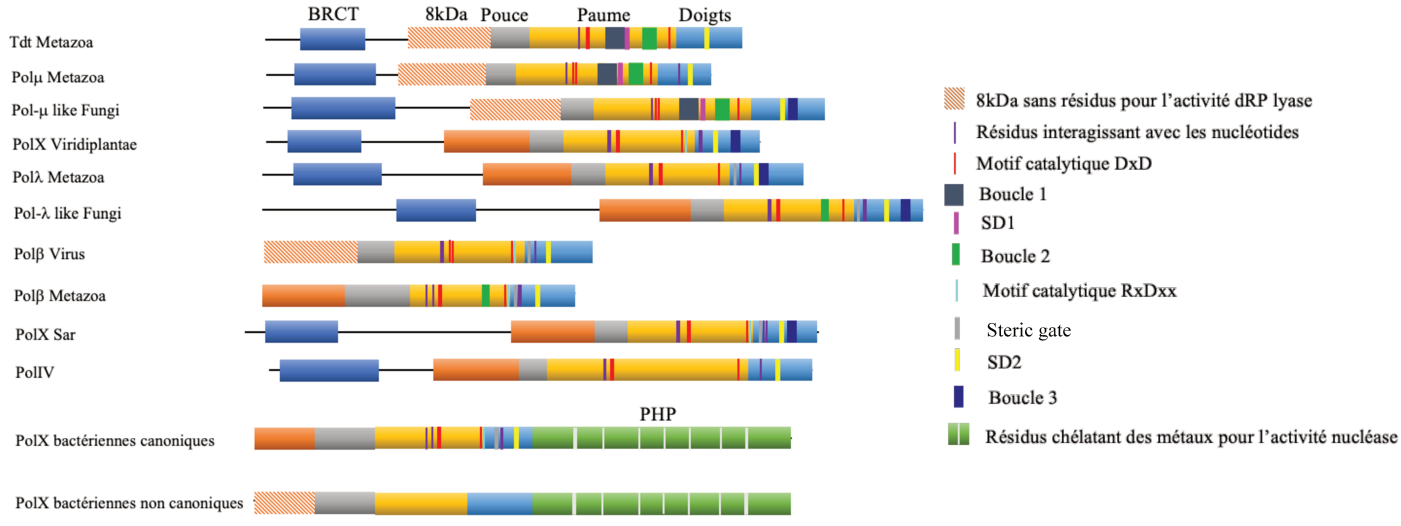


Figure 81 : Schéma des PolX étudiées dans chaque cluster, indiquant leurs domaines, indiqués en haut, et leurs caractéristiques, expliquées dans la légende à droite.

L'ensemble des motifs et domaines recherchés ici ont donc permis de construire une classification des ADN polymérases X plus compréhensive et large qu'auparavant, tout en respectant la classification déjà connue. Cependant, il existe une grande variabilité dans certains groupes, en particulier chez les *Fungi*, qui pourrait être étudiée de façon approfondie dans des travaux dédiés. En effet, les deux clusters d'ADN polymérases X de *Fungi* étudiés ici pourraient probablement être divisés en deux, en étudiant en profondeur leurs différences. De plus, contrairement aux travaux sur les ADN polymérases de la famille A publiés récemment et présentés dans un chapitre additionnel de cette thèse, les séquences utilisées pour le clustering présenté ici ont été obtenues en utilisant la séquence de l'ADN polymérases Xa de *P. tetraurelia* et une recherche par BLAST pour obtenir des séquences proches. Pour les ADN polymérases A, l'ensemble des séquences annotées comme ADN polymérases A avec HMMER ont été utilisées, ce qui semble plus exhaustif. Cette même méthode pourrait être utilisée pour confirmer les résultats présentés ici, si l'algorithme HMMER est capable de reconnaitre des motifs de ADN polymérases X, y compris chez des séquences distantes comme les ADN polymérases X bactériennes non canoniques.

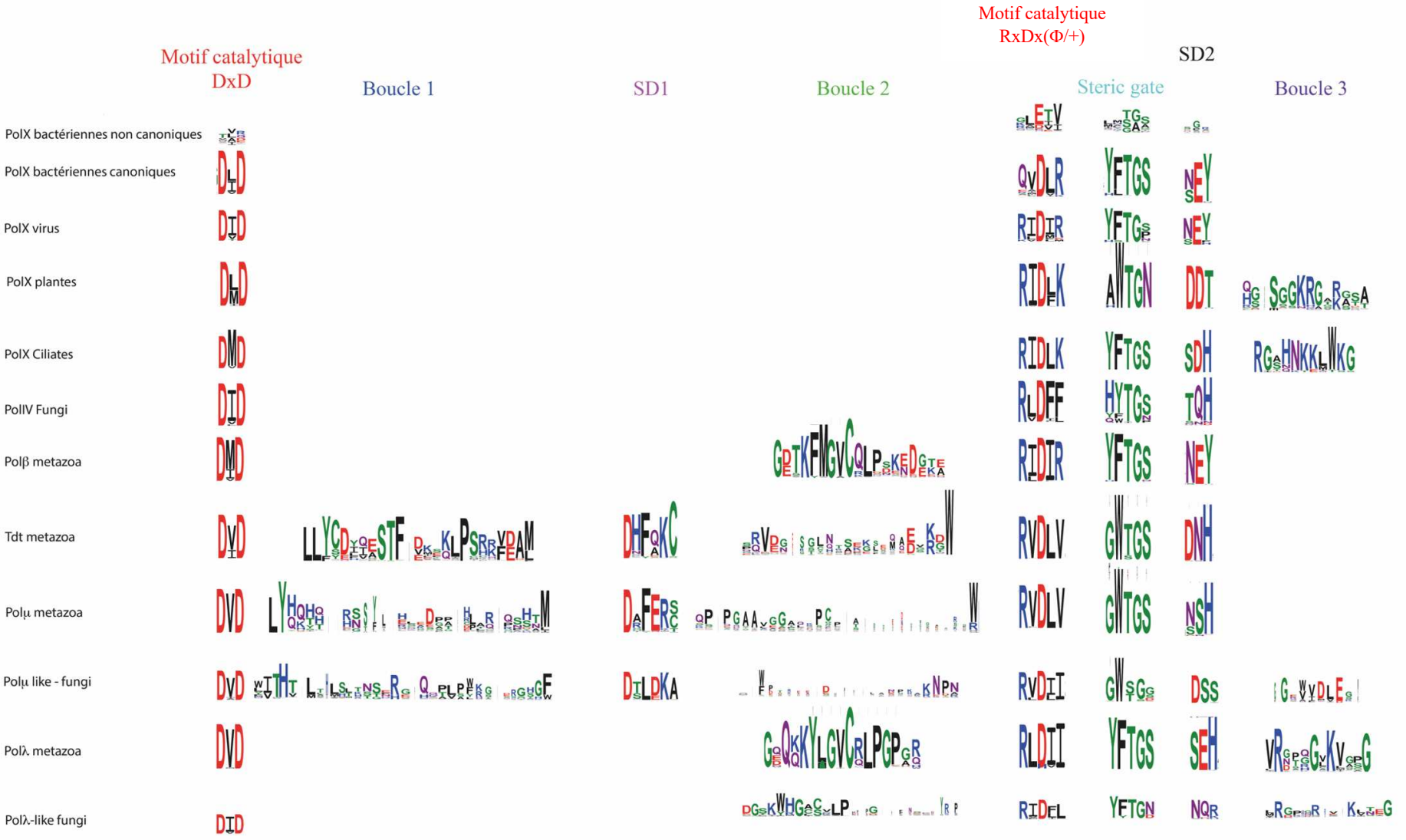


Figure 82 : Motifs d'intérêt des séquences de tous les groupes d'ADN polymérases X. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre)

3 Les ADN polymérases X de *Paramecium tetraurelia*

L'ensemble des travaux présentés jusqu'ici ont permis d'actualiser la classification des ADN polymérases X, en étendant cette classification à l'ensemble des séquences d'ADN polymérases X connues à ce jour. Cette partie se focalise sur les ADN polymérases X impliquées dans la réparation des cassures double brins programmées chez *P. tetraurelia*.

Des travaux récents du Dr Julien Bischerour à l'I2BC (non encore publiés) ont montré que quatre protéines annotées comme des ADN polymérases X étaient présentes chez *P. tetraurelia*, et impliqués dans la réparation des CDB programmées. Ces 4 ADN polymérases X ont été nommées ADN polymérases Xa (ID ParameciumDB : PTET.51.1.P0210235), b (PTET.51.1.P0360066), c (PTET.51.1.P0460033) et d (PTET.51.1.P1010039). Les ADN polymérases Xa et b partagent 90% d'identité de séquence, et les ADN polymérases Xc et d partagent 84%, et ces deux sous-groupes partagent 71% d'identité de séquence.

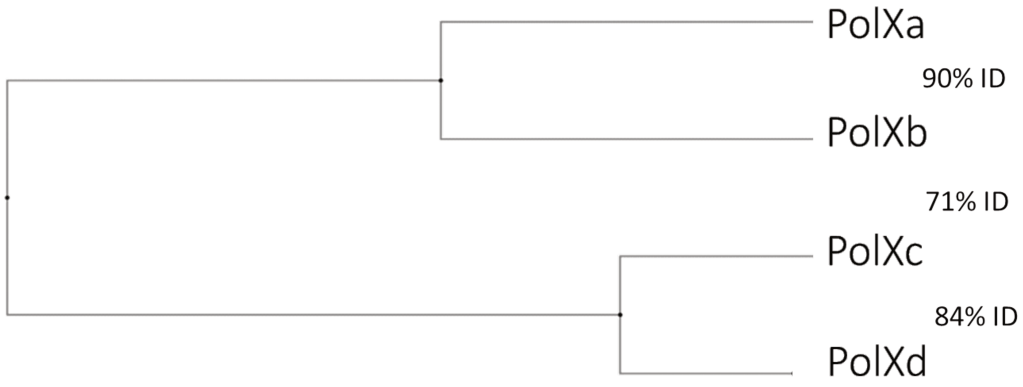


Figure 83 : Arbre généré par BLAST montrant les liens entre les 4 ADN polymérases X de Paramecium tetraurelia, et leurs pourcentages d'identité de séquences.

Connaissant l'histoire évolutive de *Paramecium tetraurelia*, on peut dire que ces quatre séquences proches sont des ohnologues : elles sont issues de gènes ayant évolué indépendamment, à partir d'un gène unique. Cependant, ces quatre ADN polymérases X ont des différences, parmi lesquelles leurs profils d'expression. En effet, comme indiqué en figure 84, l'ADN polymérase Xa est surexprimée au cours du cycle de vie de *P. tetraurelia*, en particulier lors des étapes de fragmentation du macronoyau de la cellule mère et de développement du macronoyau de la cellule fille. L'ADN polymérase Xb est elle aussi légèrement surexprimée, dès la méiose. Les ADN polymérases Xc et d ne sont quant à elle pas surexprimées.

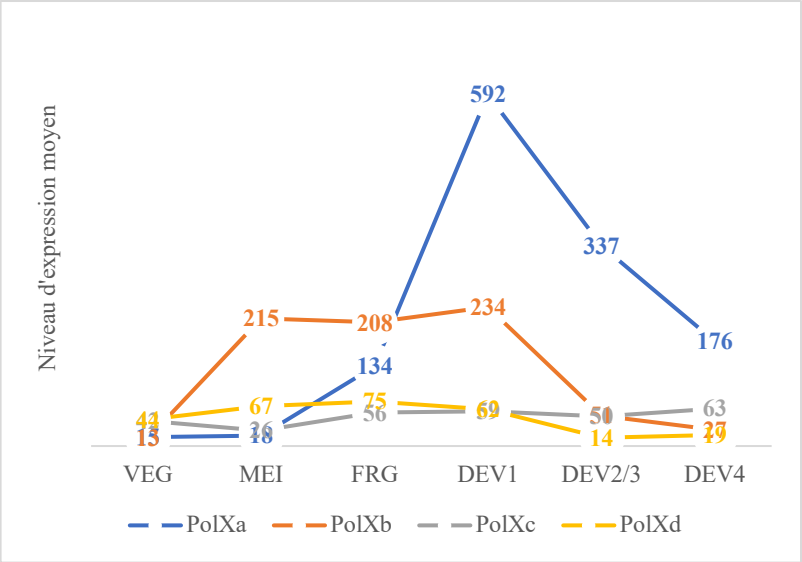


Figure 84 : Niveau d'expression moyen des 4 ADN polymérases X de P. tetraurelia au cours du cycle de vie de la cellule, mesuré par RNA-seq. VEG : cycle végétatif ; MEI : méiose des MIC, FRG : fragmentation du MAC, DEV1/2/3/4 : développement du MAC. D'après Arnaiz O et al. BMC Genomics. 2017.

L'objectif des travaux décrits dans cette partie a été d'analyser en détails les séquences de ces ADN polymérases, afin de dégager des hypothèses pouvant expliquer leur fidélité.

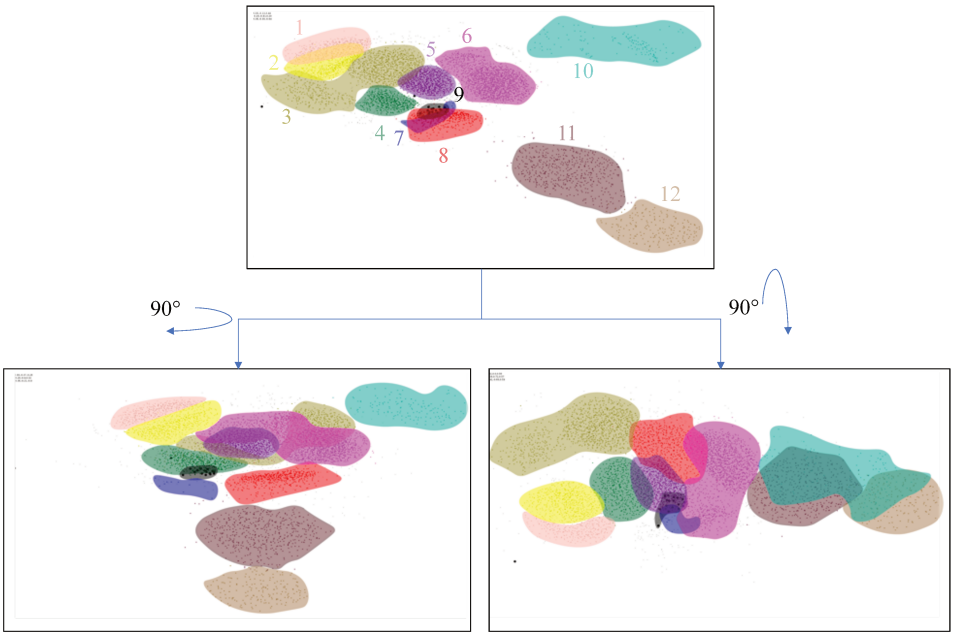
3.1 Matériel et méthodes

La majorité des résultats analysés ici sont tirés de ceux de la partie précédente sur la classification des ADN polymérases X. Pour comparer plus en profondeur les ADN polymérases X de *P. tetraurelia* avec des ADN polymérases X connues, un alignement de séquences a été réalisé grâce à PSI-Coffee, incluant de nouvelles séquences. Les structures de leurs domaines catalytiques ont ensuite été comparées sur ChimeraX (Prédiction AlphaFold2 de l'ADN polymérase X a présentée dans la partie précédente ; codes PDB : ADN polymérase λ = 1rzt ; ADN polymérase β =5tb8). Une comparaison centrée sur le site actif de l'ADN polymérase β (code PDB : 1bpy) et l'ADN polymérase X a ensuite été réalisée. Cette fois, c'est une prédiction SwissModel de l'ADN polymérase X a qui a été utilisée, car les rotamères proposés pour les chaines latérales des résidus d'intérêt semblaient plus cohérents.

3.2 Résultats

3.2.1 Les ADN polymérases de *Paramecium* forment un groupe à part au sein des ADN polymérases X

Comme indiqué dans la partie précédente, les ADN polymérases X provenant du clade appelé Sar (incluant *P. tetraurelia*, les ciliés appartenant à la famille des *Alveolata*.) sont un groupe à part. Ce groupe, en noir sur la figure 85, se place à proximité des ADN polymérases β et λ des métazoaires, respectivement indiquées en rouge (cluster 8) et en violet (cluster 5).



N° du cluster	*Groupe d'ADN polymérases X*
1	Tdt de métazoaires
2	ADN polymérases μ de métazoaires
3	ADN polymérases μ like de Fungi
4	ADN polymérases X de Viridiplantae
5	ADN polymérases λ de métazoaires
6	ADN polymérases λ like de Fungi
7	ADN polymérases β de virus
8	ADN polymérases β de métazoaires
9	ADN polymérases X de Sar
10	ADN polymérases IV de levures
11	ADN polymérases X canoniques de bactéries
12	ADN polymérases X non canoniques de bactéries

Figure 85 : Clustering des séquences d'ADN polymérase X. Un ensemble de 7250 séquences de PolXa été obtenu d'après la base de données du NCBI en janvier 2023. Une distribution 3D en clusters a été générée par CLANS en utilisant les score de similarités entre chaque paire de séquence comme valeurs d'attraction.

Comme l'indique l'alignement de séquences montré en figure 86, ce sous-groupe d'ADN polymérases X se caractérise par la présence en N-terminal d'un domaine BRCT, suivi d'un long linker dont la fonction et le repliement ne sont pas connus, précédant le domaine 8 kDa, qui présente tous les résidus permettant une activité dRP lyase et la reconnaissance de groupements 5'P.

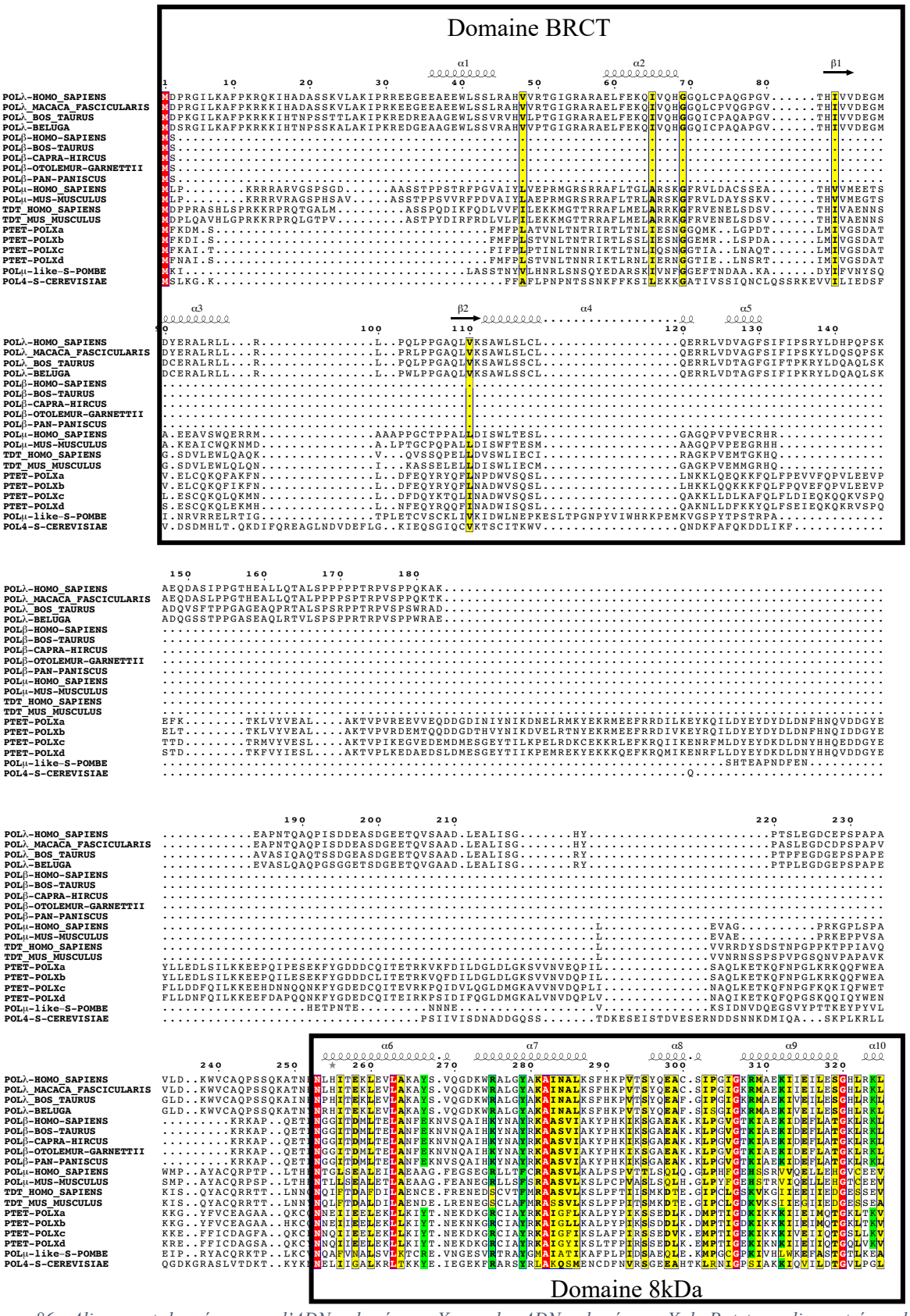


Figure 86 : Alignement des séquences d'ADN polymérases X avec les ADN polymérases X de P. tetraurelia centré sur les domaines BRCT et 8kDa, obtenu par PSI-Coffee. Les résidus impliqués dans la reconnaissance des groupements 5'P (et activité dRP lyase)sont indiqués en ***vert*** *. Les structures secondaires de l'ADN polymérase λ humaine (PDB 2JM5 et 7M43) sont indiquées au-dessus de l'alignement.*

Le domaine catalytique porte quant à lui un *steric gate*, un motif SD2 unique (SDH) et une boucle 3, en plus des résidus catalytiques, comme l'indique la figure 87.

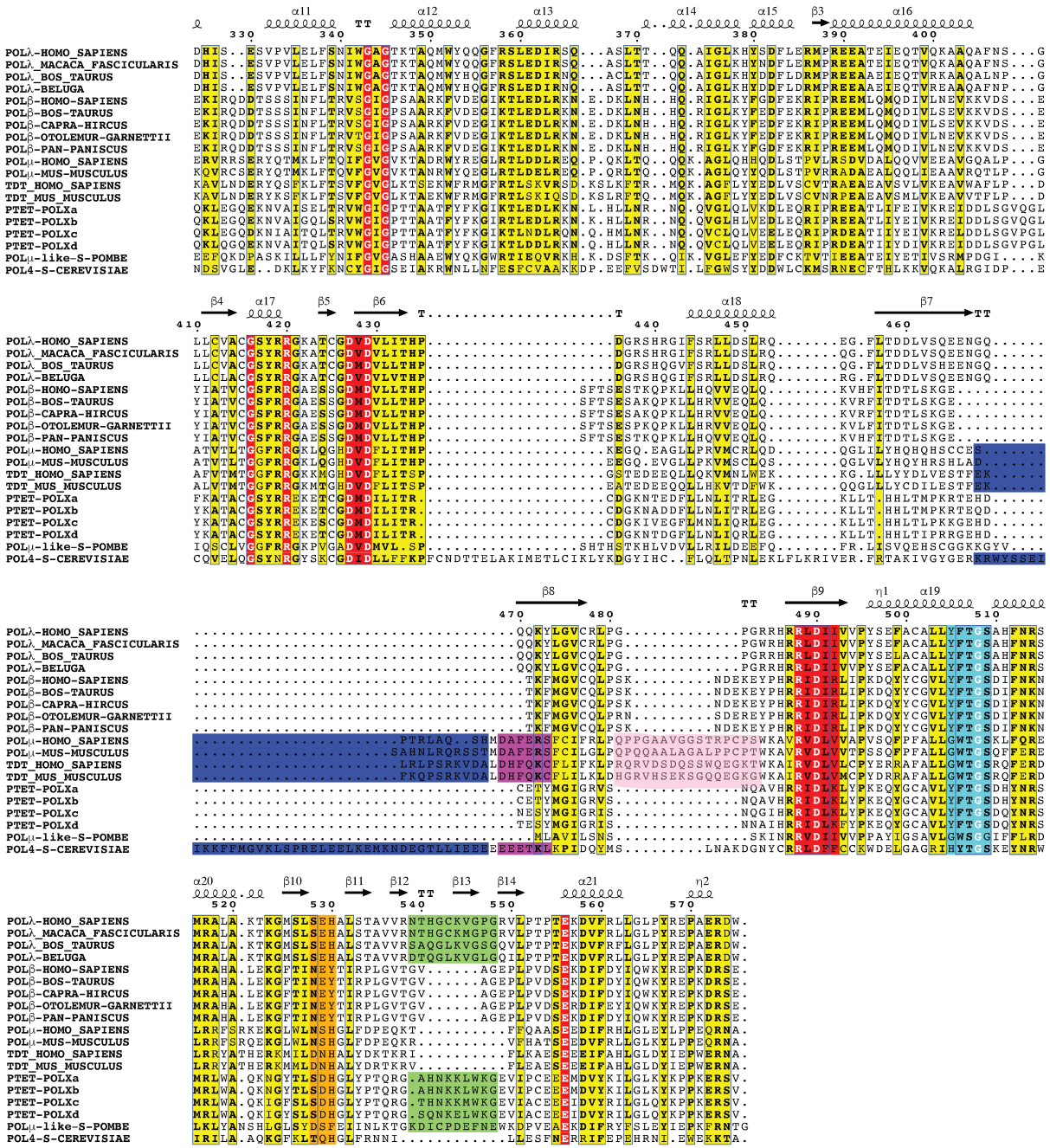


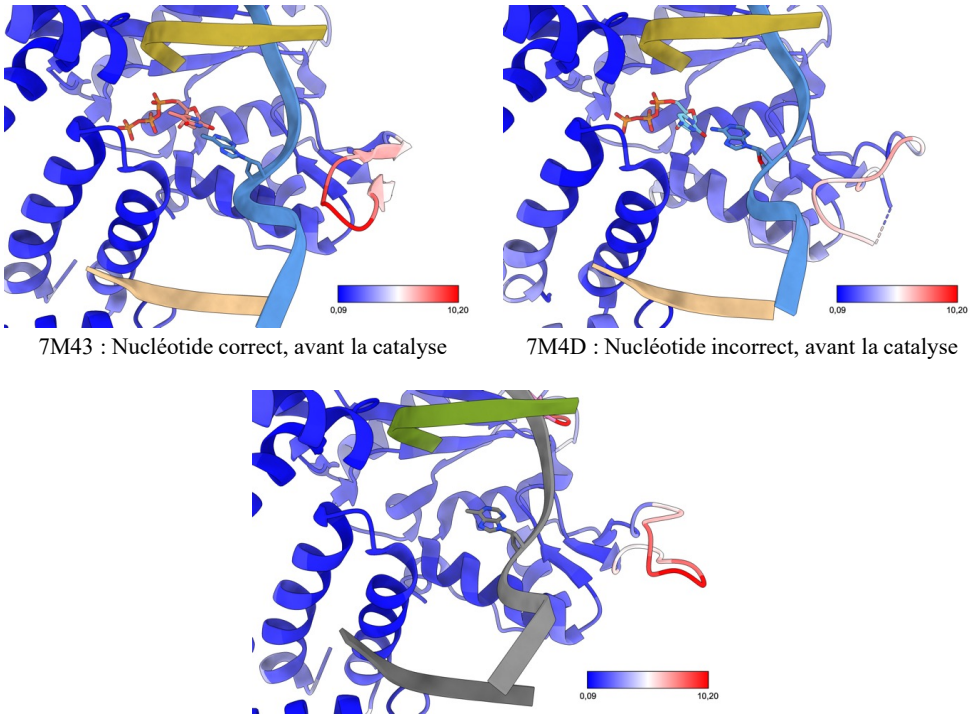
*Figure 87 : d'ADN polymérases X avec les ADN polymérases X de P. tetraurelia centré sur le domaine polymérase, obtenu par PSI-Coffee. Les éléments caractéristiques sont indiqués en couleurs : motifs catalytiques DxD et RxDx(Φ/+) en **rouge** ; boucle 1 en **bleu**, motif SD1 en **rose** ; boucle 2 en rose pâle ; steric gate en **cyan**, motif SD2 en **orange** et boucle 3 en **vert clair** . Les structures secondaires de la Polλ humaine (PDB 2JM5 pour le domaine BRCT et 7M43 pour le domaine catalytique) sont indiquées au-dessus de l'alignement. Pour annoter les boucles 1 et 2 et le motif SD1, la Tdt de souris a été utilisée comme référence. Pour la boucle 3, c'est l'ADN polymérase λ humaine qui a été utilisée comme référence.*

Les alignements présentés en figures 86 et 87 seront utilisés comme références dans les paragraphes suivants.

3.2.2 Les ADN polymérases X de *Paramecium tetraurelia* sont proches des ADN polymérases λ de métazoaires, et pourraient partager avec elles un mécanisme de fidélité

A première vue, ces ADN polymérases X semblent être similaires aux ADN polymérases λ : le principal argument est la présence du domaine BRCT et du *steric gate*, mais étant donnée la ressemblance entre les domaines polymérases des ADN polymérases λ et β (comme le *steric gate* ou la présence des résidus impliqués dans la reconnaissance des groupements 5' phosphate et l'activité dRP lyase), il faut s'attarder sur les détails des séquences. Par exemple, le motif SD2 des ADN polymérases X de *Paramecium* (SDH) est plus proche de celui de l'ADN polymérase λ (SEH) que de l'ADN polymérase β (NEY). De plus, la présence d'une boucle 3 indique une similitude avec les ADN polymérases λ.

On sait encore peu de choses au sujet de cette boucle 3. Dans un article publié en 2022 (Jamsen *et al.*, 2022), en cherchant à observer l'incorporation de nucléotides incorrects par l'ADN polymérase λ, les auteurs ont observé que l'ADN se place incorrectement dans le site actif en présence d'un nucléotide incorrect. En effet, le nucléotide *template* est situé 2,5 Å en aval de son emplacement optimal, et sa base est tournée de 22°, ce qui ne permet pas l'hybridation avec un nucléotide entrant. En cherchant à comprendre cette observation, ils se sont rendu compte que la boucle 3 avait un placement très différent selon la nature du nucléotide entrant. Si le nucléotide entrant est correct, la boucle 3 est au contact de l'ADN tout au long de la catalyse (comme le montrent les structures PDB 7M43 à 7M4C), et peut stabiliser son positionnement au sein du domaine catalytique. En revanche, si le nucléotide entrant est incorrect, cette boucle est déstructurée pendant la catalyse et n'est pas au contact de l'ADN (PDB 7M4D à 7M4K): celui-ci se place donc mal, ce qui rend plus difficile l'incorporation du nucléotide incorrect. Ce mécanisme est encore mal compris, mais il pourrait être lié à la fidélité de l'ADN polymérase λ, qui est comparable à celle de l'ADN polymérase β mais apparemment dépendante d'une séquence riche en sérines et prolines situées en N-terminal du domaine catalytique (Bebenek *et al.*, 2003; Fiala *et al.*, 2006; García-Díaz *et al.*, 2002; Ramadan *et al.*, 2002; Yamtich and Sweasy, 2010).



7M43 : Nucléotide correct, avant la catalyse

7M4D : Nucléotide incorrect, avant la catalyse

1XSL : Pas de nucléotide

Figure 88 : La nature du nucléotide entrant a un impact sur le placement de l'ADN et de la boucle 3 chez l'ADN polymérase λ humaine. L'ADN polymérase est colorée selon son RMSD après superposition des trois structures. En haut à gauche : vue centrée sur le nucléotide template avant une incorporation correcte. Ici le nucléotide entrant et l'ADN sont sur le même plan et leur orientation peut permettre l'incorporation. La boucle 3 est au contact de l'ADN template et stabilise son placement dans le domaine catalytique. En haut à droite : vue centrée sur le nucléotide template avant une incorporation incorrecte. Ici, le nucléotide template est décalé de 2,5 Å en aval de la cassure et tourné de 22° : il n'est plus dans un plan permettant l'hybridation avec le nucléotide entrant. Ici, la boucle 3 est dans une position différente, à distance de l'ADN, et est flexible. En bas : vue centrée sur le nucléotide template au sein du site actif de l'ADN polymérase λ, en absence de nucléotide entrant. Le nucléotide template est placé comme en présence d'un nucléotide incorrect, et la boucle 3 n'est pas au contact de l'ADN.

Cette boucle 3 est conservée dans peu de groupes d'ADN polymérases X : les ADN polymérases λ de métazoaires, λ like de champignons, les ADN polymérases X de *Viridiplantae* et de Sar. Elle semble d'ailleurs être davantage chargée positivement chez *Paramecium*. Par conséquent, il semble possible que les ADN polymérases X de *Paramecium* partagent ce mécanisme de l'ADN polymérase λ. Si c'est le cas, il pourrait être à l'origine de la fidélité des ADN polymérases X de *Paramecium*.

3.2.3 Les ADN polymérases X de *P. tetraurelia* pourraient utiliser un mécanisme similaire à celui de l'ADN polymérase β pour améliorer leur fidélité

3.2.3.1 L'induced-fit mechanism de l'ADN polymérase β

Dans son motif RxDx(Φ/+), l'ADN polymérase β porte une arginine en position 5 (RIDIR). Cette arginine est à l'origine d'un mécanisme à l'origine de la fidélité de cette enzyme, meilleure que celle des autres ADN polymérases de la famille X (Beard *et al.*, 2014). En effet, elle a pour particularité de changer d'état conformationnel au cours de la catalyse : en absence

d'ADN et d'un dNTP correct dans le site actif, cette arginine (R258) stabilise une conformation inactive et ouverte de l'enzyme, en étant liée à un des aspartates catalytiques (D192) par un pont salin, ce qui empêche ce dernier de coordonner les ions magnésium du site actif, donc la catalyse ne peut pas avoir lieu. Mais si l'ADN et le bon dNTP se fixent dans le site actif, la liaison entre R258 et D192 est rompue par le basculement de la chaine latérale de la phénylalanine du *steric gate* (F272). R258 est alors stabilisée par les résidus du motif SD2 (NEY), D192 peut alors coordonner les ions magnésium, et la catalyse peut avoir lieu. Ces changements conformationnels sont des points de contrôle : si l'ADN et le dNTP ne sont pas correctement placés, la polymérase ne se ferme pas et ne catalyse pas la réaction.

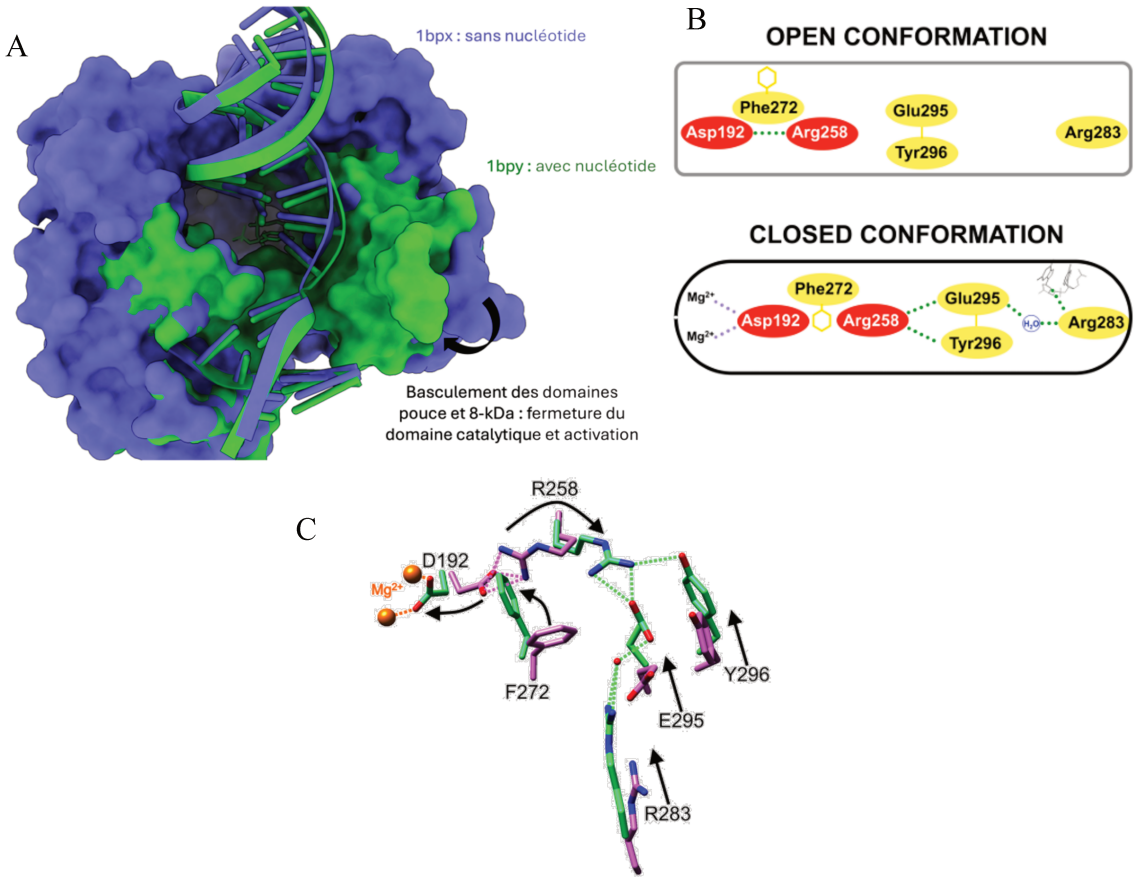
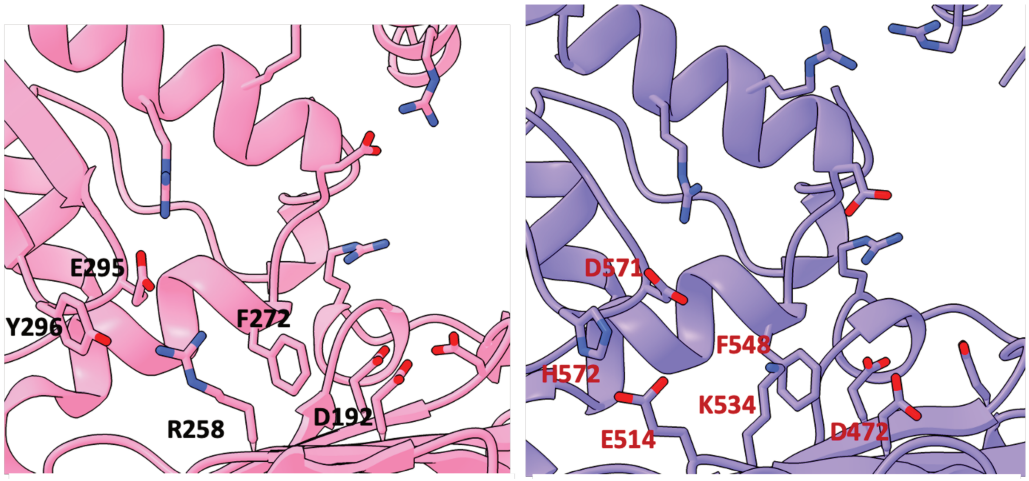


Figure 89 : Changements conformationnels induits par la fixation des substrats dans le site catalytique de l'ADN polymérase β. (A) Fermeture du domaine catalytique lors de la fixation du nucléotide entrant. En absence de nucléotide entrant (PDB 1BPX en bleu) l'ADN polymérase β est sous forme ouverte et inactive ; lors de sa fixation, elle se ferme et passe sous forme active (PDB 1BPY en vert) (B) Schéma du mécanisme de transition ouvert/ fermé. En conformation ouverte, R283 n'interagit pas avec les autres résidus, mais en conformation fermée, elle interagit avec la base modèle, le nucléotide en amont et Q295. Par conséquent, le sous domaine N est repositionné vers le site actif par une série d'interactions impliquant R283 et D192 qui coordonnent les ions Mg2+ du site actif. Cela s'accompagne également d'interactions de Q295/Y296 avec R258 forme fermées. F272 interfère transitoirement avec les interactions entre D192 et R258, permettant son interaction avec Q295/Y296. (C) Ces mêmes interactions sont représentées d'un point de vue structural, avec en rose la conformation ouverte et en vert la conformation fermée. Figure d'après Biochemistry 2014, 53, 17, 2768–2780(Beard and Wilson, 2014) et Journal of Biological Chemistry 2014, 289, 45, 21411-31422(Beard et al., 2014).

3.2.3.2 *Un mécanisme possiblement partagé par les ADN polymérases X de Paramecium*

En comparant la structure connue de l'ADN polymérase β humaine avec les prédictions de structure des ADN polymérases X de *P. tetraurelia*, il est apparu qu'un mécanisme équivalent pourrait exister chez *Paramecium*. En effet, ces ADN polymérases présentent des équivalents de tous les résidus impliqués dans le mécanisme *induced-fit* de l'ADN polymérase β: l'aspartate catalytique est conservé, tout comme la phénylalanine du *steric gate* (F272 chez l'ADN polymérase β, F548 chez les ADN polymérases X de *P. tetraurelia*) ; l'arginine 258 de l'ADN polymérase β est remplacée par une lysine en position 534, et les résidus NEY du motif SD2 sont remplacés par des résidus SDH, pouvant jouer un rôle de stabilisation de la lysine.



ADN polymérase β
(PDB 1BPY)

ADN polymérase X a
(Prédiction SwissModel)

Figure 90 : Comparaison entre la structure de l'ADN Polymérase β humaine (PDB 1BPY en rose, à gauche) et l'ADN polymérase X a de Paramecium tetraurelia (Prédiction SwissModel en bleu, à droite), centrée sur le site actif et les résidus impliqués dans la transition conformationnelle chez l'ADN polymérase β. Les résidus impliqués sont indiqués.

3.2.3.3 *Les résidus impliqués dans ce mécanisme chez les ADN polymérases β sont uniques et conservés chez toutes les ADN polymérases β like*

Cependant, pour affirmer que la seule présence de ces résidus équivalents soit le signe d'un mécanisme équivalent, j'ai d'abord cherché si toutes les ADN polymérases β ou leurs équivalents portaient ces résidus : si ces résidus sont présents chez toutes les ADN polymérases β, qui disposent de ce mécanisme, alors ces résidus sont une signature de ce mécanisme.

Pour cela, j'ai utilisé les analyses de séquences réalisées plus tôt avec CLANS.

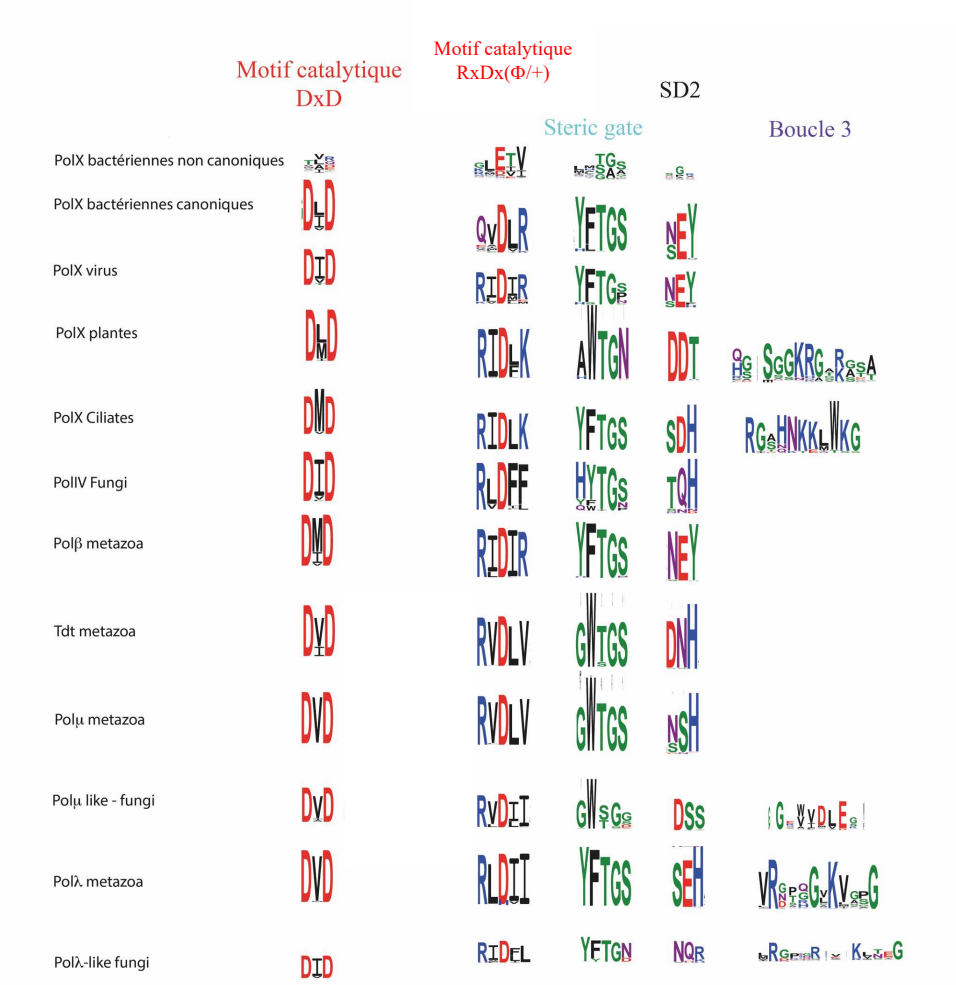


Figure 91 : Motifs caractéristiques des séquences des 12 clusters, centrés sur les motifs catalytiques, le steric gate et le motif SD2. La taille de chaque résidu est proportionnelle à sa fréquence, et la couleur des résidus dépend de leur catégorie chimique (rouge : résidu acide ; bleu : résidu basique ; vert : résidu polaire ; noir : résidu hydrophobe ; violet : résidu neutre).

En comparant les logos réalisés pour résumer les séquences de chaque cluster (figure 91), on constate que 5 clusters portent un résidu chargé positivement en position 5 du motif RxDx(Φ/+), et un résidu chargé négativement en position 2 du motif SD2 : les ADN polymérases β de métazoaires, les ADN polymérases β virales, les ADN polymérases X bactériennes canoniques, les ADN polymérases X de Sar, mais aussi les ADN polymérases X de plantes. Cependant, en étendant la comparaison à la recherche de la phénylalanine du *steric gate*, je me suis rendu compte que les ADN polymérases X de plantes ont à cette position un tryptophane. Les seuls groupes d'ADN polymérases X présentant l'ensemble des résidus impliqués dans ce mécanisme sont donc les ADN polymérases β de métazoaires, de virus, et les ADN polymérases bactériennes canoniques, soit les groupes proches de celui disposant du mécanisme *induced-fit*. J'en ai conclu que cet ensemble de résidus est une signature des ADN polymérases X disposant de ce mécanisme.

3.2.4 Le linker séparant les domaines BRCT et catalytique

Enfin, malgré ces éléments permettant de relier les ADN polymérases de *P. tetraurelia* aux ADN polymérases λ et β, un élément reste unique chez ces enzymes : la séquence liant le domaine BRCT au domaine catalytique. Comme indiqué précédemment, parmi les 12 clusters d'ADN polymérases X obtenus et étudiés, trois présentent une séquence plus longue que la moyenne : les deux clusters d'ADN polymérases de *Fungi*, et celui des ADN polymérases X de Sar. Les longueurs des séquences de ce dernier clusters sont moins variables, ce qui suggère que cette moyenne élevée n'est pas due à des séquences longues de façon aberrante. Surtout, en comparaison avec les autres ADN polymérases X (figure 85), cette grande longueur semble provenir de la séquence liant les domaines BRCT et catalytiques.

Parmi les ADN polymérases X déjà connues ayant un domaine BRCT, il n'y a que pour les ADN polymérases λ de métazoaires que la séquence équivalente a été étudiée. En effet, ce groupe de polymérases porte une séquence non structurée riche en sérines et prolines, dont le rôle n'est pas encore clair, malgré des études suggérant un ciblage pour des modifications post traductionnelles protégeant la polymérase de la dégradation, un rôle modulateur de l'activité en situation de synthèse translésionnelle, et même un rôle dans la fidélité de la polymérase (Fiala *et al.*, 2006; Garcia-Diaz *et al.*, 2005a).

Or comme l'indique l'alignement de séquences en figure 85, la séquence liant le domaine BRCT au domaine catalytique chez les ADN polymérases de *Paramecium* ne semble pas particulièrement riche en sérines et prolines, mais semble en revanche très conservée, et surtout plus longue que chez les autres ADN polymérases X. Le rôle de ce linker n'est pas encore connu, mais il est possible d'émettre plusieurs hypothèses. Premièrement, ce long linker pourrait au sein de la machinerie NHEJ spécialisée de *Paramecium* permettre à la polymérase de reconnaitre ses partenaires plus facilement, en les « cherchant » plus loin. Comme nous l'avons vu précédemment, les prédictions structurales d'AlphaFold et de RosettaFold semblent indiquer que ce linker n'est pas structuré. C'est possible, mais pas certain : le fonctionnement de ces algorithmes de prédiction repose sur l'existence de séquences similaires pour lesquelles la structure est connue, donc s'il n'existe pas telle séquence, l'algorithme peut avoir du mal à prédire correctement la structure. Il est aussi possible que ce linker ne soit structuré qu'en présence de ses partenaires, ou en présence d'ADN. En effet, la machinerie NHEJ spécialisée de *Paramecium* a plusieurs avantages sur le NHEJ classique : l'ADN adopte toujours la même forme lors de la coupure par PiggyMac, la machinerie NHEJ est recrutée immédiatement, et par

conséquent tout le processus est toujours le même. Il est donc possible que ce long linker ait pour rôle de se lier aux partenaires protéiques ou à l'ADN pour rigidifier l'ensemble.

3.3 Conclusion : Les explications possibles de la fidélité des ADN polymérases X de *Paramecium tetraurelia*

Pour rappel, la question principale des travaux de ma thèse est la suivante : comment expliquer que les ADN polymérases X impliquées dans la réparation des cassures double brins programmées chez *Paramecium tetraurelia* sont aussi fidèles ? Au cours des travaux décrits dans ce premier chapitre, j'ai pu définir deux hypothèses.

La première hypothèse est celle liant la présence de la boucle 3 chez les ADN polymérases X de *Paramecium* à un mécanisme permettant de stabiliser l'ADN dans le site actif, uniquement en présence d'un nucléotide correct. En effet, si comme chez l'ADN polymérase λ **cette boucle permet de stabiliser l'ADN template au sein du site actif**, sa conservation chez les ADN polymérases X de *Paramecium* pourrait indiquer sa **participation à la fidélité de ces polymérases**.

La seconde hypothèse repose sur l'existence possible d'un mécanisme équivalent au mécanisme *induced fit* expliquant la fidélité des ADN polymérases β. Par conséquent, si les ADN polymérases X de *Paramecium* utilisent un mécanisme similaire, cela pourrait aussi être une explication de leur fidélité. Cela a donc conduit à l'hypothèse suivante : **les ADN polymérases X de *Paramecium* utilisent un mécanisme d'activation à deux étapes, qui leur confère une grande fidélité**. Celui-ci reposerait sur une lysine en position 5 du motif RxDx(Φ/+) (K534), sur la phénylalanine du *steric gate* (F548), et sur l'ensemble des résidus du motif SD2 (S570, D571 et H572).

Au cours des expériences détaillées dans les chapitres suivants, j'ai pu étudier ces deux hypothèses, et montrer que les deux mécanismes observés chez les ADN polymérases β et λ ne sont pas cumulables (en tout cas pas tels quels), et que les ADN polymérases X de *Paramecium* portent des formes variantes mécanismes retrouvés chez les ADN polymérases β et λ.

Chapitre 2

Étude expérimentale des ADN polymérases X de *Paramecium tetraurelia* et de leur fidélité

1 Matériel et Méthodes

1.1 Production, purification et études biochimiques et enzymatiques des ADN polymérases de la famille X de *Paramecium tetraurelia*

1.1.1 Les constructions des ADN polymérases X de *Paramecium tetraurelia* étudiées

Pour l'étude des ADN polymérases X de *Paramecium tetraurelia*, le premier objectif a été de produire au moins un représentant de chaque sous-groupe (a/b et c/d), sous deux formes : complète (FL ou *Full-Length*), ou uniquement le domaine catalytique (sans le domaine BRCT ni le linker). Toutes les protéines étudiées dans ces travaux ont été exprimées en système bactérien, fusionnées à une étiquette (*tag*) de 14 histidines, clivable à l'aide de la protéase TEV (*Tobacco Etch Virus*, virus de la gravure du tabac).

Seule une construction a pu être étudiée sous forme complète, pour l'ADN polymérase Xd (identifiant de la séquence sur ParameciumDB : PTET.51.1.P1010039), nommée PolXdFL. PolXdFL a un poids moléculaire de 76588 Da avec son tag de 14 histidines, et de 72345 Da après clivage de ce tag.

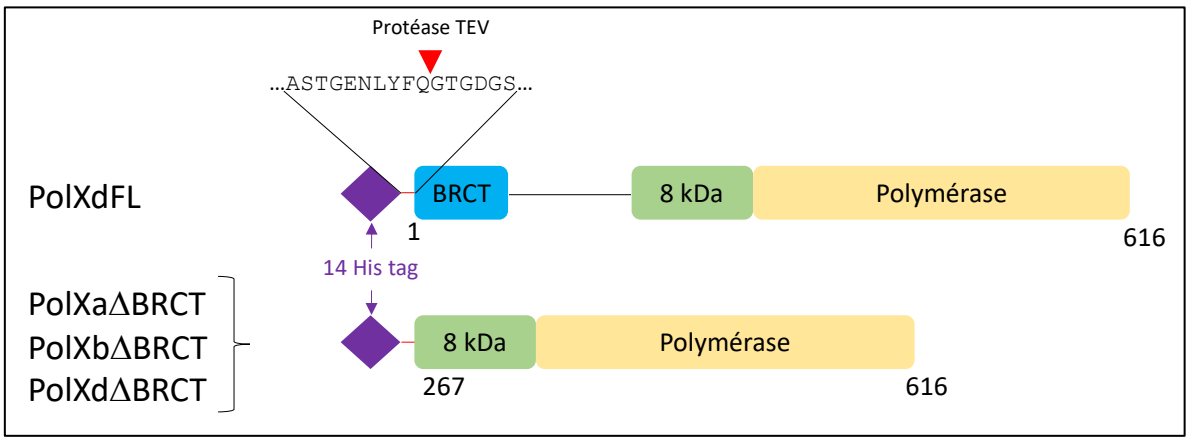


Figure 92 : Constructions des ADN polymérases X de Paramecium tetraurelia produites et étudiées. Le losange violet indique l'étiquette de 14 histidines. La séquence indiquée en haut est celle liant cette étiquette à la séquence de la protéine produite. Le site de clivage de la protéase TEV est indiqué avec un triangle rouge. Les domaines BRCT, 8 kDa et polymérase sont indiqués respectivement dans des rectangles bleu, vert et jaune. Les résidus inclus dans chaque construction sont indiqués sous la construction associée.

Plusieurs constructions, n'incluant que le domaine catalytique, ont été produites et purifiées dans ces travaux. La purification de ces protéines tronquées est supposée faciliter la cristallisation de ces protéines, comme cela a pu être fait pour d'autres ADN polymérases de la famille X (Gouge *et al.*, 2013; Jamsen *et al.*, 2022; Moon *et al.*, 2007). En effet, le domaine

BRCT et le linker sont trop flexibles pour permettre à ces protéines de cristalliser. Trois constructions tronquées d'ADN polymérases X de *Paramecium tetraurelia* ont pu être obtenues, pour les ADN polymérases Xa (PTET.51.1.P0210235), Xb (PTET.51.1.P0360066) et Xd. Elles ont donc été nommées respectivement PolXaΔBRCT, PolXbΔBRCT et PolXdΔBRCT, et comprennent toutes les résidus 267 à 616 de chacune des protéines sauvages. PolXaΔBRCT, lorsqu'elle est fusionnée à son étiquette de 14 histidines, a une masse moléculaire de 45026 Da (40798 Da après clivage de ce tag). Pour PolXbΔBRCT, cette valeur est de 45016 Da avec le tag, et de 40773 Da après son élimination. Enfin, pour PolXdΔBRCT, la masse moléculaire avant clivage du tag est de 45243 Da (et de 40999 Da après ce clivage).

1.1.2 Les constructions des ADN polymérases X humaines produites pour les tests enzymatiques

1.1.2.1 L'ADN Polymérase β

L'ADN polymérase β humaine (Uniprot P06746) a été produite sous sa forme sauvage complète. Avec le tag, son poids moléculaire est de 42895 Da. Après clivage du tag, sa masse moléculaire est de 38652 Da.

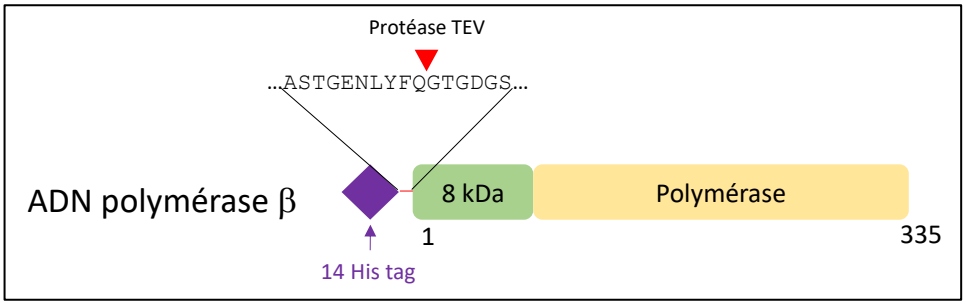


Figure 93 : Construction de l'ADN polymérase β humaine produite et étudiée. Le losange violet indique l'étiquette de 14 histidines. La séquence indiquée en haut est celle liant cette étiquette à la séquence de la protéine produite. Le site de clivage de la protéase TEV est indiqué avec un triangle rouge. Les domaines 8 kDa et polymérase sont indiqués respectivement dans des rectangles vert et jaune. Les résidus inclus sont indiqués sous ces rectangles.

1.1.2.2 L'ADN Polymérase λ

Dans la première partie de ces travaux, l'ADN polymérase λ humaine (Uniprot Q9UGP5) a été produite sous sa forme complète de 68069 Da en présence du tag 14 histidines et de 63825 Da sans ce tag.

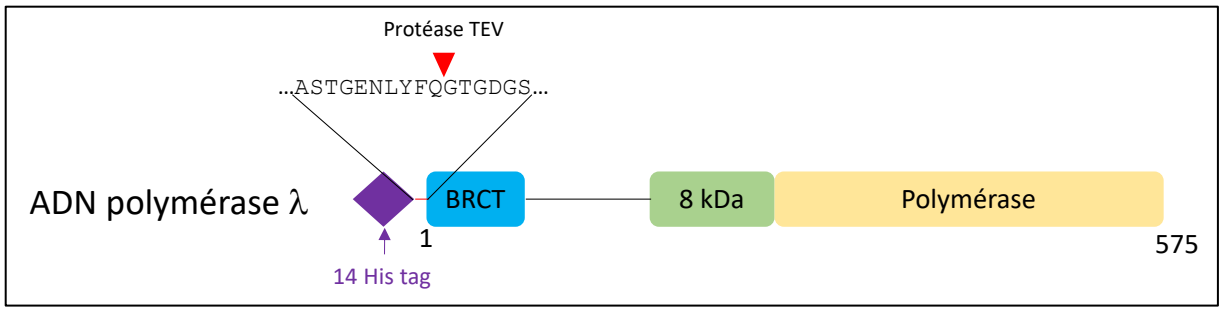


Figure 94 : Construction de l'ADN polymérase λ humaine produite et étudiée. Le losange violet indique l'étiquette de 14 histidines. La séquence indiquée en haut est celle liant cette étiquette à la séquence de la protéine produite. Le site de clivage de la protéase TEV est indiqué avec un triangle rouge. Les domaines BRCT, 8 kDa et polymérase sont indiqués respectivement dans des rectangles bleu, vert et jaune. Les résidus inclus sont indiqués sous ces rectangles.

1.1.3 Préparation des plasmides d'expression

1.1.3.1 PolXdFL

Le gène synthétique permettant d'exprimer PolXdFL a été optimisé pour son expression dans *E. coli* et synthétisé grâce au service ThermoFisher GeneArt. Les séquences de reconnaissance des enzymes de restriction BamHI (5'-GGATCC-3') et NotI (5'-GCGGCCGC-3') ont été ajoutées aux extrémités 5' et 3' du gène d'intérêt respectivement.

Deux méthodes de clonage ont été utilisées pour ce gène : le clonage par restriction-ligation et par la méthode de *Gibson assembly*.

1.1.3.1.1 Clonage par restriction-ligation

Une fois reçu, le plasmide portant le gène (résistant à l'ampicilline) a été utilisé pour transformer par choc thermique des bactéries chimiocompétentes *E. coli* DH5α (voir Annexe 1.2.1 page III et 1.3 page V). L'objectif à cette étape est de surproduire le plasmide dans les bactéries en les cultivant dans du milieu LB (*Lysogeny Broth*) à 37°C en présence d'ampicilline (100 µg/mL). Le plasmide a ensuite été récupéré et purifié (kit NucleoSpin Plasmid, Macherey Nagel), et le gène a été cloné dans le plasmide LS05 (un plasmide pRSF-Duet modifié, voir Annexe 1.1, page II) par la méthode de *STRU-cloning* (Bellini *et al.*, 2011) (voir Annexe 1.4.1, page VI) en utilisant les enzymes de restriction BamHI et NotI (New England Biolabs).

Après transformation de bactéries *E. coli* Top10 (Annexe 1.2.2, page IV) et culture en présence de kanamycine (50 µg/mL), l'insertion du gène d'intérêt a été vérifiée par PCR (*Polymerase Chain Reaction* ou amplification génique) sur colonie, en utilisant les amorces permettant le séquençage du plasmide LS05 : ACYCDuetUP1 (5'-GGATCTCGACGCTCTCCCT-3') et DuetDown1 (5'-GATTATGCGGCCGTGTACAA-3'),

chacune à une concentration de 250 nM. L'ADN polymérase utilisée était la DreamTaq (ThermoFisher) à une concentration de 25 mU/µL, avec un mélange de dNTPs à une concentration finale de 200 µM pour chaque dNTP. Plusieurs colonies ont été placées dans 10 µL de milieu LB, et 1 µL de ces cultures a été utilisé pour une réaction de PCR dans 50 µL. Chaque réaction a été incubée 3 min à 95°C, suivie de 25 cycles (30 secondes à 95°C, 30 sec à 55°C, 1 min à 72°C) et l'incubation a été terminée par 5 min d'élongation à 72°C.

Après la PCR, les échantillons ont été mélangés avec du SybrGreen 2X et séparés par électrophorèse sur gel d'agarose 1%. Le gel a été visualisé sous lumière ultraviolette, et les échantillons présentant des bandes de la taille attendue pour le gène d'intérêt ont été choisis. Les colonies bactériennes associées ont été repiquées dans du milieu LB liquide en présence de kanamycine, cultivées une nuit à 37°C, et l'ADN en a été extrait, purifié et séquencé (service Eurofins NightXpress) avec les mêmes amorces que pour la PCR.

Si les résultats de séquençage étaient corrects, c'est-à-dire si le gène d'intérêt avait bien été intégré dans le plasmide avec la séquence attendue, le plasmide purifié était alors conservé à -20°C pour être utilisé plus tard.

Les plasmides obtenus pour PolXdFL par cette méthode ont pu être utilisés dans un premier temps, mais les rendements obtenus lors des purifications étaient très faibles, et il a été décidé de recommencer ce clonage, par la méthode de *Gibson assembly* (Gibson *et al.*, 2009).

1.1.3.1.2 Clonage par Gibson Assembly

Cette méthode est décrite en détails dans l'annexe 1.4.2, page VI. En résumé, une PCR a été réalisée avec des amorces (créées selon les règles décrites dans l'Annexe 1.4.3, page VIII) permettant de linéariser le plasmide LS05 receveur et d'isoler le gène d'intérêt. Cette PCR a été réalisée avec l'ADN polymérase Q5 High Fidelity Hot Start 1X (New England Biolabs), avec ou sans ajout de GC enhancer. Les amorces ont été utilisées à une concentration de 500 nM, et la masse de plasmide utilisée était d'environ 10 ng pour une réaction faite dans un volume final de 50 µL. Les mélanges ont été incubés 3 min à 98°C, puis ont suivi 30 cycles (10 sec à 98°C, 30 sec à différentes températures testées en gradient entre 55 et 72 °C, 30 sec/kb à 72°C) et la PCR s'est terminée par 2 min d'élongation à 72°C. La taille des ADN amplifiés par PCR a été contrôlée par électrophorèse sur gel d'agarose 1%, comme indiqué précédemment.

Le kit NEBuilder HiFi DNA Assembly (New England Biolabs) a ensuite été utilisé pour insérer le gène de PolXdFL dans le plasmide LS05. Des bactéries DH5α ont été transformées avec le plasmide obtenu, et ont été cultivées une nuit à 37°C sur gélose en présence de kanamycine, puis des colonies ont été repiquées dans 5 mL de milieu LB liquide et cultivées à nouveau une nuit, toujours à 37°C en présence de kanamycine. L'ADN a ensuite été purifié et séquencé, comme dans la méthode précédente, et si la séquence du plasmide était correcte, il a été stocké à -20°C pour une utilisation future.

1.1.3.2 Les constructions n'incluant que le domaine catalytique (PolXaΔBRCT, PolXbΔBRCT, PolXdΔBRCT)

Les gènes synthétiques permettant d'exprimer les constructions «ΔBRCT » ont été optimisés et synthétisés grâce au service ThermoFisher GeneArt. Pour PolXaΔBRCT et PolXdΔBRCT, la méthode de *Gibson Assembly* (Gibson *et al.*, 2009) a été utilisée, de la même manière que pour PolXdFL.

PolXbΔBRCT a pu être cloné correctement par la méthode de restriction-ligation, de la même façon que pour le clonage initial de PolXdFL.

1.1.3.3 Les ADN polymérase β et λ humaines

Les plasmides LS05 permettant l'expression des ADN polymérases β et λ humaines ont été synthétisés par GenScript. Une fois reçus, ils ont été utilisés pour transformer des bactéries compétentes DH5α, de façon à les produire en grandes quantités, les purifier et les séquencer comme décrit précédemment. Une fois leurs séquences validées, ces plasmides ont été stockés à -20°C.

1.1.4 Production des ADN polymérases X de *Paramecium tetraurelia* et *Homo sapiens* en système bactérien

Pour produire en système bactérien de grandes quantités des protéines étudiées, un protocole commun a été utilisé.

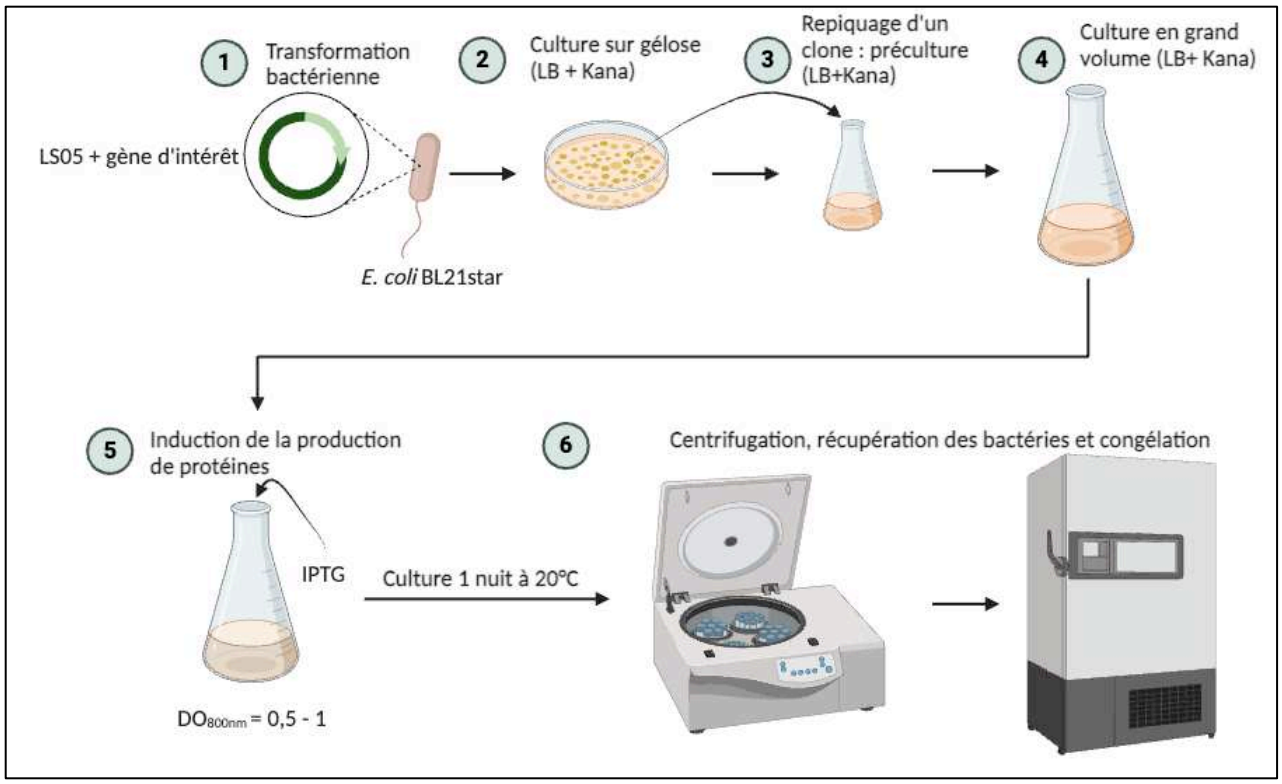


Figure 95 : Schéma du protocole général utilisé pour produire les protéines en système bactérien. Des bactéries E.coli BL21star sont transformées avec le plasmide LS05 portant le gène d'intérêt, et cultivées sur milieu LB gélosé et supplémenté en kanamycine. Une colonie est repiquée dans un petit volume de LB liquide supplémenté en kanamycine, puis cette préculture est utilisée pour ensemencer de grands volumes de culture. Quand la DO_{800nm} atteint une valeur située entre 0,5 et 1, de l'IPTG est ajouté pour induire la production de protéines par les bactéries. Après une nuit à 20°C, les bactéries sont récupérées par centrifugation et congelées. Figure créée sur BioRender.

Pour chaque construction, des bactéries *E. coli* BL21star (DE3) (Annexe 1.2.3, page III) ont été transformées avec le plasmide LS05 contenant le gène à exprimer. Après avoir été cultivées une nuit à 37°C sur milieu LB gélosé en présence de kanamycine, une colonie a été repiquée dans du milieu LB liquide en présence de kanamycine et cultivée une nuit à 37°C. Le lendemain, cette préculture a été utilisée pour inoculer du milieu LB-kanamycine en grand volume (de 2 L à 16 L, par erlenmeyers de 2L), à raison de 25 mL de préculture saturée par litre de culture. Ces cultures ont ensuite été incubées à 37°C sous agitation (130 rpm), et leur densité optique à 600 nm (DO_{600}) a été mesurée régulièrement. Lorsqu'elle était inférieure à 0,5 la culture à 37°C était poursuivie. Lorsqu'elle se situait entre 0,5 et 1, l'IPTG (isopropyl β-D-1-thiogalactopyranoside) était ajouté à une concentration finale de 1 mM pour induire la production de la protéine d'intérêt dans les bactéries, et les cultures ont été incubées à 20°C sous agitation jusqu'au lendemain. Les cultures ont été récupérées, centrifugées 20 min à 3000 g, les culots ont été resuspendus dans du milieu de culture et centrifugés à nouveau, puis conservés à -20°C.

1.1.5 Purification des protéines produites

Toutes les purifications présentées dans ces travaux ont suivi les mêmes étapes principales, et ont été réalisées à 4°C.

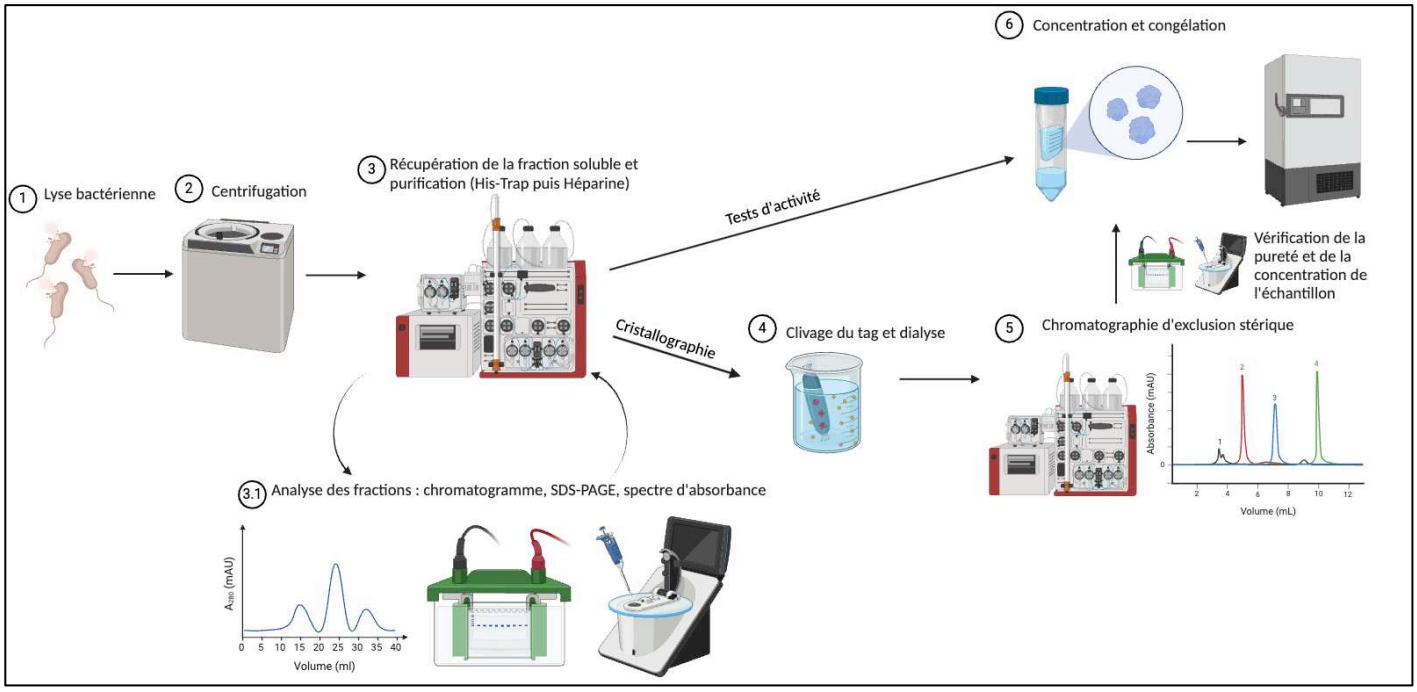


Figure 96 : Schéma du protocole général utilisé pour purifier les protéines étudiées. Après l'étape de lyse bactérienne, l'échantillon est centrifugé et la fraction soluble est injectée sur des colonnes His-Trap et Héparine pour purifier la protéine d'intérêt. À chaque étape, la qualité et la quantité de protéine purifiée sont estimées par SDS-PAGE et mesure d'absorbance à 280 nm. Dans le cas des protéines purifiées dans le but de faire des tests d'activité, les échantillons sont ensuite concentrés et conservés à -20°C. Dans le cas de protéines produites dans le but d'étudier leur structure, le tag de 14 histidines est supprimé lors d'une dialyse en présence de protéase TEV, puis la protéine est purifiée par chromatographie d'exclusion stérique. Les fractions présentant la protéine d'intérêt sont analysées par SDS-PAGE et leur concentration est mesurée, puis elles sont concentrées et conservées à -20°C. Figure créée sur BioRender

Après la lyse des bactéries, toutes les purifications ont été réalisées avec des colonnes (His-Trap FF 5 mL, HiTrap Heparin HP 5 mL, S200/S75, Cytiva) branchées sur un système Äkta Pure (Cytiva), permettant de suivre les chromatographies (suivi de l'absorbance à 280 nm (A_{280nm}), de la conductance, des gradients appliqués, etc). Dans certains cas (chromatographies faisant suite à une incubation avec la protéase TEV), une étape a pu être réalisée en chambre froide à l'aide d'une pompe péristaltique.

Les tampons utilisés lors des purifications sont indiqués en Annexe 1.6 (page X), et les principes de toutes ces étapes sont détaillés en Annexe 1.7 (page X).

1.1.5.1 Lyse des bactéries et récupération des protéines solubles

Pour lyser les bactéries ayant produit les protéines, deux méthodes ont pu être utilisées : la sonication et la *French Press* (ou *CellDisruptor*). Les principes de ces deux méthodes sont

respectivement détaillés en Annexe 1.7.1 et 1.7.2 (pages X et XI). Dans les deux cas, les bactéries ont tout d'abord été décongelées et resuspendues dans un tampon de lyse (tampon A supplémenté avec un cocktail d'antiprotéases et de la benzonase, une endonucléase permettant d'éliminer les acides nucléiques), à raison de 50 mL de tampon de lyse par litre de culture. La méthode de French Press n'a été utilisée que pour PolXdFL, en raison de sa tendance à agréger lors de la sonication. Après nettoyage de l'appareil avec du tampon de lyse, la suspension bactérienne a subi 3 cycles à une pression de 1,4 kbar dans l'appareil. Pour toutes les autres constructions, la méthode de sonication a été utilisée. La suspension bactérienne a été refroidie par un mélange d'eau et de glace, et la sonication a été faite en 5 cycles (1 seconde ON/ 1 seconde OFF, pendant une minute, suivie d'une minute de pause).

Les suspensions bactériennes ont ensuite été centrifugées 30 min à 20 000 g, pour séparer les protéines insolubles qui forment un culot des protéines solubles, qui restent alors en solution. Les fractions solubles ont été récupérées et filtrées à 0,22 μm.

1.1.5.2 Chromatographies d'affinité His-Trap

Pour chaque purification, l'extrait soluble filtré a été injecté sur une colonne His-Trap de 5 mL contenant une résine NiNTA par la pompe S de l'appareil Äkta Pure, à un débit de 3 mL/min. Après injection, la résine a été lavée avec 5 volumes de tampon A, puis l'élution a été réalisée avec un gradient de tampon B, permettant d'injecter sur la colonne des quantités croissantes d'imidazole (ces étapes ont été réalisées à un débit de 5 mL/min). Le suivi de l'absorbance à 280 nm par l'appareil de purification a permis de savoir dans quelles fractions de 1 mL la concentration en protéines était la plus élevée, et ces fractions ont été analysées en SDS-PAGE (Laemmli, 1970) (Annexe 1.7.4.1, page XIV).

1.1.5.3 Chromatographies sur résine Héparine

Les fractions contenant la protéine d'intérêt ont été rassemblées et diluées dans du tampon C, jusqu'à ce que leur concentration en NaCl soit inférieure à celle du tampon d'équilibration de la résine Héparine (tampon D). Elles ont ensuite été injectées de la même manière que pour l'étape His-Trap, et la résine a été lavée avec 5 volumes de tampon D. L'élution a été réalisée avec un gradient de tampon E, donc avec une concentration croissante de NaCl. Les fractions contenant le plus de protéine d'intérêt d'après le chromatogramme et l'analyse SDS-PAGE ont été rassemblées.

Dans le cas des protéines exprimées en vue de tests d'activité, la purification a pu se terminer après une étape de concentration (par diafiltration) de la protéine en tube concentrateur (Vivaspin ou Amicon) avec un seuil de poids moléculaire inférieur d'au moins 10 kDa au poids moléculaire de la protéine purifiée. Une fois la protéine purifiée, elle a été congelée dans l'azote liquide et conservée à -20°C.

Pour les protéines purifiées en vue d'essais de cristallogenèse, deux étapes ont été ajoutées : un clivage du tag de 14 histidines par la protéase TEV et une chromatographie d'exclusion stérique.

1.1.5.4 Clivage de l'étiquette 14 histidines

Après l'étape de chromatographie sur résine Héparine, les fractions rassemblées ont été mélangées avec des aliquots de protéase TEV à 10 mg/mL (1 aliquot de 100 µL pour 15 mL d'échantillon). Ce mélange a été dialysé une nuit à 4°C dans le tampon de stockage de la protéine d'intérêt (tampon F) supplémenté avec 2 mM de β-mercaptoéthanol, sous agitation. La dialyse a été réalisée soit avec un tube de dialyse, soit avec une cassette de dialyse, dans les deux cas avec un seuil de poids moléculaire inférieur à celui de la protéine d'intérêt sans son tag. Le lendemain, l'échantillon dialysé a été récupéré et injecté sur la colonne His-Trap, comme indiqué précédemment. Les fractions, en particulier celles issues du lavage de la colonne et de la charge de l'échantillon, ont été analysées par SDS-PAGE, et celles contenant la protéine d'intérêt sans son tag de 14 histidines ont été conservées et concentrées.

1.1.5.5 Chromatographies d'exclusion stérique

Les chromatographies d'exclusion stérique ont été réalisées avec des colonnes Superdex200 10/300 ou 16/60. Dans tous les cas, un échantillon concentré a été injecté sur la colonne équilibrée avec du tampon F, et les protéines ont été séparées avec un débit allant de 0,5 à 1 mL/min, selon les spécifications de la colonne. Les fractions correspondant aux pics d'absorbance à 280 nm ont été analysées par SDS-PAGE, et celles contenant les protéines d'intérêt ont été rassemblées, concentrées, congelées dans l'azote liquide et conservées à -20°C.

1.1.5.6 Contrôle de la qualité des protéines purifiées

La qualité des protéines purifiées a pu être vérifiée au cours des purifications par SDS-PAGE (Annexe 1.7.4.1, page XIV) et par mesure du spectre d'absorbance des échantillons entre

220 nm et 350 nm (NanoDrop) (Annexe 1.7.4.2, page XIV), qui permet d'obtenir des informations sur la concentration de la protéine, sur l'éventuelle contamination des échantillons par des acides nucléiques, et sur l'agrégation des protéines contenues dans ces échantillons.

En fin de purification, la qualité des échantillons a pu être vérifiée de façon plus approfondie dans certains cas, grâce à la PlateForme de Biophysique Moléculaire et d'Interactions (PFBMI, Institut Pasteur). Deux méthodes additionnelles ont alors été utilisées : une mesure précise de la masse moléculaire par spectrométrie de masse MALDI-TOF (*Matrix Assisted Laser Desorption Ionization – Time Of Flight* (Annexe 1.7.4.3, page XV)) et une mesure de la dispersité des espèces en solution par DLS (*Dynamic Light Scaterring* ou diffusion dynamique de la lumière) (Annexe 1.7.4.4, page XVI).

1.1.6 Caractérisation de l'activité des ADN polymérases X de *Paramecium tetraurelia*

Après les avoir purifiées, la première étape de l'étude expérimentale des ADN polymérases X de *Paramecium tetraurelia* a eu pour but de déterminer les contextes dans lesquels ces polymérases peuvent être actives ainsi que de les caractériser d'un point de vue cinétique.

1.1.6.1 Contextes testés

Dans un premier temps, l'objectif a été de déterminer les activités possibles des ADN polymérases X de *P. tetraurelia*, comme cela a pu être fait pour d'autres polymérases (Moran and Wilson, 2022). Plusieurs substrats ont donc été choisis et préparés.

- Extension d'amorce: le contexte principal des ADN polymérases réplicatives.

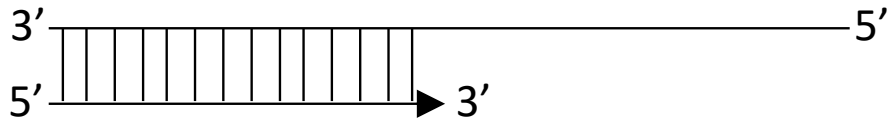


Figure 97 : Schéma du substrat ADN mimant un contexte d'extension d'amorce. L'amorce à étendre est indiquée par une flèche.

- « *Microhomology mediated end joining* » (« MMEJ ») : un contexte mimant une cassure double brins, avec une homologie de 10 paires de bases (pb) entre les ADN situés en amont et en aval de la cassure, permettant à la polymérase d'ajouter un nucléotide grâce au brin matrice situé sur le brin d'ADN en *trans* : si la polymérase

doit étendre l'ADN du duplexe situé en amont du dommage, elle prend son information modèle sur le duplexe en aval. Ici l'homologie est théoriquement assez grande pour que les deux duplexes d'ADN soient liés et forment un quadruplexe en solution, et l'amorce située en aval de la cassure porte un groupement phosphate en 5', nécessaire pour une réparation correcte par l'ADN polymérase λ.

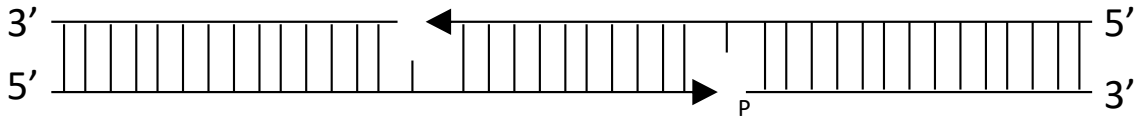


Figure 98 : Schéma du substrat ADN mimant un contexte de MMEJ. L'amorce à étendre est indiquée par une flèche, et le brin amorce en aval du dommage porte un groupement phosphate en 5', indiqué par un P.

- NHEJ : un contexte similaire, avec une microhomologie de seulement 2 pb (ne permettant pas à l'ADN seul de former un quadruplexe en solution) et un groupement phosphate en 5' de l'amorce en aval.

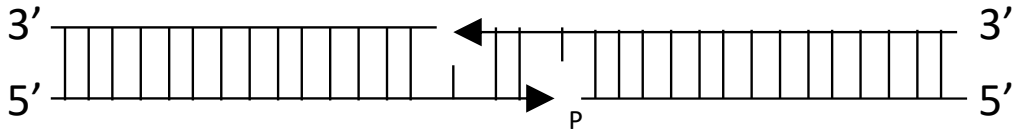


Figure 99 : Schéma du substrat ADN mimant un contexte de NHEJ. L'amorce à étendre est indiquée par une flèche, et le brin amorce en aval du dommage porte un groupement phosphate en 5', indiqué par un P.

- « *Gap-filling* » : ici, le brin matrice est continu, et le brin amorce est en deux parties séparées par un trou (gap) de 1 nucléotide. La polymérase doit donc ajouter 1 nucléotide. Ce contexte est celui principalement rencontré par l'ADN polymérase β, mais il a aussi été utilisé pour caractériser l'ADN polymérase λ. Le brin amorce en aval du dommage porte un groupement 5'P.

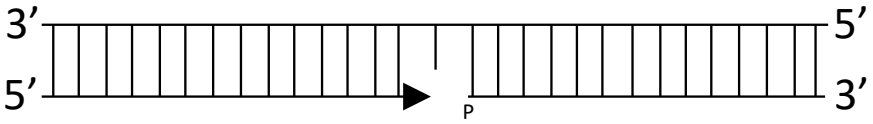


Figure 100 : Schéma du substrat ADN mimant un contexte de gap-filling. L'amorce à étendre est indiquée par une flèche, et le brin amorce en aval du dommage porte un groupement phosphate en 5', indiqué par un P.

- « NHEJ-cis » : le contexte rencontré par les ADN polymérases X de *P. tetraurelia* lors de la réparation des cassures double brins programmées (Figure 44). Ce contexte diffère du NHEJ décrit plus haut car la microhomologie de 2 pb est strictement

composée d'un dinucléotide 5'-TA-3', et surtout l'ADN polymérase lit le brin *template* en *cis*, c'est-à-dire que pour étendre l'ADN en amont de la cassure, elle utilise comme modèle l'ADN directement complémentaire, lui aussi situé en amont de la cassure (comme dans le contexte d'une extension d'amorce). Cependant, pour que l'extension ait lieu de façon correcte, l'enzyme ne doit ajouter qu'un nucléotide : si elle en ajoute plusieurs, cela suppose que la microhomologie n'a pas été conservée, et que le duplexe en aval a été déplacé. Ce duplexe en aval comporte un groupement 5'P.

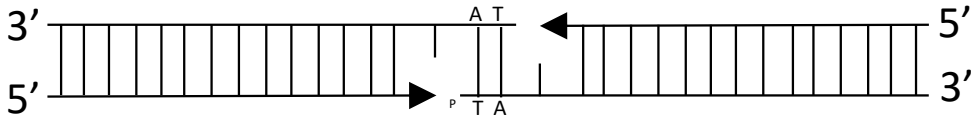


Figure 101 : Schéma du substrat ADN mimant un contexte de NHEJ-cis. L'amorce à étendre est indiquée par une flèche, et le brin amorce en aval du dommage porte un groupement phosphate en 5', indiqué par un P.

- Terminal transférase : Ici, l'ADN polymérase doit étendre une amorce plus longue que son *template*, voire une amorce sans template. La Tdt est la seule ADN polymérase de la famille X à être active naturellement dans ce contexte, mais l'ADN polymérase μ peut aussi être active, en présence d'ions Mn^{2+}.

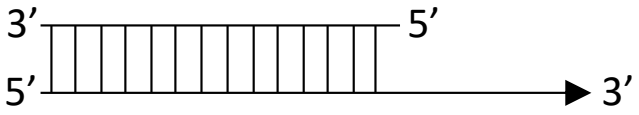


Figure 102 : Schéma du substrat ADN mimant un contexte nécessitant l'activité terminal transférase. L'amorce à étendre est indiquée par une flèche, et le brin amorce en aval du dommage porte un groupement phosphate en 5', indiqué par un P.

1.1.6.2 Préparation des substrats d'ADN testés

Les oligonucléotides utilisés (détaillés en Annexe 2, page XVII) ont été synthétisés par Eurogentec, Eurofins ou Biomers. Dans tous les cas, ils ont été suspendus dans du tampon TE (10 mM Tris-HCl pH8, 1 mM EDTA) à une concentration de 100 μM et conservés à -20°C.

Ces oligonucléotides ont ensuite été mélangés dans un tampon d'hybridation (10 mM Tris-HCl pH8, 1 mM EDTA, 50 mM NaCl), pour obtenir une concentration finale de 10 μM de duplexe (extension d'amorce, terminal transférase), triplex (gap-filling) ou quadruplex (NHEJ, NHEJ-cis et MMEJ). Les mélanges ont été incubés à 95°C pendant 10 min puis refroidis progressivement (-1°C/min) jusqu'à 20°C, puis conservés à -20°C. Chaque complexe d'ADN

comportait un marquage fluorescent FAM en 5' sur l'amorce devant être étendue par les ADN polymérases.

1.1.6.3 Préparation des ADN polymérases testées

Dans ces tests, les ADN polymérases testées ont été les suivantes : PolXaΔBRCT, PolXbΔBRCT, PolXdΔBRCT, PolXdFL, ADN polymérase λ et μ humaines (cette dernière a été produite et purifiée par le Dr Sophia Missoury au laboratoire).

Toutes les ADN polymérases testées ont été préparées pour ces tests de la même manière : après avoir été décongelées à 4°C, elles ont été centrifugées 10 min à 21 000g à 4°C pour éliminer les agrégats éventuels, puis leur concentration molaire a été calculée à partir de leur absorbance à 280 nm mesurée au Nanodrop. Elles ont ensuite été diluées dans du tampon d'activité 1X jusqu'à 10 fois la concentration utilisée dans les tests.

1.1.6.4 Tests d'activité et obtention des résultats

Les tests d'activité ont tous été réalisés dans un volume de 10 μL, soit en barrette de tubes PCR, soit en plaque 96 puits, avec 1 μM d'ADN hybridé. Des nucléotides ont été ajoutés à une concentration finale de 250 μM, ainsi que les ADN polymérases et du tampon d'activité (1X final). Pour la grande majorité des tests d'activité, le tampon utilisé était composé de 50 mM de Tris-HCl pH 7,5, 50 mM de KCl, 5 mM de $MgCl_2$, 1 mM de DTT et 5% de glycérol. Cependant, dans le cadre des tests de l'activité terminal transférase, le $MgCl_2$ a été remplacé par du $MnCl_2$, le manganèse étant l'ion permettant à l'ADN polymérase μ utilisée ici en contrôle d'avoir une activité nucléotidyltransférase (Martin *et al.*, 2013). Ces deux tampons ont été préparés 10 fois concentrés (10X) et utilisés à une concentration de 1X lors des tests.

Les mélanges ont été incubés pendant 30 min à 27°C pour les ADN polymérases de *P. tetraurelia* (la température optimale de croissance de l'organisme), ou à 37°C pour les ADN polymérases humaines. Les réactions ont été arrêtées par ajout de 20 μL de solution d'arrêt (10 mM EDTA, 98% formamide, 1 mg/ml bleu de bromophénol), chauffées à 95°C pendant 10 min, et conservées à -20°C.

Pour être analysées, les réactions arrêtées ont été chauffées à nouveau 10 min à 95°C puis déposées sur des gels urée-polyacrylamide (Albright and Slatko, 1994). Des témoins négatifs ont été utilisés comme marqueurs de taille, et correspondaient à l'ADN testé n'ayant

pas été en contact avec les ADN polymérases. Les électrophorèses ont été réalisées à 2000V (et 40 mA par gel) pendant des durées allant jusqu'à 4h. Les résultats d'électrophorèses ont été révélés en scannant les gels au Typhoon FLA 9000, et traités avec le logiciel ImageJ.

1.1.6.4.1 Problèmes rencontrés et solutions employées

Dans certains cas, les résultats sur gels ont montré des bandes additionnelles, correspondant à des tailles d'ADN inhabituelles, y compris pour les témoins négatifs, non traités par les enzymes testées. Nous avons d'abord pensé à un problème de conservation des stocks d'oligonucléotides, mais des analyses par spectrométrie de masse ont montré que les oligonucléotides utilisés avaient bien la masse attendue et n'étaient pas dégradés.

Ces bandes provenaient d'une mauvaise dénaturation des échantillons : les complexes d'ADN se séparaient mal malgré la présence de formamide et le chauffage à 95°C, donc leur migration sur gel était modifiée. La solution utilisée a été l'ajout d'un excès (10X) d'un ADN compétiteur au moment d'arrêter les réactions, identique à l'amorce étendue par les ADN polymérases mais sans marquage fluorescent. Ainsi, le brin *template* lié à l'amorce marquée s'en détache lors du chauffage et se lie préférentiellement à l'ADN non marqué (présent en excès).

1.1.7 Test de l'activité dRP lyase d'une ADN polymérase X de *Paramecium tetraurelia*

1.1.7.1 Principe de l'expérience

L'objectif ici est de placer une ADN polymérase X de *P. tetraurelia* dans un contexte où il est possible de constater les effets de son activité dRP lyase. J'ai donc choisi de reconstituer *in vitro* un équivalent du système BER *short-range*. Dans ce cas précis, après que l'ADN polymérase ait supprimé le groupement dRP, elle peut ajouter un nucléotide, et une ADN ligase peut rétablir la liaison phosphodiester, donc l'ADN retrouve sa taille initiale.

J'ai donc reconstitué un duplexe d'ADN (la séquence est détaillée dans la figure 103) présentant une déoxyuridine (dU) sur un brin marqué en 3' par fluorescence FAM. Celui-ci a été mélangé avec les composants du kit USER3 (New England Biolabs). Ce kit contient une Uracile-Glycosylase, qui a pour rôle d'éliminer le dU, laissant un site abasique. La seconde enzyme du kit, l'endonucléase IV, a pour rôle d'éliminer ce site abasique, en laissant un espace de 1 nt, suivi en aval d'un groupement dRP.

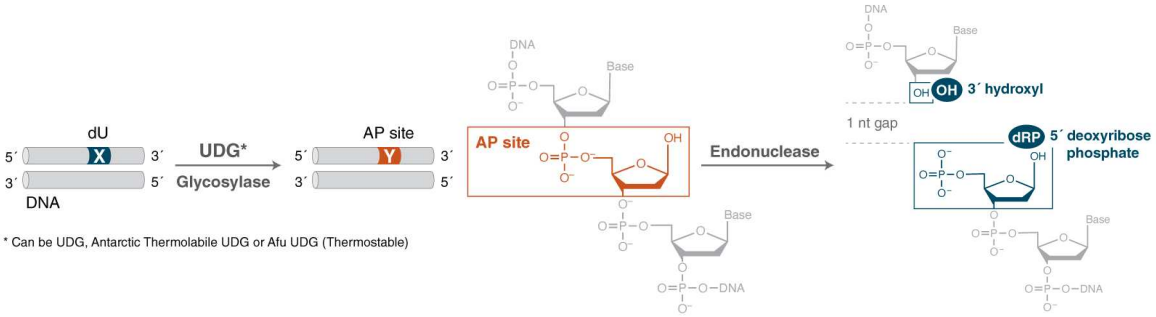


Figure 103 : Activité des enzymes du kit USER3 (New England Biolabs).Une uracile glycosylase a pour rôle de supprimer le dU présent sur l'ADN, en le remplaçant par un site abasique (AP). L'endonucléase IV a ensuite pour rôle de supprimer le site abasique en laissant un espace de 1 nt suivi d'un groupement dRP.

L'ADN polymérase testée est ensuite ajoutée avec des nucléotides, et son activité dRP lyase est supposée éliminer le groupement dRP, lui permettant ensuite d'ajouter un nucléotide. L'ADN ligase du phage T4 est ensuite ajoutée avec de l'ATP, et a pour rôle de rétablir la liaison phosphodiester manquante, ce qui rend sa taille initiale à l'ADN. Ainsi, on peut suivre grâce à la fluorescence de l'ADN marqué sa coupure (après utilisation du kit USER3), puis le rétablissement de sa taille initiale, qui ne peut avoir lieu que si la polymérase a supprimé le groupement dRP et ajouté un nucléotide, et si l'ADN ligase a pu rétablir la liaison phosphodiester, ce qui n'est pas possible en présence du dRP.

Figure 104 : Schéma des étapes du test de l'activité dRP lyase. La première étape (suppression du dU et remplacement pas un groupement dRP) est réalisée par les composants du kit USER3 (New England Biolabs). Les deux activités indiquées en gras sont celles de la polymérase (si elle dispose bien de l'activité dRP lyase, comme c'est le cas pour l'ADN polymérase β). L'étape de ligation (en gris) est réalisée par la T4 DNA ligase, avec de l'ATP.

1.1.7.2 Test d'activité

Les oligonucléotides ont été hybridés de la même façon que précédemment, de façon à obtenir un mélange à 50 µM. 4 pmoles d'ADN hybridé ont été mélangées aux enzymes du kit

USER3 et au tampon ThermoPol (20 mM Tris-HCl pH 8,8, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton® X-100), et l'ensemble a été incubé à 65°C pendant 2h.

En parallèle, PolXaΔBRCT et l'ADN polymérase β humaine ont été décongelées, centrifugées et diluées dans le tampon d'activité jusqu'à une concentration de 500 nM.

Dans un volume de 10 µL, 1 µM d'ADN traité par USER3 a été mélangé à 250 µM de dNTPs, 400 unités d'ADN ligase T4 et 50 nM de chaque ADN polymérase, dans un tampon d'activité (50 mM Tris-HCl pH 7,5, 5 mM $MgCl_2$, 50 mM KCl, 1 mM DTT et 1 mM ATP). Le mélange a été incubé à 27°C pendant 30 min, puis la réaction a été arrêtée et le résultat a été visualisé sur gel comme indiqué précédemment. Des témoins négatifs ont été réalisés : un premier n'incluant que l'ADN initial non traité par les enzymes du kit USER3, un second avec l'ADN traité, une réaction sans ADN polymérase, et une réaction sans ligase. Le témoin positif était quant à lui le test réalisé avec l'ADN polymérase β.

1.1.8 Caractérisation cinétique des ADN polymérases X de *Paramecium tetraurelia*

Pour comparer les ADN polymérases X de *Paramecium tetraurelia* aux ADN polymérases λ et β humaines, l'objectif suivant a été de les caractériser du point de vue cinétique. Dans la littérature, cela a été réalisé pour les deux enzymes humaines en contexte de gap-filling (Chagovetz *et al.*, 1997; Garcia-Diaz *et al.*, 2004). Bien que ce contexte ne corresponde pas à la réalité biologique pour les ADN polymérases X de *P. tetraurelia* (ni pour l'ADN polymérase λ), j'ai donc choisi de faire cette caractérisation dans ce contexte de cassure simple brin, afin de pouvoir comparer mes résultats à ceux de la littérature scientifique.

Ici, seules deux enzymes de *P. tetraurelia* ont été caractérisées : PolXaΔBRCT et PolXdΔBRCT, l'objectif étant de comparer les deux sous-groupes a/b et c/d. Pour chaque ADN polymérase, ces tests ont été réalisés au moins 3 fois.

1.1.8.1 Test enzymatique

Les mêmes oligonucléotides hybridés que pour les tests de gap-filling présentés précédemment ont été utilisés, à une concentration finale de 1 µM. L'ADN a été mélangé à 5 nM de chacune des ADN polymérases testées, avec du tampon d'activité (50 mM Tris-HCl pH 7,5, 50 mM KCl, 5 mM $MgCl_2$, 1 mM DTT et 5% glycérol). Le nucléotide supposé être incorporé dans ce contexte (dGTP) a été ajouté à des concentrations allant de 200 nM à 10 µM,

et l'ensemble a été incubé à 27°C pendant 10 min. Les réactions ont été arrêtées comme précédemment, et déposées sur gel urée-PAGE pour être analysées.

1.1.8.2 Obtention des résultats bruts

Une fois les gels visualisés, ils ont été analysés avec le logiciel ImageJ. Chaque piste a été analysée grâce à la fonction *Plot lanes*. Une ligne de base a été tracée pour chaque piste de façon à séparer les pics d'intensité des bandes, de la manière la plus constante possible. La fonction *Wand - Magic tool* a ensuite été utilisée pour mesurer les aires sous la courbe pour chaque pic, correspondant à l'intensité des bandes mesurées.

Pour chaque piste, le traitement a ensuite été réalisé sur Microsoft Excel par la méthode suivante (Gahlon and Sturla, 2019) : une mesure du bruit de fond a été soustraite aux intensités des bandes +1 (bande avec ajout d'un nucléotide), puis l'intensité des bandes +1 a été divisée par la somme de l'intensité des bandes 0 (bande initiale sans ajout de nucléotides) et +1. Cela a permis d'obtenir l'intensité des bandes +1 relative à l'intensité totale des bandes. Cette mesure correspondant à la quantité d'ADN présent dans la bande +1 rapportée à la quantité totale d'ADN (0 et +1), la multiplier par la quantité totale d'ADN dans le test permet de connaitre la quantité d'ADN que l'ADN polymérase a étendu, donc la concentration de produit. Cette quantité de produit formé a ensuite été divisée par la durée du test en minutes afin d'obtenir la mesure de vélocité de l'enzyme (en nM/min) pour chaque piste, donc pour chaque concentration de dGTP testée.

1.1.8.3 Traitement des résultats

Ces valeurs ont ensuite été traitées grâce au logiciel GraphPad Prism 10. Les données ont été analysées par régression non linéaire de façon à obtenir une approximation du modèle de Michaelis-Menten (Johnson and Goody, 2011) ($v_i = v_{max} \frac{[S_0]}{K_M+[S_0]}$), avec pour objectif d'obtenir les constantes K_m et k_{obs} (Annexe 3, page XIX) des deux enzymes (donc $v_i = [E_t]k_{obs} \frac{[S_0]}{K_M+[S_0]}$), en définissant initialement la concentration totale d'enzyme.

1.1.8.4 Critique de la méthode

Dans le cas de la caractérisation de ces enzymes, il a été possible d'obtenir des résultats assez reproductibles. Cependant, dans d'autres cas j'ai pu être confronté à des traitements de données similaires qui ont posé plusieurs problèmes. La première étape de préparation des gels

peut être source de résultats variables : l'expérience au sein du laboratoire a montré que les mêmes échantillons déposés sur deux gels différents peuvent aboutir à des résultats légèrement différents, en termes de contraste, de finesse des bandes, etc. Il est également difficile de traiter toujours exactement de la même façon les images de gels sur ImageJ : l'étape critique est ici le tracé de la ligne de base, car c'est à cette étape qu'on peut introduire beaucoup d'approximations qui se répercutent sur l'analyse des résultats *a posteriori*. Certains résultats obtenus par cette méthode doivent donc être observés avec prudence, et en gardant un esprit critique. C'est pourquoi les résultats présentés ci-après ont été obtenus en reproduisant plusieurs fois les mêmes expériences (de 3 à 12 réplicats).

Cependant, d'autres méthodes peuvent permettre d'obtenir des résultats plus reproductibles induisant moins de variabilité. On peut par exemple citer les travaux réalisés par Kelly Zatopek (Zatopek *et al.*, 2020) et Leonardo Betancurt-Anzola (Betancurt-Anzola *et al.*, 2023), qui avec un appareil séquenceur Sanger modifié et une électrophorèse capillaire obtiennent des chromatogrammes très résolutifs permettant une analyse de brins d'ADN au nucléotide près, avec une ligne de base très reproductible. De plus, cette méthode permet un bien meilleur débit, puisque l'analyse d'un échantillon ne prend que quelques minutes, et que ce type d'appareil peut analyser plusieurs dizaines d'échantillons à la suite.

1.1.9 Comparaison de la fidélité des ADN polymérases X de *Parameciu tetraurelia* et de l'ADN polymérase λ humaine

La première question de ces travaux était la suivante : les ADN polymérases X de *P. tetraurelia* sont-elles à l'origine de la grande fidélité du système de réparation des CDB programmées chez cet organisme ?

Un de mes objectifs a donc été de tester la fidélité de ces ADN polymérases X, en particulier en comparaison avec l'ADN polymérase λ. La majorité des travaux réalisés sur l'ADN polymérase λ se reposaient sur des études « Pré *steady-state* » de constructions tronquées de cette ADN polymérase, impliquant des temps expérimentaux très courts, qu'il m'était impossible techniquement de réaliser. Cependant, une étude publiée en 2004 (Fiala *et al.*, 2004) indiquait un protocole incluant des temps plus longs qu'il m'a été possible de suivre pour observer des incorporations erronées de l'ADN polymérase λ.

Le substrat utilisé était celui utilisé dans les tests de gap-filling. Les ADN polymérases ont été testées en conditions de *single turnover*, donc en excès d'enzyme par rapport aux concentrations d'ADN. 120 nM des différentes ADN polymérases ont donc été mélangés à 30 nM d'ADN et 120 µM de chaque nucléotide incorrect. Le tampon utilisé contenait 50 mM de Tris-HCl pH 8,4, 2% de glycérol, 5 mM de $MgCl_2$, 100 mM de NaCl, 5 mM de DTT, 0,1 mg/mL de BSA, et 0,1 µM d'EDTA. L'ensemble a été incubé 50 minutes à 37°C pour les constructions de l'ADN polymérase λ humaine, et à 27°C pour les ADN polymérases X de *Paramecium tetraurelia*. Quatre ADN polymérases ont été testées : l'ADN polymérase λ humaine, une construction tronquée (construction λmut, décrite en détails dans la seconde partie de ces travaux), PolXaΔBRCT et PolXdFL. Les échantillons ont ensuite été déposés sur gel urée-PAGE et analysés comme précédemment.

1.1.10 Essais de cristallogenèse des ADN polymérases X de *Paramecium tetraurelia*

Dans le cas de PolXaΔBRCT et PolXdΔBRCT, plusieurs essais de cristallisation ont été réalisés, avec l'aide de la plateforme de cristallographie de l'Institut Pasteur.

1.1.10.1 Essais réalisés avec PolXaΔBRCT

Trois campagnes de criblage de conditions de cristallisation ont été réalisées pour PolXaΔBRCT, avec différents partenaires (ADN et nucléotides) inspirés d'expériences proches ayant permis d'obtenir les structures cristallographiques des ADN polymérases β (PDB 1RZT (García-Díaz *et al.*, 2002)) et λ (PDB 7M43 (Jamsen *et al.*, 2022)) humaines.

Tableau 4 : Complexes utilisés dans les campagnes de criblage de conditions de cristallisation pour PolXaΔBRCT.

Concentration de la protéine	**Oligonucléotides utilisés et structure mimée**		**Nucléotide utilisé**
8 mg/mL	5'-CAGTG-3' 5'-pCGTCG-3' 5'-CGACGACACTG-3'	Gap-filling (1 nt)	dTpCpp
10 mg/mL	5'-CAGTAC-3' 5'-pGCCG-3' 5'-CGGCAGTACCTG-3'	Gap-filling (1 nt) PDB 7M43	dTTP
10 mg/mL	5'-GTGCG-3' 5'-pGCCG-3' 5'-CGGCAACGCAC-3'	Gap-filling (2 nt) PDB 1RZT	dTTP

Dans chaque cas, les oligonucléotides ont été hybridés comme décrit précédemment. Ils ont été mélangés avec PolXaΔBRCT et les nucléotides indiqués dans le tableau ci-dessus à des ratios de concentrations molaires [protéine : ADN : nucléotide] de [1 :1,2 :2]. D'autres échantillons ont été préparés, en ne mélangeant que la protéine et l'ADN, ou avec la protéine seule.

Ces mélanges ont ensuite été déposés dans des plaques Greiner 96 puits, pour former des gouttes de 400 nL (200 nL de préparation de protéine, 200 nL de solution précipitante) en goutte assise, à l'aide d'un robot Genesis Workstation 150 (Tecan). Les solutions précipitantes provenaient des 7 kits suivants : Crystal Screen 1 et 2 (Hampton), JBScreen Wizard 1 et 2 (Jena Biosciences), Structure Screen 1 et 2 (Molecular Dimensions), JBSCreen 1 à 8 (Jena Biosciences), Salt RX 1 et 2 (Hampton) et PEGion 1 et 2 (Hampton).

Les plaques préparées ont été stockées à 18°C, et des photos de chaque goutte ont été réalisées régulièrement à l'aide d'un robot Formulatrix RockImager.

1.1.10.2 Essais réalisés avec PolXdΔBRCT

Des campagnes de criblage de conditions de cristallisation ont également été réalisées pour PolXdΔBRCT, avec les mêmes substrats que pour PolXaΔBRCT et d'autres.

Tableau 5 : Complexes utilisés dans les campagnes de criblage de conditions de cristallisation pour PolXdΔBRCT.

Concentration de la protéine	**Oligonucléotides utilisés et structure mimée**		**Nucléotide utilisé**
10 mg/mL	5'-CTGATGCGC-3' 5'-pGTCGG-3' 5'-CCGACGGCGCATCAG-3'	Gap-filling (1 nt)	ddCTP
	5'-GCTGATGCGC-3' 5'-pGTCGG-3' 5'-CCGACGGCGCATCAGC-3'		
	5'-CGCTGATGCGC-3' 5'-pGTCGG-3' 5'-CCGACGGCGCATCAGCG-3'		
10 mg/mL			
10 mg/mL	5'-CAGTG-3' 5'-pCGTCG-3' 5'-CGACGACACTG-3'	Gap-filling (1 nt)	ddTTP
10 mg/mL	5'-CAGTAC-3' 5'-pGCCG-3' 5'-CGGCAGTACCTG-3'	Gap-filling (1 nt) PDB 7M43	dTTP
10 mg/mL	5'-GTGCG-3' 5'-pGCCG-3' 5'-CGGCAACGCAC-3'	Gap-filling (2 nt) PDB 1RZT	dTTP

Ces mélanges ont été réalisés, mélangés avec les solutions de cristallisation et observés régulièrement de la même façon que pour PolXaΔBRCT.

1.2 Productions, purifications et études enzymatiques et structurales de versions mutantes de l'ADN polymérase λ humaine

1.2.1 Constructions mutantes de l'ADN polymérase λ humaine étudiées

1.2.1.1 Les mutations utilisées communes à toutes les constructions produites

D'après des travaux de 2022 (Jamsen *et al.*, 2022), trois mutations permettent soit de faciliter la cristallisation de l'ADN polymérase λ humaine, soit d'améliorer la résolution obtenue lors des expériences de diffraction des rayons X : la délétion de l'extension N-terminale (domaine BRCT et linker, résidus 0-241), trop flexible pour permettre d'obtenir des cristaux ; la modification de la boucle 1 (remplacement des résidus 464-472 par l'équivalent chez l'ADN polymérase β : KGET) pour améliorer la résolution obtenue lors des expériences de diffraction des rayons X ; la mutation C544A qui permet de faciliter la cristallisation. Une construction mutante de l'ADN polymérase λ humaine portant l'ensemble de ces mutations a été produite et nommée λmut.

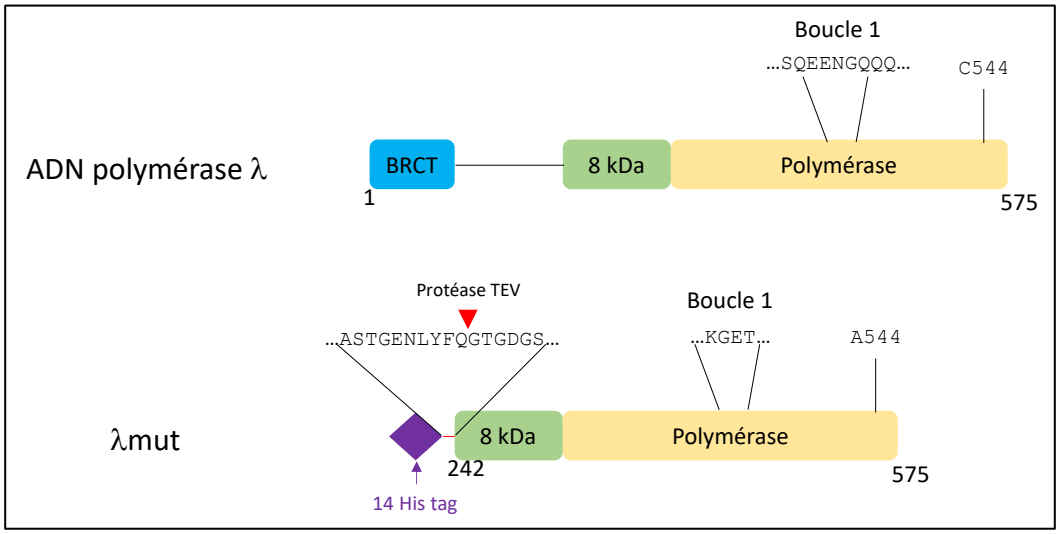


Figure 105 : Comparaison de l'ADN polymérase λ humaine sauvage et du mutant λmut produit et étudié. Le losange violet indique l'étiquette de 14 histidines. La séquence indiquée au-dessus haut est celle liant cette étiquette à la séquence de la protéine produite. Le site de clivage de la protéase TEV est indiqué avec un triangle rouge. Les domaines BRCT, 8 kDa et polymérase sont indiqués respectivement dans des rectangles bleu, vert et jaune. Les résidus inclus dans chaque construction sont indiqués sous la construction associée. Les résidus mutés dans la construction λmut sont indiqués au-dessus de celle-ci, et leurs équivalents dans la protéine sauvage sont indiqués sur celle-ci.

1.2.1.2 Les mutations visant à conférer à l'ADN polymérase λ le mécanisme de fidélité de l'ADN polymérase β ou des ADN polymérases X de Paramecium tetraurelia

Quatre constructions mutantes de l'ADN polymérase λ ont été produites. Deux visaient à conférer le mécanisme connu retrouvé chez l'ADN polymérase β et ont été nommées λmutR (mutation I493R) et Polβ–like (mutation I493R et mutation des résidus 529 à 531 (motif SD2) en NEY). Les deux autres constructions visaient à tester de la même manière le possible mécanisme des ADN polymérases X de *Paramecium tetraurelia* : λmutK (mutation I493K) et *Ptet-like* (mutations I493K et E530D).

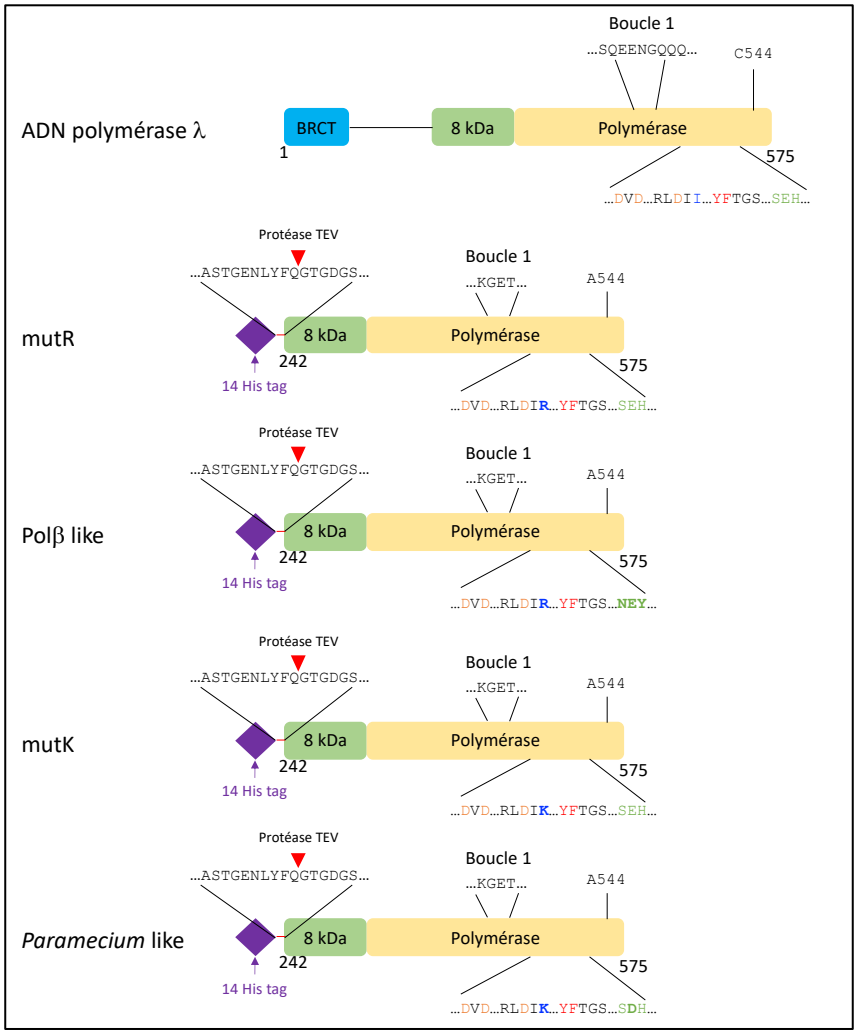


Figure 106 : Comparaison de l'ADN polymérase λ humaine sauvage et des mutants λmutR, Polβ like, λmutK et Ptet-like. Le losange violet indique l'étiquette de 14 histidines. La séquence indiquée au-dessus haut est celle liant cette étiquette à la séquence de la protéine produite. Le site de clivage de la protéase TEV est indiqué avec un triangle rouge. Les domaines BRCT, 8 kDa et polymérase sont indiqués respectivement dans des rectangles bleu, vert et jaune. Les résidus inclus dans chaque construction sont indiqués sous la construction associée. Les mutations communes à toutes les constructions sont indiquées au-dessus de celles-ci, et leurs équivalents dans la protéine sauvage sont indiqués sur celle-ci. Les mutations spécifiques de chaque mutant sont indiquées en gras sous ce mutant, et leurs équivalents dans la protéine sauvage sont indiqués sur celle-ci. Les résidus d'intérêt sont indiqués : résidus catalytiques en orange, résidus impliqués dans l'activation/inactivation du site catalytique en bleu, résidus du steric gate en rouge, et résidus du motif SD2 en vert.

Tableau 6 : Poids moléculaires des cinq constructions mutées de l'ADN polymérase λ étudiées

Nom de la construction mutante	Poids moléculaire avant clivage du tag (Da)	Poids moléculaire après clivage du tag (Da)
λmut	40947	36703
mutR	40990	36875
Polβ–like	41043	36800
mutK	40962	36847
Ptet-like	40948	36704

1.2.2 Préparation des plasmides d'expression

Pour obtenir l'ensemble des constructions mutantes présentées ci-dessus, des mutagenèses dirigées ont été réalisées successivement, en partant du plasmide LS05 portant le gène de l'ADN polymérase λ humaine, utilisé dans la première partie de ces travaux.

Pour la mutation Δ1-241 (nommée ΔBRCT), une PCR a été réalisée comme précédemment avec l'ADN polymérase Q5 avec des amorces s'hybridant à chaque extrémité de la séquence à éliminer (le principe est détaillé en Annexe 1.5.1, page VIII). Cela a pour but d'amplifier le plasmide, à l'exception de la partie située entre les deux amorces, qui correspond aux résidus 1 à 241. Après la PCR, la taille des fragments a été contrôlée par électrophorèse en gel d'agarose 1%. Lorsque le plasmide linéarisé avait la taille attendue, il a été incubé avec le mélange d'enzymes KLD (pour Kinase, Ligase, DpnI, (New England Biolabs)) pendant 5 min à température ambiante. Après cette réaction, des bactéries DH5α ont été transformées avec le plasmide recircularisé, et ont été cultivées à 37°C sur un milieu LB gélosé, puis des colonies ont été repiquées en milieu liquide. L'ADN plasmidique a été purifié et séquencé, comme dans les méthodes précédentes, et si sa séquence était correcte, il a été stocké à -20°C.

Pour l'ensemble des autres mutations, une méthode basée sur la PCR avec l'ADN polymérase Q5 a également été utilisée, avec des amorces crées pour amplifier l'ensemble du plasmide en intégrant la mutation à réaliser (principe en Annexe 1.5.2, page VIII). Comme précédemment, la taille du plasmide amplifié et linéarisé a été contrôlée par électrophorèse en gel d'agarose 1%, puis le mélange KLD a été utilisé pour le recircularisé. Ce plasmide a été produit en bactéries DH5α, purifié, et séquencé, comme précédemment.

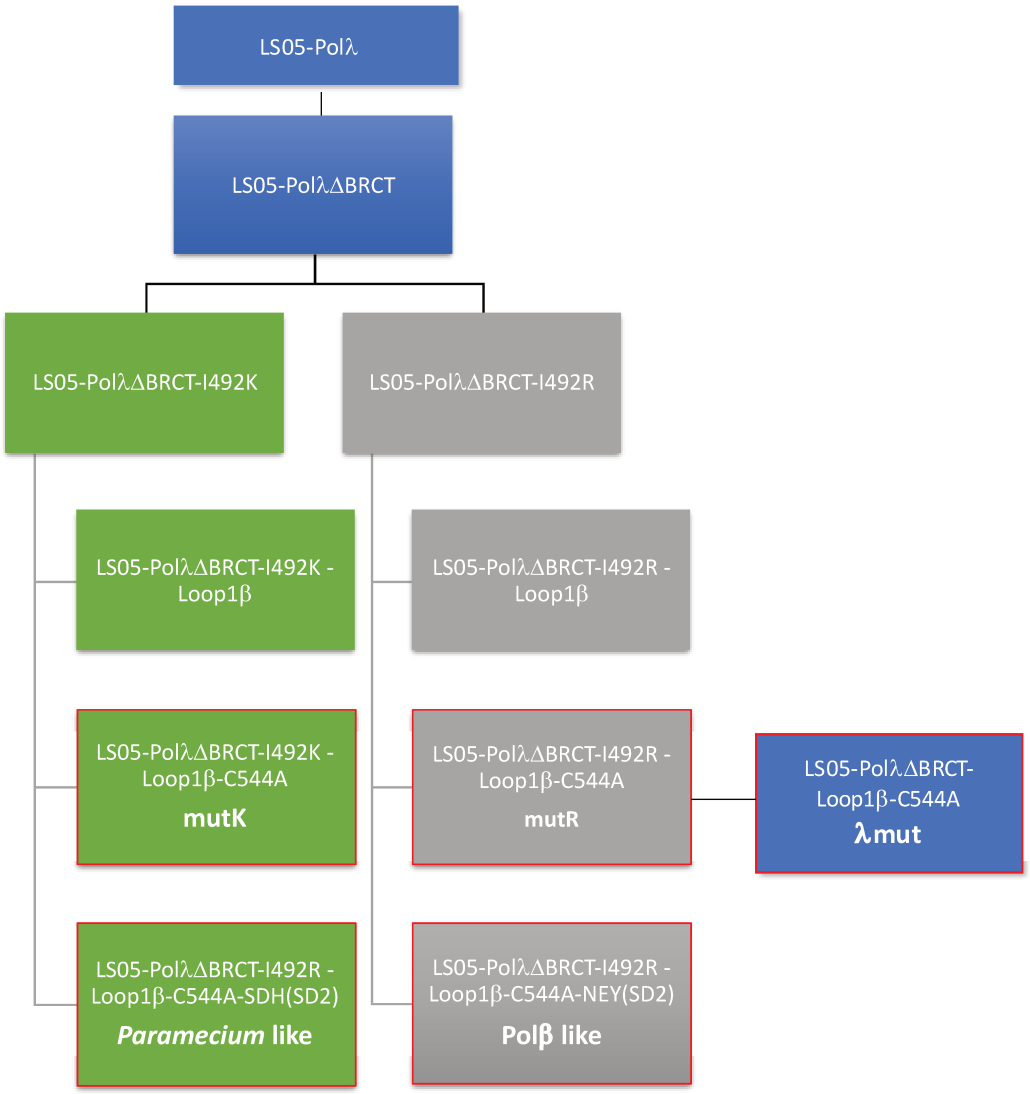


*Figure 107 : Ordre de réalisation des mutagenèses visant à obtenir les mutants λmutK, λmutR, Polβ–like, Ptet-like et λmut (encadrés en rouge, leur nom est indiqué en **gras**)*

1.2.3 Production des constructions mutées de l'ADN polymérase λ humaine en système bactérien et purification

Pour produire en système bactérien de grandes quantités des protéines étudiées, le même protocole que précédemment a été utilisé (Chapitre 2, 1.1.4, page 135).

Les purifications de ces constructions ont également été réalisées suivant les mêmes étapes décrites précédemment (Chapitre 2, 1.1.5, page 137). Les tampons utilisés sont décrits en Annexe 1.6, page X.

1.2.4 Caractérisation enzymatique des constructions mutées de l'ADN polymérase λ humaine : cinétique et fidélité

Ici, l'objectif a été de caractériser l'activité enzymatique des constructions mutantes de l'ADN polymérase λ humaine, pour comprendre l'impact des mutations réalisées sur l'activité et la fidélité de cette enzyme. Ces constructions ont été testées de la même façon que PolXaΔBRCT et PolXdΔBRCT.

Les mêmes oligonucléotides hybridés que pour les tests de gap-filling présentés précédemment ont été utilisés, à une concentration finale de 200 nM. Cet ADN a été mélangé à 5 nM de chacune des ADN polymérases testées, avec du tampon d'activité (50 mM Tris-HCl pH 8, 10 mM $MgCl_2$, 1 mM DTT, 0,1 mg/mL BSA et 4% glycérol). Du dGTP a été ajouté à des concentrations croissantes allant de 200 nM à 10 µM, et l'ensemble a été incubé à 37°C pendant 5 min. Pour les tests avec les autres dNTPs, ceux-ci ont été ajoutés à des concentrations croissantes allant de 200 nM à 10 mM, et l'ensemble a été incubé à 37°C pendant 1h. Les réactions ont été arrêtées comme précédemment, et déposées sur gel urée-PAGE pour être analysées. L'analyse cinétique a été réalisée comme précédemment (Chapitre 2, 1.1.8, page 146).

1.2.5 Caractérisation structurale des complexes étudiés

1.2.5.1 Cristallisation des complexes protéine-ADN-nucléotide

Figure 108 : Schéma du protocole général utilisé pour obtenir des cristaux des constructions étudiées. L'échantillon est d'abord préparé, en mélangeant les différents partenaires (protéine, ADN et nucléotide), puis est utilisé pour préparer une plaque de conditions de cristallisation en goutte suspendue. Une fois la plaque préparée, elle est incubée à 4°C et fréquemment observée pour surveiller l'apparition de cristaux. Figure créée sur BioRender

1.2.5.1.1 Substrats utilisés

L'ADN utilisé était un substrat de *gap-filling*, avec un brin *template* de 11 nucléotides (nts) (5′- CGGCAGTACTG−3′) hybridé à une amorce « amont » de 6 nts (5′-CAGTAC-3′) et

une amorce « aval » phosphorylée en 5’ de 4 nts (5’-pGCCG-3’). Ces oligonucléotides ont été synthétisés et purifiés (HPLC) par Biomers. Après avoir été suspendus dans du tampon TE (10 mM Tris-HCl pH 7,5, 1 mM EDTA) pour obtenir une concentration de 10 mM, les trois oligonucléotides ont été mélangés au ratio 1 :1 :1 dans un tampon d’hybridation, pour obtenir un complexe concentré à 3,33 mM. Le mélange a été chauffé à 95°C pendant 10 min puis refroidi progressivement (1°C/min) jusqu’à 20°C. Il a ensuite été conservé à -20°C.

Le nucléotide utilisé était du dTTP à 100 mM (Sigma), car la base modèle sur l’ADN est un dA. Dans le cas de la construction *Ptet-like*, l’expérience de cristallisation a également été réalisée avec du dCTP à 100 mM (Sigma).

1.2.5.1.2 Préparation des complexes protéine-ADN-nucléotide

Pour préparer les complexes, chaque protéine a été décongelée à 4°C, centrifugée 10 min à 21 000 g à 4°C pour éliminer d’éventuels agrégats, puis sa concentration a été calculée à partir de son absorbance à 280 nm mesurée au NanoDrop en utilisant des dilutions sérielles. Lorsque leur concentration n’était pas suffisante, certains mutants ont été reconcentrés.

Un mélange de 120 µL a ensuite été préparé, contenant 16 mg/mL de protéine (430 µM), ainsi que deux fois plus d’ADN (860 µM). Ce mélange a été incubé à 4°C pendant 2h, puis 2 mM de dTTP (ou de dCTP) ont été ajoutés, ainsi que 10 mM de $CaCl_2$. Après une autre incubation de 2h, ce mélange a été utilisé pour préparer des gouttes de cristallisation dans différentes conditions.

1.2.5.1.3 Conditions de cristallisation testées

À partir de cette étape, jusqu’à l’affinement des données cristallographiques, toutes les expériences ont été réalisées avec l’aide du Dr Ahmed Haouz et de la plateforme de cristallographie de l’Institut Pasteur. Le principe de la cristallogenèse est détaillé en Annexe 4.1 (page XX).

Les conditions de cristallisation utilisées étaient proches de celles utilisées dans l’article de 2022 (Jamsen *et al.*, 2022) (20 mM bicine pH 7,5, 14–20% PolyPure PEG, 300 mM Na-K tartrate). Cependant, le PolyPure PEG (PolyEthylène Glycol) n’étant pas disponible commercialement, nous avons choisi de modifier le protocole en utilisant 4 PEG de poids moléculaires différents, pour générer une matrice sur une plaque 24 puits (avec 1 mL de solution

par puits). Ces plaques VDX ont été préparées à l'aide d'un robot MatrixMaker (Emerald Biosystems). Dans tous les cas, la cristallisation a été réalisée à 4°C.

Tableau 7 : Première matrice de conditions de cristallisation testée, produite en plaque VDX 24 puits par la Plateforme de Cristallographie de l'Institut Pasteur

	1	2	3	4	5	6
A	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	14%		PEG 600			20%
B	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	14%		PEG 1000			20%
C	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	14%		PEG 10K			20%
D	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	14%		PEG 20K			20%

Dans certains cas, lorsque de nombreux cristaux de petite taille ont été obtenus, il a été nécessaire de préparer une nouvelle matrice, pour optimiser les conditions de cristallisation. Dans ces conditions, des mélanges de PEG de tailles différentes (PEG Smear) ont été utilisés

Tableau 8 : Deuxième matrice de conditions de cristallisation testée pour l'optimisation de la cristallisation de certaines constructions, produite en plaque VDX 24 puits par la Plateforme de Cristallographie de l'Institut Pasteur

	1	2	3	4	5	6
A	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	10%		PEG Smear	Low		22,5%
B	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	10%		PEG Smear	Medium		22,5%
C	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	10%		PEG Smear	High		22,5%
D	20 mM bicine pH 7,5 300 mM Na-K tartrate					
	10%		PEG Smear	Broad2		22,5%

Enfin, dans le cas du mutant λmutK, de nombreux cristaux de petite taille ont été obtenus dans ces conditions, mais n'ont pas permis d'atteindre une résolution satisfaisante (7 Å). Il a donc été décidé de les reproduire, en ajoutant à la meilleure condition (20 mM bicine pH 7,5, 300 mM Na-K tartrate, 17,5% PEG 20K) des additifs. Ce screening a été réalisé par la plateforme de cristallographie de l'Institut Pasteur. Le kit utilisé était le kit Additive Screen HT (Hampton Research). Afin de réduire le nombre de cristaux pouvant apparaitre, la concentration de protéine a été abaissée à 10 mg/mL, et la concentration d'ADN a été adaptée en conséquence.

1.2.5.1.4 Préparation des gouttes

Pour lancer les expériences de cristallogenèse, les complexes préparés ont ensuite été mélangés aux conditions de chaque puits, à différents ratios de volumes [complexe : solution de cristallisation] (1 µL:1 µL, 1:2, 2:1). Ces mélanges ont été déposés sur des lamelles de verre, et scellés sur les puits correspondants, pour une cristallisation par échange de vapeur en goutte suspendue. Les plaques VDX 24 puits ont été conservées à 4°C, et analysées régulièrement pour observer l'apparition de cristaux.

Dans le cas de l'optimisation de conditions pour la cristallisation du mutant λmutK, une plaque 96 puits a été préparée, avec des gouttes de 400 nL (200 nL de préparation de protéine, 200 nL de solution précipitante), en goutte assise dans une plaque Greiner 96 puits, à l'aide d'un robot Genesis Workstation 150 (Tecan).

1.2.5.2 Collecte et intégration des données de diffraction

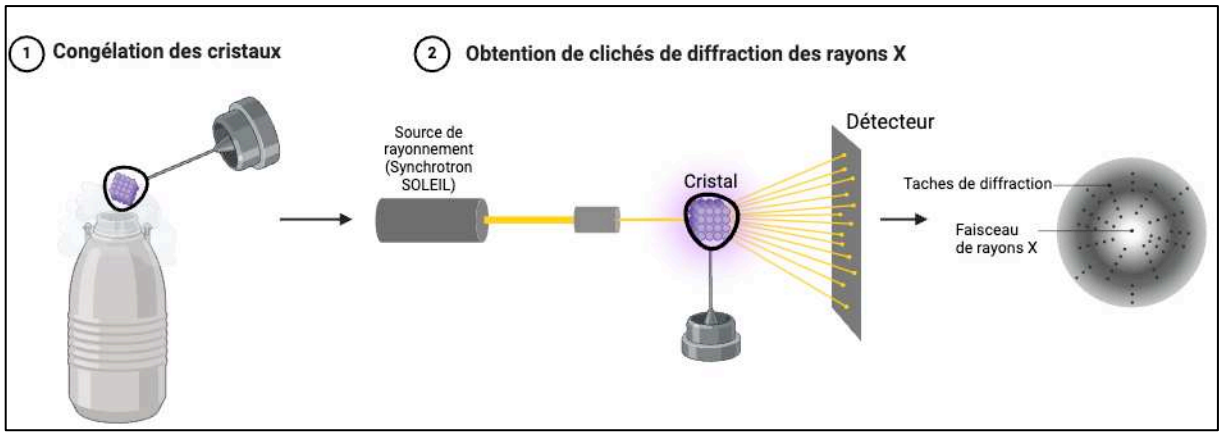


Figure 109 : Schéma global du protocole utilisé pour obtenir les clichés de diffraction des rayons X pour les protéines étudiées. Une fois les cristaux obtenus, ils sont congelés dans l'azote liquide après trempage dans une solution protectrice, et sont amenés au synchrotron SOLEIL. Le rayonnement synchrotron est utilisé pour émettre un faisceau de rayons X qui est projeté sur le cristal. Des clichés de diffraction sont obtenus et enregistrés par le détecteur. Figure créée sur BioRender

Une fois des cristaux obtenus, ils ont été récupérés si leur taille le permettait. Cette étape très délicate a été réalisée par Ahmed Haouz, avec des boucles de nylon de différentes tailles, selon la taille des cristaux. Une fois chaque cristal récupéré, il a été « lavé » avec un passage dans sa solution de cristallisation, puis passé dans une solution cryoprotectrice (25% éthylène glycol pour tous les cristaux sauf ceux obtenus pour λmutK après le criblage d'additifs qui ont été congelés avec 25% de glycérol) et congelé dans de l'azote liquide (à -195°C).

Les cristaux congelés ont été analysés sur les lignes Proxima-1 et Proxima-2 du Synchrotron SOLEIL (St Aubin). Ils ont été placés par un robot sous un flux d'azote gazeux, et centrés dans le faisceau en faisant varier les positions de la tête goniométrique servant de support aux boucles. Un faisceau de rayons X avec une longueur d'onde $\lambda = 0{,}980112$ Å a été projeté sur les cristaux, et les clichés de diffraction ont été enregistrés avec un détecteur Eiger (Dectris Ltd.). 3600 clichés de diffraction ont été obtenus, par rotation du cristal d'un angle de 0,1° entre chaque image. Les intensités des tâches de diffractions ont ensuite été intégrées par Ahmed Haouz à l'aide du script *xdsme* basé sur le logiciel XDS (Kabsch, 2010). Ce script permet d'indexer les données en déterminant la maille cristalline, le groupe d'espace (voir Annexe 4.2, page XXI), puis d'intégrer et de moyenner les intensités des facteurs de structures, conduisant aux statistiques de collection du jeu de données.

Il était intéressant ici de connaître le groupe d'espace des cristaux obtenus. En effet, les expériences réalisées se basaient sur d'autres et n'incluaient que de petites modifications : théoriquement, on s'attend donc à obtenir une organisation similaire des molécules au sein du cristal. Si ce n'est pas le cas, cela peut supposer des changements d'organisation des protéines dans l'espace.

Les données obtenues avec certains cristaux étaient anisotropes et ont été traitées avec le serveur Staraniso par Pierre Legrand. Une fois les données brutes traitées, il a été possible de les utiliser pour résoudre les structures des complexes.

1.2.5.3 Construction des modèles structuraux

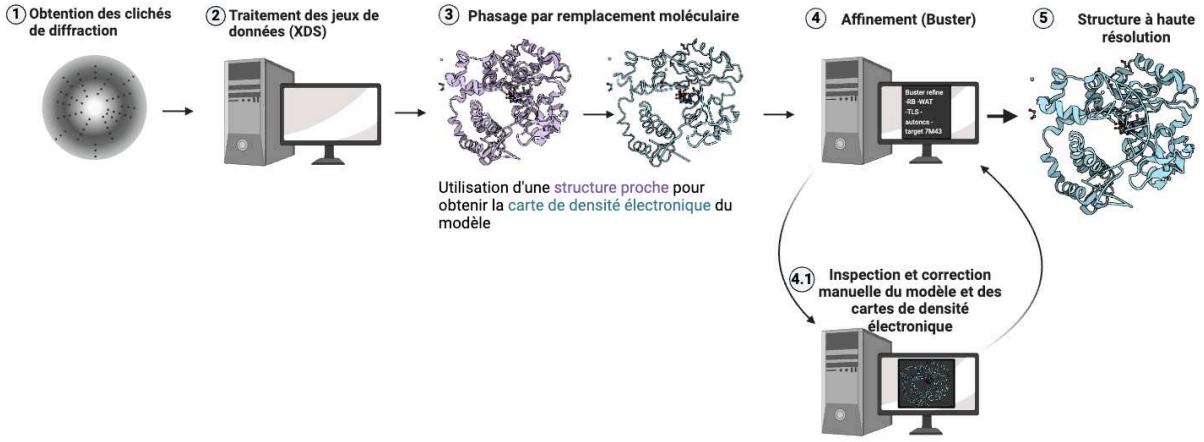


Figure 110 : Schéma général du protocole utilisé pour l'obtention des modèles structuraux des protéines étudiées. Une fois les clichés de diffraction obtenus, ils sont traités à l'aide du logiciel XDS, puis le phasage est réalisé par remplacement moléculaire, ce qui permet d'obtenir un modèle initial et une carte de densité électronique. Le modèle est ensuite affiné via un processus itératif : le logiciel Buster permet au modèle de coïncider le mieux possible avec les données expérimentales (la carte de densité électronique), et le modèle produit est ensuite corrigé manuellement. Lorsque le modèle obtenu coïncide avec les données expérimentales, le modèle final est obtenu. Figure créée sur BioRender

1.2.5.3.1 Phasage par remplacement moléculaire

Pour obtenir une carte de densité électronique correspondant aux clichés de diffractions obtenus, il faut obtenir la phase des réflexions (voir Annexe 4.3, page XXI). Ici, les expériences réalisées visent à obtenir les structures de mutants, et sont basées sur des expériences ayant déjà permis d'obtenir la structure d'une version sauvage de l'ADN polymérase λ. Par conséquent, la structure d'une protéine très proche existe (PDB 7M43), et a donc été utilisée pour réaliser un remplacement moléculaire. Le remplacement moléculaire a été réalisé avec le logiciel Phaser (Evans, 2006), après détermination du coefficient de Matthews pour connaitre le nombre de molécules par unité asymétrique (voir Annexe 4.4, page XXII). À l'issue de cette étape, une carte de densité est obtenue et le modèle initial doit être affiné pour correspondre autant que possible aux données expérimentales.

1.2.5.3.2 Affinement

Ici, tous les affinements ont été réalisés à l'aide du logiciel Buster (Bricogne G. *et al.* (2017). BUSTER version 2.10.4. Cambridge, United Kingdom: Global Phasing Ltd.).

Le rôle de ce logiciel est de calculer des facteurs de structure à partir du modèle (F_{calc}) et de les comparer aux facteurs de structure observés dans les données expérimentales (F_{obs}). Il fait cela de manière itérative, en modifiant le modèle initial pour diminuer la différence entre

les deux. Les modifications apportées doivent respecter des contraintes de géométrie protéique et chimique.

Pour les affinements réalisés ici, plusieurs fonctions de Buster ont été utilisées (leurs rôles sont détaillés en Annexe 4.5, page XXII) : l'affinement *Rigid-Body*, l'utilisation des informations de symétries non cristallines, l'utilisation de tenseurs TLS (*Translation Rotation and Screw rotation*), l'utilisation d'une cible d'affinement, et le placement de molécules d'eau.

À chaque cycle, le logiciel calcule deux cartes de densité électronique : $2F_{obs}$-F_{calc} (densité positive indiquant par où le modèle doit passer) et F_{obs}-F_{calc} (densités positives et négatives indiquant les incohérences entre le modèle et les données expérimentales). Ces deux cartes et le modèle sont inspectés en utilisant le logiciel Coot. Le modèle est ajusté manuellement de façon itérative. La conformation de la chaine principale doit correspondre à des angles ϕ et φ énergétiquement favorables pour tous les résidus, en accord avec le diagramme de Ramachandran (Ramachandran *et al.*, 1963).

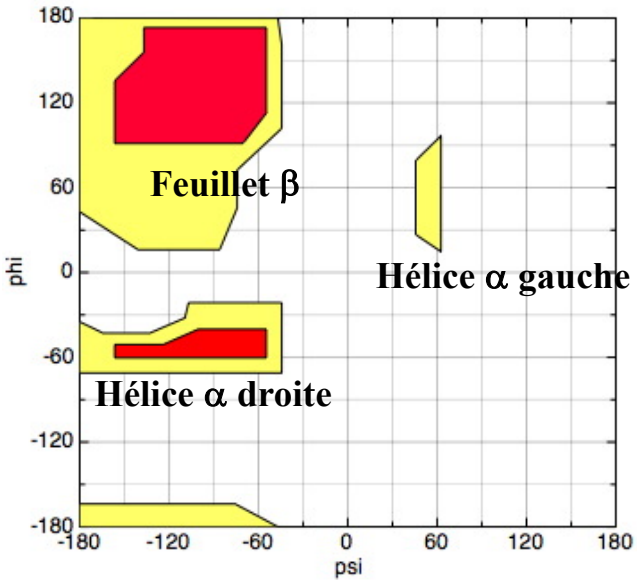


Figure 111 : Diagramme de Ramachandran. D'après https://www.cryst.bbk.ac.uk/PPS95/course/3_geometry/rama.html

Après chaque cycle d'affinement et de traitement du modèle à la main, un autre peut être commencé, et ainsi de suite jusqu'à obtenir un modèle satisfaisant. L'accord entre modèle obtenu et données expérimentales est indiqué par deux valeurs : R_{work} et R_{free} (voir Annexe 4.6, page XXIII).

Lorsque le modèle est en accord avec les données expérimentales, les valeurs de R_{work} et R_{free} baissent. La valeur de R_{free} est calculée sur un ensemble de réflexions qui ne sont pas

utilisées pour l'affinement ; elle doit diminuer au cours des affinements de la même façon que R_{work}, mais l'écart entre ces deux valeurs doit être aussi faible que possible (<5%). On considère que l'affinement est correct quand les valeurs de R_{work} et R_{free} sont les plus basses possibles, et correspondent aux valeurs obtenues pour d'autres protéines de la même gamme de résolution (Read *et al.*, 2011).

1.2.5.4 Analyse des modèles structuraux obtenus

Une fois obtenues, les structures cristallographiques affinées ont été visualisées sur le logiciel UCSF ChimeraX (version 1.6.dev202301310903). Elles ont également pu être comparées, et les figures associées ont été créées sur ce logiciel.

1.3 Production, purification et étude enzymatique de constructions mutantes de l'ADN polymérase X a de *Paramecium tetraurelia* et de l'ADN polymérase λ humaine

1.3.1 Constructions mutantes étudiées

1.3.1.1 Construction mutante de l'ADN polymérase X a de Paramecium tetraurelia

Cette construction est basée sur la construction PolXaΔBRCT étudiée précédemment. Dans cette nouvelle construction appelée PolXaΔBRCT-Loop3β, la séquence de la boucle 3 a été remplacée par son équivalent chez l'ADN polymérase β. Les résidus 581 à 588 ont donc été remplacés par la séquence GVA. Cette construction a un poids moléculaire de 44262 Da en présence de son tag de 14 histidines. Cette étiquette n'a jamais été supprimée dans ces travaux.

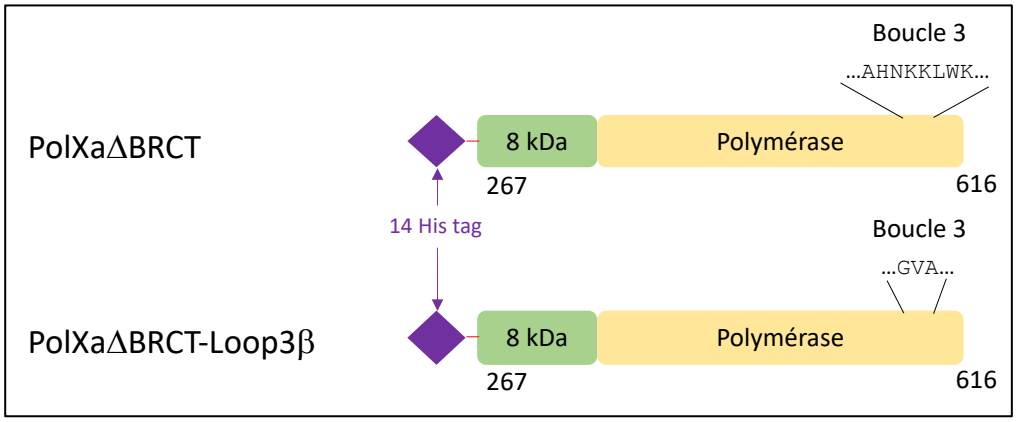


Figure 112 : Comparaison de la construction PolXaΔBRCT et de la construction PolXaΔBRCT-Loop3β. Le losange violet indique l'étiquette de 14 histidines. Les domaines 8 kDa et polymérase sont indiqués respectivement dans des rectangles vert et jaune. Les résidus inclus dans chaque construction sont indiqués sous la construction associée. La séquence de la boucle 3 est indiquée pour chaque construction.

1.3.1.2 Construction mutante de l'ADN polymérase λ humaine

Cette construction est basée sur la construction de l'ADN polymérase λ complète étudiée précédemment (Chapitre 2, 1.1.2.2, page 132). Dans cette nouvelle construction appelée λLoop3β, la séquence de la boucle 3 a été remplacée par son équivalent chez l'ADN polymérase β. Ainsi, les résidus 539 à 547 ont été remplacés par la séquence GVA. Cette construction a un poids moléculaire de 67402 Da en présence de son étiquette de 14 histidines. Cette étiquette n'a jamais été supprimée dans ces travaux.

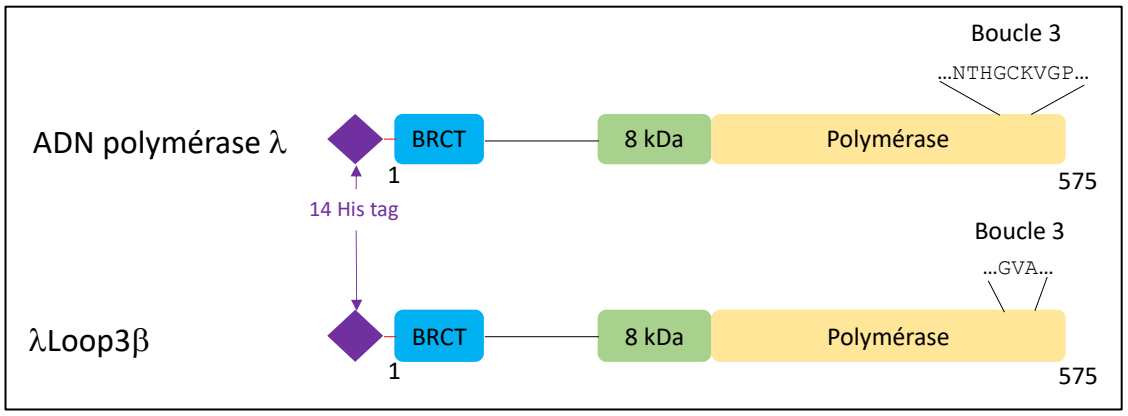


Figure 113 : Comparaison de la construction initiale de l'ADN polymérase λ et de la construction λLoop3β. Le losange violet indique l'étiquette de 14 histidines. Les domaines 8 kDa et polymérase sont indiqués respectivement dans des rectangles vert et jaune. Les résidus inclus dans chaque construction sont indiqués sous la construction associée. La séquence de la boucle 3 est indiquée pour chaque construction.

1.3.2 Préparation des plasmides d'expression

Pour obtenir les plasmides permettant d'exprimer ces constructions mutantes, des mutagenèses dirigées ont été réalisées de la même manière que précédemment (Chapitre 2, 1.2.2, page 154). Pour chaque mutant, une PCR a été réalisée en utilisant des amorces s'hybridant de part et d'autre de la séquence à éliminer sur un plasmide parental (LS05-PolXaΔBRCT et LS05-ADN polymérase λ). Une fois la taille de l'ADN amplifié contrôlée par électrophorèse en gel d'agarose 1%, le mélange d'enzymes KLD a été utilisé pour circulariser les plasmides, qui ont ensuite été surproduits par des bactéries *E. coli* DH5α, purifiés et séquencés. Les plasmides portant la mutation ont été conservés à -20°C.

1.3.3 Production et purification des constructions mutées

Pour produire en système bactérien de grandes quantités des protéines étudiées, le même protocole que précédemment a été utilisé (Chapitre 2, 1.1.4, page 135).

Les purifications de ces constructions ont également été réalisées suivant les mêmes étapes que décrites précédemment (Chapitre 2, 1.1.5, page 137). Les tampons utilisés sont décrits en Annexe 1.6 (page X).

1.3.4 Comparaison de la fidélité des ADN polymérases X mutées avec les constructions non mutées

L'objectif ici a été de déterminer si les mutations réalisées sur la boucle 3 avaient un impact sur la fidélité de l'ADN polymérases X a et de l'ADN polymérase λ. Les constructions mutées ont été testées et comparées aux constructions sauvages (Chapitre 2, 1.1.9, page 148).

120 nM des quatre ADN polymérases ont été mélangés à 30 nM d'ADN et 120 µM de chaque nucléotide incorrect. Le tampon utilisé contenait 50 mM de Tris-HCl pH 8,4, 2% de glycérol, 5 mM de $MgCl_2$, 100 mM de NaCl, 5 mM de DTT, 0,1 mg/mL de BSA, et 0,1 µM d'EDTA. L'ensemble a été incubé 50 minutes à 37°C pour les constructions de l'ADN polymérase λ humaine, et à 27°C pour les ADN polymérases X de *Paramecium tetraurelia*. Les échantillons ont ensuite été déposés sur gel urée-PAGE et analysés comme précédemment.

2 Résultats

2.1 Caractérisation biochimique et enzymatique des ADN polymérases de la famille X de *Parамecium tetraurelia*

Dans le but de caractériser les ADN polymérases X de *Paramecium tetraurelia*, celles-ci ont été produites et purifiées, ainsi que les ADN polymérases β et λ humaines. Une caractérisation enzymatique a été réalisée, pour déterminer dans quels contextes ces enzymes sont actives. Certaines ADN polymérases X ont d'autres activités, comme les ADN polymérases β et λ qui portent une activité dRP lyase. Cette activité a été testée pour la construction PolXaΔBRCT, puis deux constructions (PolXaΔBRCT et PolXdΔBRCT) ont été caractérisées du point de vue cinétique. Enfin, la fidélité des ADN polymérases X de *Paramecium tetraurelia* a été comparée à celle de l'ADN polymérase λ, pour répondre à la première question de ces travaux : ces ADN polymérases sont-elles particulièrement fidèles ?

2.1.1 Expression et purification des ADN polymérases X de *Paramecium tetraurelia* et *Homo sapiens* étudiées

Les tampons utilisés dans les purifications sont détaillés en Annexe 1.6, page X.

2.1.1.1 PolXdFL

Le gène codant pour PolXdFL a été cloné dans le plasmide LS05 par la méthode de restriction-ligation, de manière à pouvoir exprimer la protéine en système bactérien. La production de cette construction a longtemps été difficile : les bactéries produisaient peu de protéine, et de façon peu soluble. Bien que le séquençage du plasmide utilisé (couvrant la séquence codant pour la protéine) fût correct, il a été décidé de refaire le clonage des gènes d'intérêt dans le plasmide LS05. En effet, il est possible que des éléments importants pour l'expression protéique aient été altérés ou manquants. Ce second clonage a été réalisé avec la méthode de *Gibson assembly*, ce qui a permis d'obtenir un plasmide fonctionnel pour l'expression de cette protéine.

Une fois le plasmide obtenu et sa séquence confirmée, PolXdFL a été produite en bactéries *E. coli* BL21star(DE3), en induisant sa production à 20°C avec 1 mM d'IPTG.

La purification a ensuite été réalisée plusieurs fois, en suivant le protocole global énoncé en Matériel et Méthodes (page 137). Je présente ici les résultats d'une purification réalisée à partir d'une production de 5L de culture bactérienne. La lyse bactérienne a été réalisée au CellD, et après centrifugation la fraction soluble a été chargée sur une colonne HisTrap équilibrée avec du tampon A, puis l'élution a été réalisée avec un gradient de tampon B.

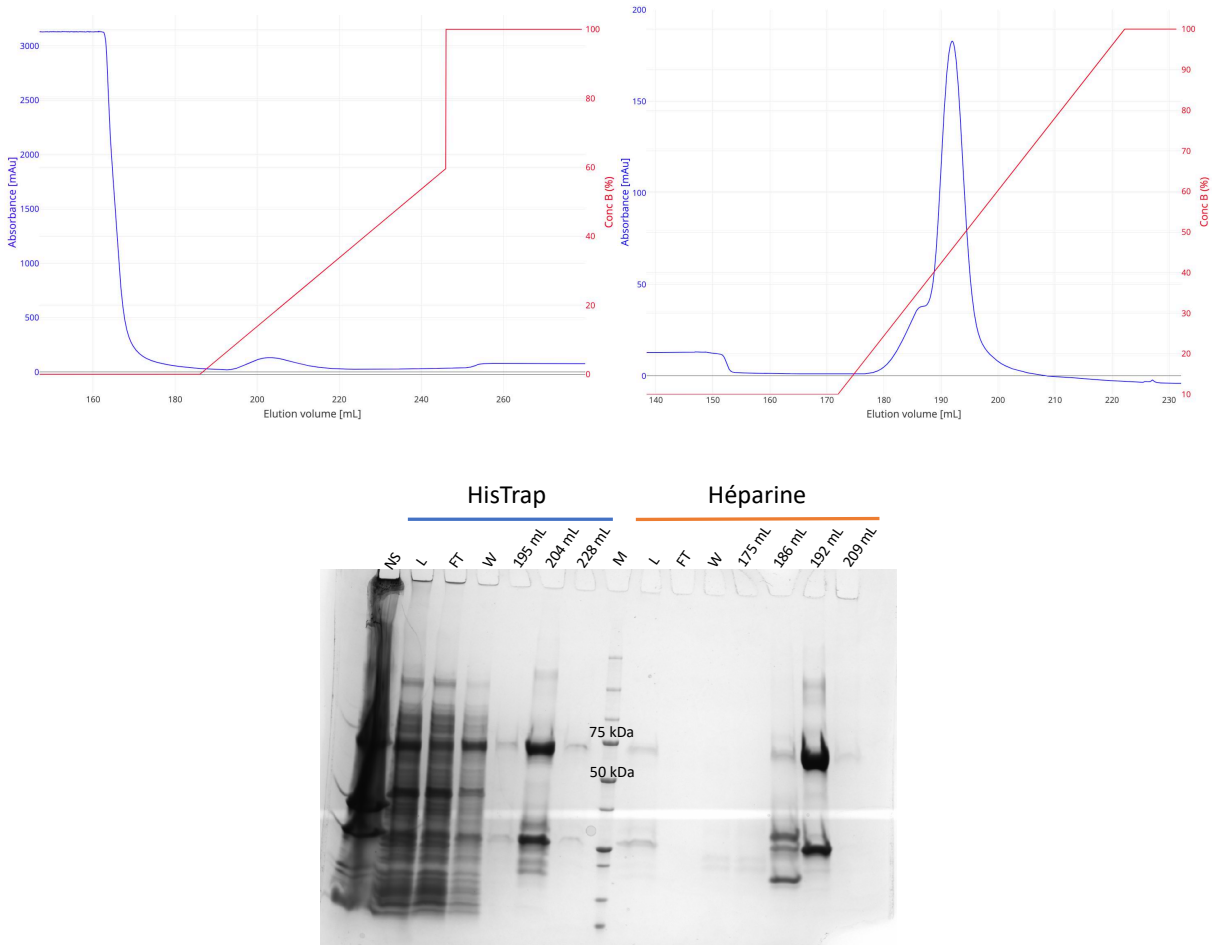


Figure 114 : Les deux premières étapes de la purification de PolXdFL. En haut à gauche : Chromatogramme de l'étape HisTrap de la purification de PolXdFL (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. En haut à droite : Chromatogramme de l'étape Héparine de la purification de PolXdFL (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. En bas : SDS-PAGE. Les fractions liées à l'étape Histrap sont indiquées par un trait bleu ; celles de l'Héparine par un trait orange ; M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 75 et 50 kDa sont indiquées ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Les fractions allant de 195 mL à 227 mL ont été rassemblées et diluées avec du tampon C (voir Annexe 1. 6, page X), et cet échantillon a été chargé sur une colonne Héparine équilibrée avec du tampon D. L'élution a été réalisée avec un gradient de tampon E. Une analyse SDS-PAGE a été réalisée avec des fractions issues de ces deux étapes, montrant une bande

correspondant au poids moléculaire attendu pour PolXdFL (75 kDa environ). La protéine d'intérêt était aussi présente dans la fraction de lavage (W), indiquant qu'elle s'était mal fixée à la résine dans ces conditions. Cependant, la majorité de la protéine d'intérêt était présente dans les fractions éluées correspondant aux pics visibles sur les chromatogrammes associés. Ces fractions montraient aussi la présence de protéines contaminantes. À l'issue de la chromatographie sur résine Héparine, les fractions de 178 mL à 208 mL ont été rassemblées, et l'ensemble montrait une absorbance à 280 nm (A_{280nm}) de 0,353. L'échantillon a été dialysé en présence de protéase TEV afin de supprimer l'étiquette de 14 histidines. Le lendemain, l'échantillon a été injecté sur la colonne His-Trap, et les fractions ont été déposées sur gel SDS-PAGE. L'ajout de la protéase semble bien avoir permis la suppression de l'étiquette, comme l'indique la légère baisse de poids moléculaire observée. Les fractions *Flow-Through* (FT) et *Wash* (W) contenant les protéines non fixées à la résine présentaient bien une bande correspondant à la protéine sans son étiquette, et l'éluat obtenu avec 500 mM d'imidazole ne présentait que les bandes correspondant aux protéines contaminantes.

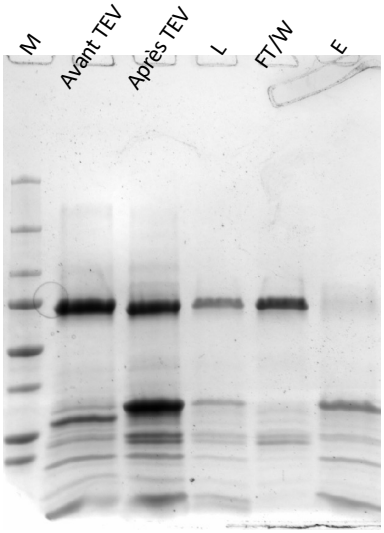


Figure 115 : SDS-PAGE de l'étape de purification à la suite du clivage du tag de PolXdFL par la protéase TEV. M : Marqueur de poids moléculaire ; Les fractions correspondant à l'échantillon avant et après l'incubation avec la TEV sont indiquées ; L : load, fraction soluble chargée sur la résine ; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; E : élution de la résine avec 500 mM d'imidazole.

Les fractions contenant la protéine ont été concentrées, jusqu'à un volume final de 5 mL à une absorbance de 2,076 à 280 nm, soit 2,17 mg/mL. L'échantillon a été aliquoté et congelé, puis stocké à -20°C. La pureté de l'échantillon était suffisante pour des tests d'activité.

2.1.1.2 PolXaΔBRCT

Une fois le plasmide LS05-PolXaΔBRCT obtenu par la méthode de *Gibson assembly* et sa séquence confirmée, la protéine a été produite en bactéries *E. coli* BL21star(DE3).

La purification a ensuite été réalisée plusieurs fois, en suivant le protocole global énoncé en Matériel et Méthodes. Je présente ici les résultats d'une purification réalisée à partir d'une production de 8L de culture bactérienne. La lyse bactérienne a été réalisée au sonicateur, et l'échantillon a suivi les mêmes étapes de purifications que PolXdFL.

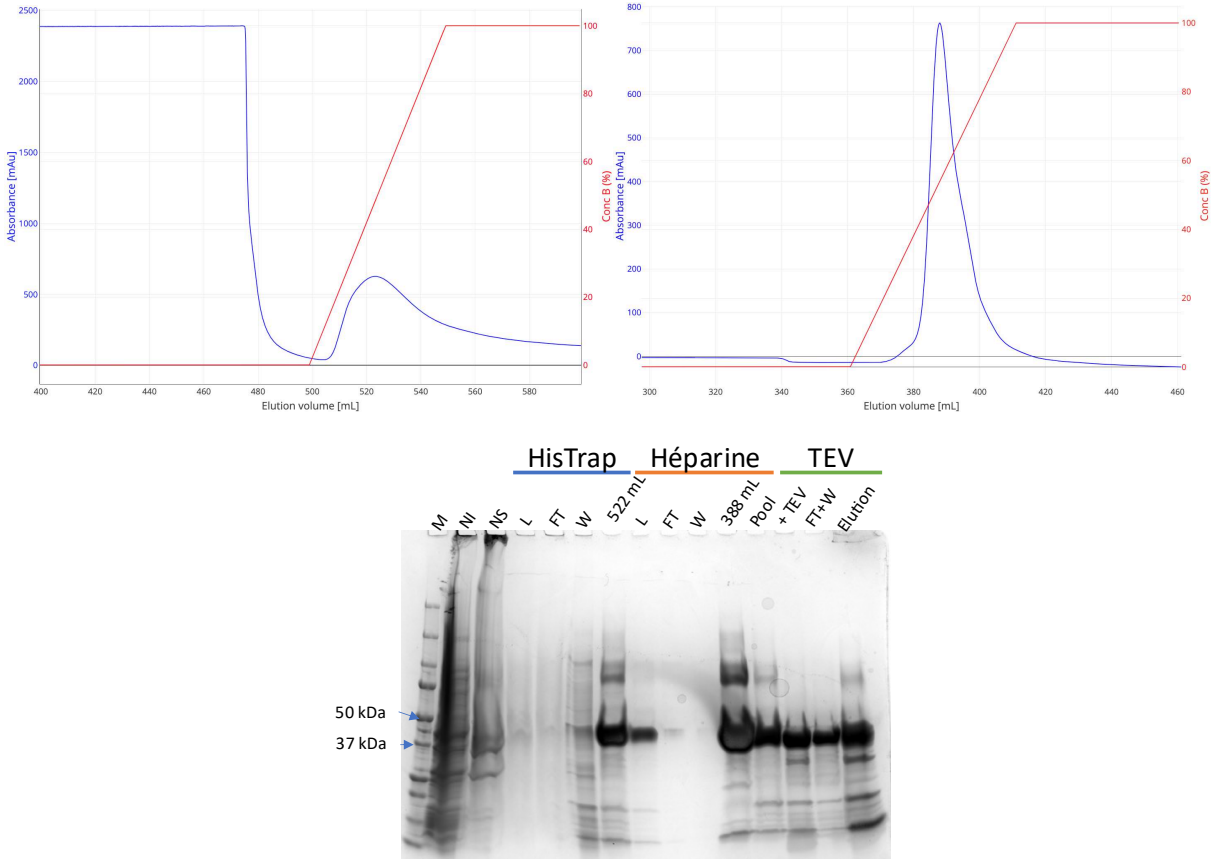


Figure 116 : Trois premières étapes de purification de PolXaΔBRCT. En haut à gauche : Chromatogramme de l'étape HisTrap de la purification de PolXaΔBRCT (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. En haut à droite : Chromatogramme de l'étape Héparine de la purification de PolXaΔBRCT (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. EN bas : SDS-PAGE. Les fractions liées à l'étape Histrap sont indiquées par un trait bleu ; celles de l'Héparine par un trait orange ; celles de l'étape de clivage du tag par la TEV sont en vert. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; NI : bactéries avant induction de la production de protéines à l'IPTG ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée ; Pool : mélange des fractions rassemblées à la fin de la chromatographie Héparine ; +TEV : mélange de protéines après dialyse avec la protéase TEV.

À l'issue de la chromatographie d'affinité His-Trap, les fractions allant de 507 mL à 560 mL ont été rassemblées et diluées avec du tampon C. Cet échantillon de 350 mL avait une

absorbance à 280 nm de 0,338, et un rapport A_{260nm}/A_{280nm} de 0,74, n'indiquant pas de contamination par des acides nucléiques. Il a été chargé sur colonne Héparine et élué avec un gradient de NaCl. Les fractions de 377 mL à 420 mL ont été rassemblées, et l'ensemble montrait une A_{280nm} de 1,133, avec un rapport A_{260nm}/A_{280nm} de 0,60. Cet échantillon a été dialysé à 4°C, après y avoir ajouté 4 aliquots de 100 µL de protéase TEV à 10 mg/mL environ. Le lendemain, l'échantillon a été récupéré et séparé sur HisTrap. Une analyse SDS-PAGE a été réalisée sur des fractions des trois étapes précédentes, et montrait une bande majoritaire correspondant à la masse moléculaire attendue avant l'ajout de la protéase (45 kDa environ), ainsi qu'une bande de plus faible poids moléculaire après l'incubation avec cette protéase (40,7 kDa). Cela indique que la protéase a bien clivé la protéine de fusion et éliminé le tag. Cependant, une bande correspondant à la masse de la protéine était aussi visible dans la fraction éluée de la His-Trap après incubation avec la protéase, signe que ce clivage n'avait pas été totalement efficace.

Les fractions contenant la protéine séparée de son tag ont été concentrées jusqu'à un volume de 4 mL, avec une absorbance à 280 nm de 7,213. Ces 4 mL ont été injectés sur une colonne S200 16/600 PG, et les fractions séparées ont été récoltées et analysées par SDS-PAGE.

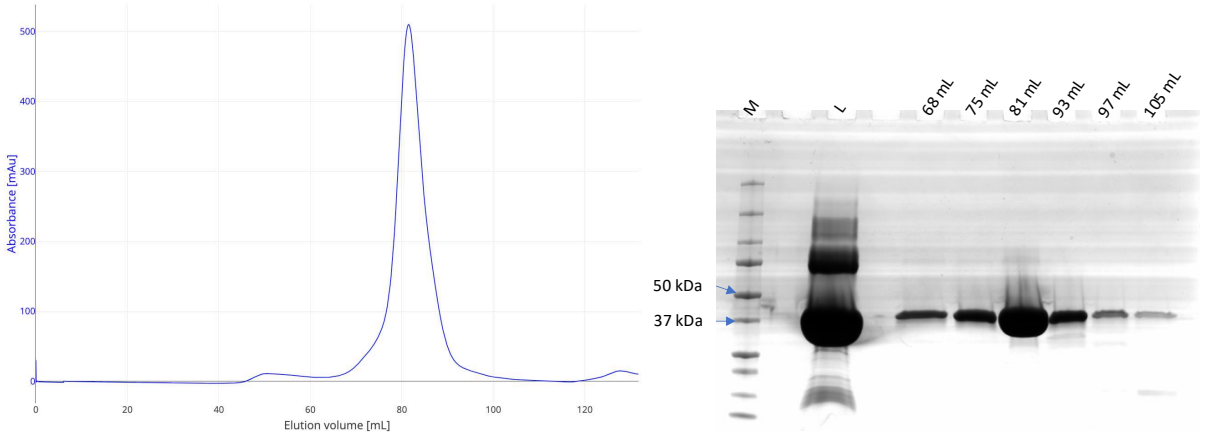
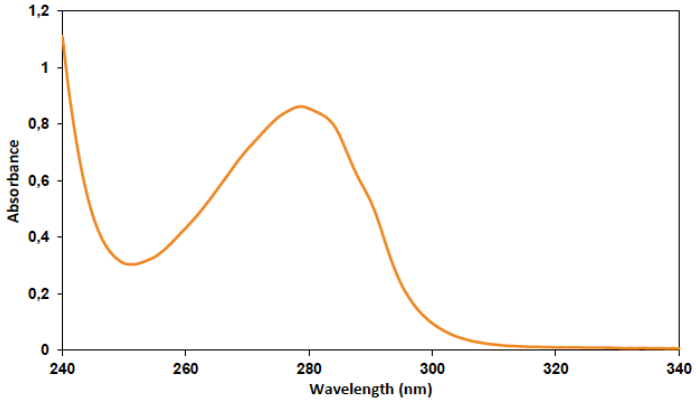


Figure 117 : Étape de chromatographie par exclusion de taille de la purification de PolXaΔBRCT. À gauche : Chromatogramme. L'absorbance à 280 nm (en mUA) est indiquée en bleu. SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; L : load, fraction soluble chargée sur la résine; Les volumes indiquent des fractions choisies d'après le chromatogramme.

Les fractions les plus pures, ont été rassemblées et concentrées, jusqu'à un volume final de 950 µL à une absorbance de 10,969 à 280 nm, soit 10,21 mg/mL. L'échantillon a été aliquoté et congelé, puis stocké à -20°C.

Suite à une autre purification de similaire de PolXaΔBRCT, un contrôle qualité de la protéine a été réalisé par la plateforme de biophysique de l'Institut Pasteur (PFBMI). Le spectre d'absorption, obtenu avec un échantillon dilué au 1/10^{e}, montre bien l'absence de

contamination aux acides nucléiques (A_{260nm}/A_{280nm} = 0,51) ainsi que l'absence d'agrégats ($A_{340nm} \approx 0$). La mesure par spectrométrie de masse montre deux types d'ions : monochargés, donnant une mesure m/z de 40797,29 (soit un poids moléculaire de 40,797 kDa); et doublement chargés, donnant un m/z de 20430,827 (soit un poids moléculaire de 40,862 kDa). La masse attendue étant de 40,798 Da, on peut considérer que la protéine obtenue est bien présente et complète. Enfin, la mesure en DLS de l'échantillon montre une population très majoritaire (>99%), ce qui indique que la protéine est bien homogène en solution dans le tampon choisi, et a un rayon hydrodynamique d'environ 3,2 nm.



Abs 260 nm	Abs 280 nm	Abs 340 nm
0,433	0,855	0,007

Figure 118 : Spectre d'absorption de l'échantillon de PolXaΔBRCT entre 240 et 340 nm (les valeurs d'absorbance à 260 nm, 280 nm et 340 nm sont indiquées en dessous)

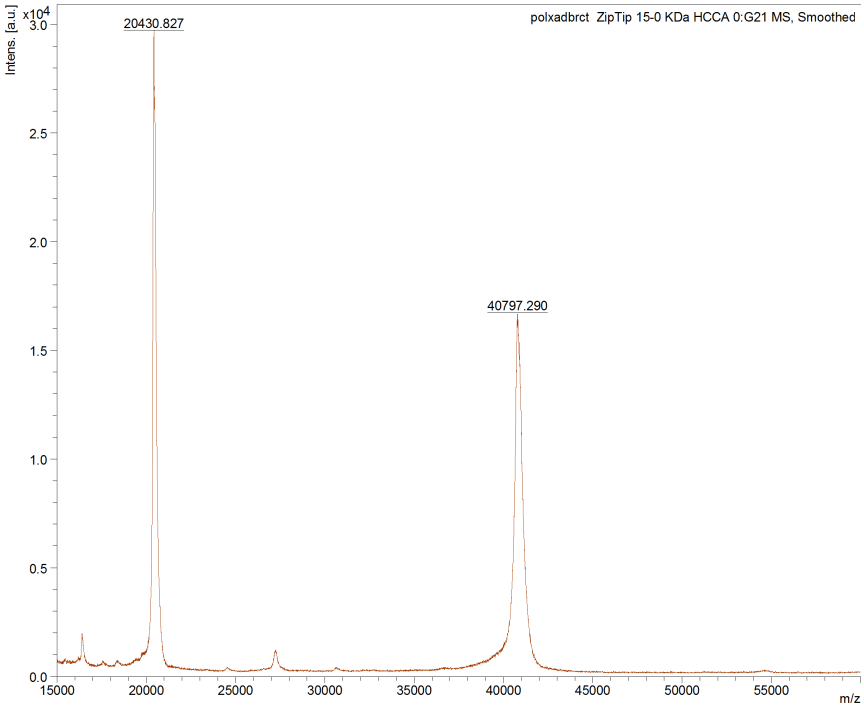
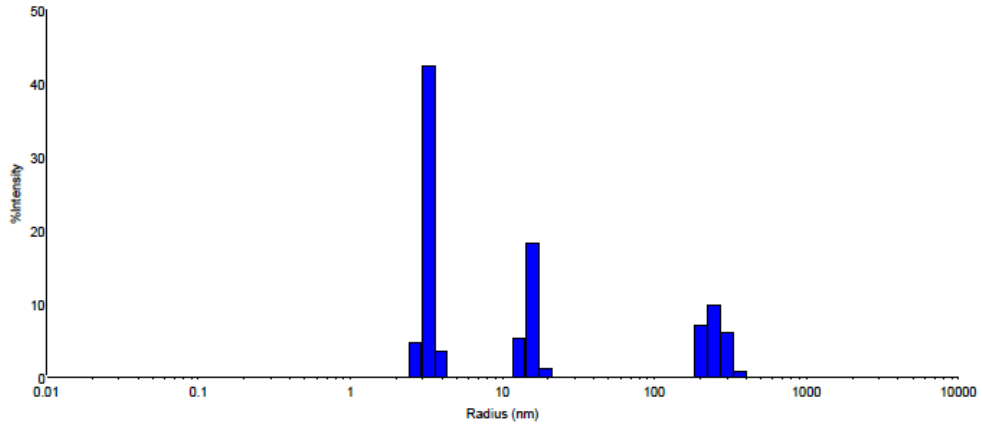


Figure 119 : Spectre de masse obtenu par MALDI-TOF de l'échantillon de PolXaΔBRCT



	Rayon (nm)	*Poids Moléculaire (kDa)*	*% d'intensité*	*% de masse*
Pic 1	3,2	52,9	50,9	99,2
Pic 2	15,2	1972,8	25	0,5
Pic 3	249,9	1372342,4	24,1	0,3

Figure 120 : Mesure de la dispersité de l'échantillon de PolXaΔBRCT. Le tableau en dessous indique les valeurs obtenues de rayon hydrodynamique, de poids moléculaire (approximatif), d'intensité, et de masse pour chaque espèce.

2.1.1.3 PolXbΔBRCT

Le plasmide LS05 permettant d'exprimer PolXbΔBRCT a été produit de la même façon que pour PolXaΔBRCT, puis a été utilisé pour produire la protéine en bactéries *E. coli* BL21star(DE3), de la même manière.

Je présente ici les résultats d'une purification réalisée à partir d'une production de 4L de culture bactérienne. Après la lyse des bactéries et une centrifugation, la fraction soluble a été chargée sur une colonne His-Trap. Des essais ont montré que lors de cette étape, l'absorbance à 260 nm restait très élevée (rapport A_{260nm}/A_{280nm} = 1,71), indiquant une forte contamination aux acides nucléiques. Il a donc été décidé d'utiliser un tampon de lavage (Annexe 1.6, page X) pour éliminer l'ADN (Schneider *et al.*, 2023).

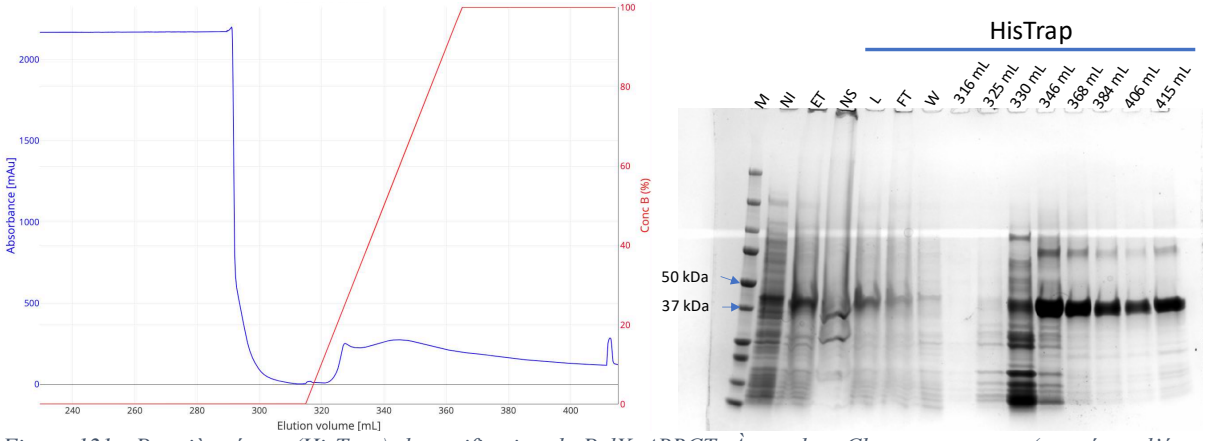


Figure 121 : Première étape (HisTrap) de purification de PolXaΔBRCT. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. À droite : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; NI : bactéries avant induction de la production de protéines à l'IPTG ; ET : extrait total, obtenu après lyse bactérienne et avant séparation des fractions solubles et insolubles ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

L'analyse SDS-PAGE a été réalisée avec les fractions de cette chromatographie, qui ont montré une bande correspondant à la masse attendue (45 kDa environ). Le pic en début d'élution (vers 325 mL) ne contenait pas de protéine, mais probablement des acides nucléiques. Le pic obtenu à partir de 330 mL environ contenait quant à lui bien la protéine d'intérêt. Les fractions allant de 336 mL à 417 mL ont ensuite été rassemblées, et diluées avant d'être chargées sur une colonne Héparine. Cet échantillon de 600 mL avait une absorbance à 280 nm de 0,19, et un rapport A_{260nm}/A_{280nm} de 0,94, n'indiquant pas de contamination aux acides nucléiques.

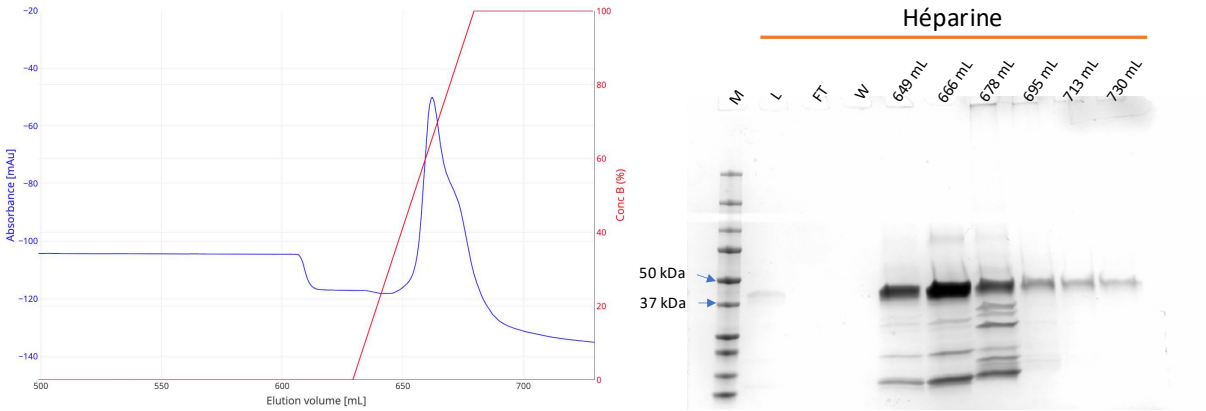


Figure 122 : Deuxième étape (Héparine) de purification de PolXaΔBRCT. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. À droite : : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Une analyse SDS-PAGE a été réalisée sur les fractions issues de la chromatographie sur colonne Héparine, et a montré une bande majoritaire correspondant à la masse moléculaire attendue (45 kDa environ). Les fractions de 654 mL à 695 mL ont été rassemblées, et l'ensemble

avait une A_{280nm} de 0,202, avec un rapport A_{260nm}/A_{280nm} de 0,65. Bien que le gel indique la présence de protéines au-delà de 695 mL, l'absorbance à 260 nm était haute dans les fractions suivantes, qui n'ont donc pas été utilisées pour la suite. Les fractions rassemblées ont été concentrées, jusqu'à un volume final de 3 mL à une absorbance de 2,617 à 280 nm, soit 2,38 mg/mL. L'échantillon a été aliquoté et congelé, puis stocké à -20°C.

2.1.1.4 PolXdΔBRCT

Le clonage du gène codant pour PolXdΔBRCT a été réalisé par *Gibson assembly*. Une fois le plasmide obtenu et sa séquence confirmée par séquençage, la protéine a été produite en bactéries *E. coli* BL21 star, en induisant la production de la protéine à 20°C avec 1 mM d'IPTG.

Les résultats présentés ici correspondent à une purification réalisée à partir de 6L de culture bactérienne. La lyse bactérienne a été réalisée au sonicateur, et après centrifugation, la fraction soluble a été chargée sur une colonne HisTrap.

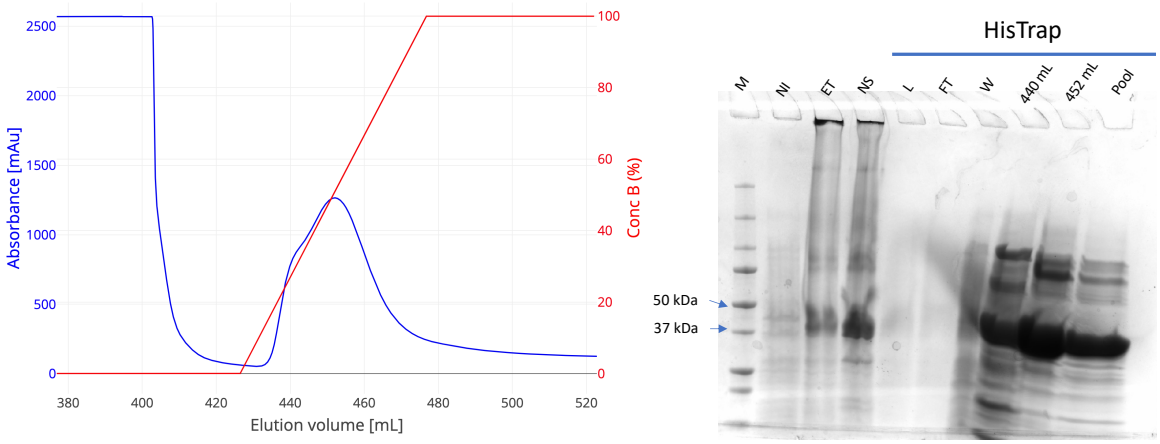


Figure 123 : Première étape (HisTrap) de purification de PolXdΔBRCT. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. À droite : SDS-PAGE. Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; NI : bactéries avant induction de la production de protéines à l'IPTG ;ET : extrait total, ensemble des protéines solubles et insolubles obtenues après la lyse bactérienne ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée ; Pool : mélange des fractions rassemblées à la fin de la chromatographie.

Les fractions éluées ont été analysées par SDS-PAGE, et montraient une bande correspondant à la masse moléculaire attendue (45 kDa environ). Cependant, une bande correspondant possiblement au poids moléculaire de la protéine était aussi présente dans la fraction non soluble, indiquant qu'une petite partie avait été perdue lors de la lyse des bactéries. Les fractions éluées de 432 mL à 481 mL ont été rassemblées et diluées jusqu'à une

concentration en NaCl de 70 mM environ. L'échantillon a été chargé sur une colonne Héparine et l'élution a été réalisée avec un gradient de NaCl.

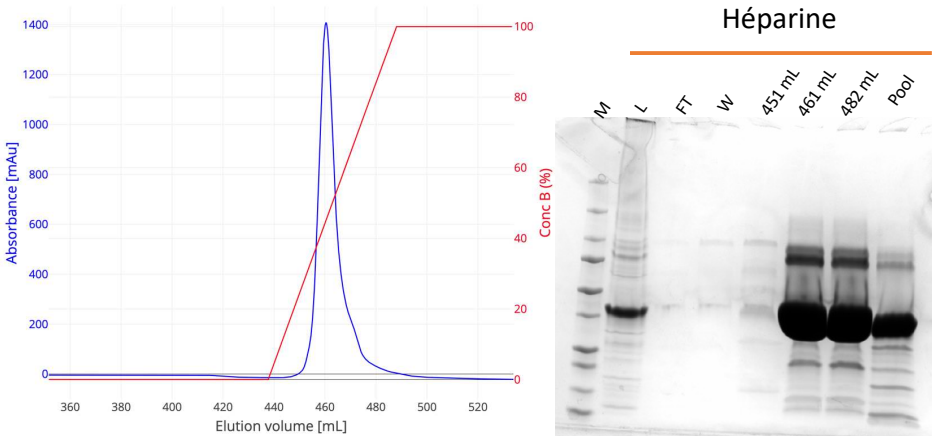


Figure 124 : Deuxième étape (Héparine) de purification de PolXdΔBRCT. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. À droite : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée ; Pool : mélange des fractions rassemblées à la fin de la chromatographie.

L'analyse SDS-PAGE a montré une bande majoritaire correspondant au poids moléculaire attendu pour cette protéine (45 kDa environ). Les fractions de 451 mL à 482 mL ont été rassemblées, et l'ensemble avait une A_{280nm} de 2,845, avec un rapport A_{260nm}/A_{280nm} de 0,56. L'ensemble a été dialysé une nuit après ajout de 100 µL de protéase TEV concentrée à 10 mg/mL. Le lendemain, le dialysat a été séparé sur une colonne His-Trap. Les protéines fixées ont été éluées avec 500 mM d'imidazole.

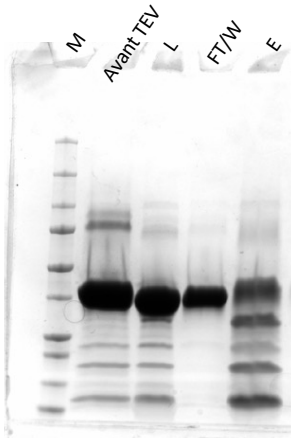


Figure 125 : SDS-PAGE de l'étape de purification à la suite du clivage du tag de PolXdΔBRCT par la protéase TEV. M : Marqueur de poids moléculaire; La fraction correspondant à l'échantillon avant l'incubation avec la TEV est indiquée; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ;E : élution de la résine avec 500 mM d'imidazole.

L'analyse SDS-PAGE indique que l'échantillon chargé sur la colonne présente très majoritairement une protéine d'un poids moléculaire inférieur à l'échantillon avant dialyse, ce

qui confirme que l'étiquette 14-histidines a été coupée. La fraction obtenue avec 500 mM d'imidazole présente cependant une bande correspondant au poids moléculaire de PolXdΔBRCT avec son tag (indiquant que la protéase n'a pas été entièrement efficace), mais également de nombreux contaminants. L'absorbance à 280 nm des fractions contenant la protéine clivée a été mesurée, et était de 1,167. Elles ont donc été concentrées jusqu'à un volume final de 3 mL avec une absorbance de 23,8 à 280 nm environ, soit 20 mg/mL environ. Deux injections de 1,5 mL ont été faites sur une colonne de chromatographie d'exclusion stérique S200 16/600 PG.

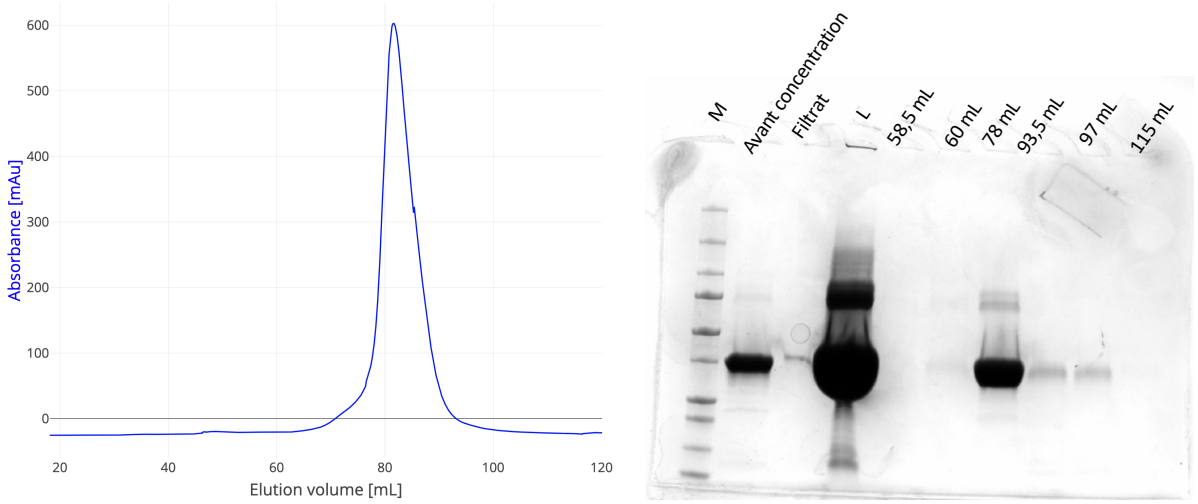


Figure 126 : Étape de chromatographie d'exclusion de taille de la purification de PolXdΔBRCT. À gauche : Chromatogramme. L'absorbance à 280 nm (en mUA) est indiquée en bleu. SDS-PAGE M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; La fraction non concentrée issue du clivage du tag 14 histidines est indiquée « Avant concentration » ; L : load, fraction soluble chargée sur la résine; Les volumes indiquent des fractions choisies d'après le chromatogramme.

Après chaque chromatographie et vérification de leur pureté par SDS-PAGE, les fractions allant de 60 mL à 93,5 mL ont été rassemblées, et avaient une A_{280nm} de 0,69. Elles ont été concentrées jusqu'à une 24,92 mg/mL, 16 aliquots de 100 µL ont été congelés et stockés, et une partie a également été diluée pour congeler 25 aliquots de 20 µL à 4,984 mg/mL.

2.1.1.5 L'ADN polymérase β humaine

Le clonage du gène codant pour l'ADN polymérase β a été réalisé par GenScript. Une fois le plasmide obtenu et sa séquence confirmée, la protéine a été produite en bactéries *E. coli* BL21star(DE3), en induisant sa production à 20°C avec 1 mM d'IPTG.

La purification a ensuite été réalisée, en suivant le protocole global énoncé en Matériel et Méthodes, à partir d'une production de seulement 2L de culture bactérienne. La lyse

bactérienne a été réalisée au sonicateur, et après centrifugation, la fraction soluble a été chargée sur une colonne HisTrap, et les protéines fixées ont été éluées.

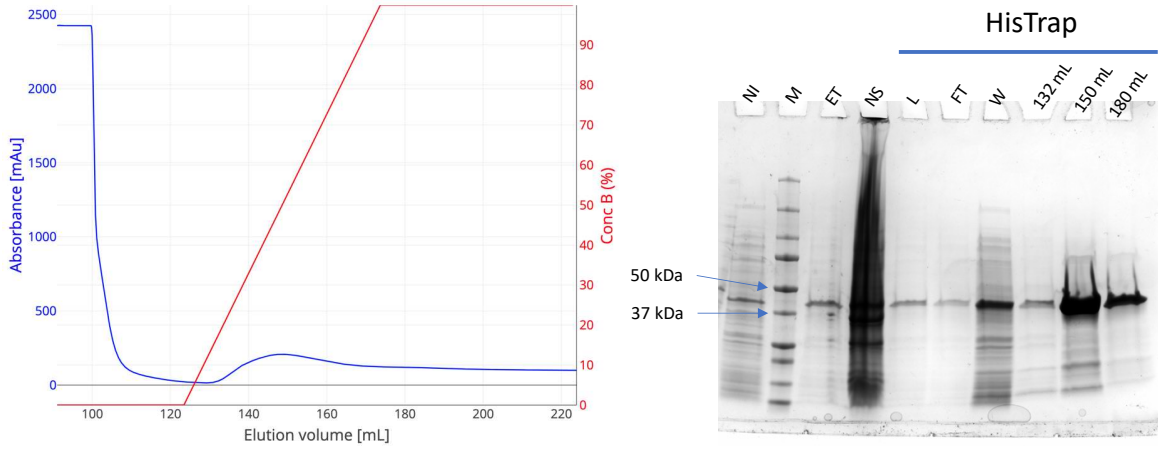


Figure 127 : Première étape (HisTrap) de purification de l'ADN polymérase β. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. À droite : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; NI : bactéries avant induction de la production de protéines à l'IPTG ;ET : extrait total, ensemble des protéines solubles et insolubles obtenues après la lyse bactérienne ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Certaines fractions ont été analysées par SDS-PAGE, et une bande correspondant au poids moléculaire attendu (43 kDa environ) était présente. Les fractions allant de 132 mL à 186 mL ont été rassemblées. La fraction issue du lavage de la colonne (W) présentait aussi une bande correspondant à la protéine d'intérêt, qui ne s'était pas fixée correctement sur la résine. La quantité de protéines y étant minoritaire et l'objectif étant des tests d'activité nécessitant de faibles quantités de protéine, cela n'a pas posé de problème. L'ensemble rassemblé avait une A_{280nm} de 0,714 avec un rapport A_{260nm}/A_{280nm} de 0,68, n'indiquant pas de contamination aux acides nucléiques. Il a été dilué et chargé sur une résine Héparine.

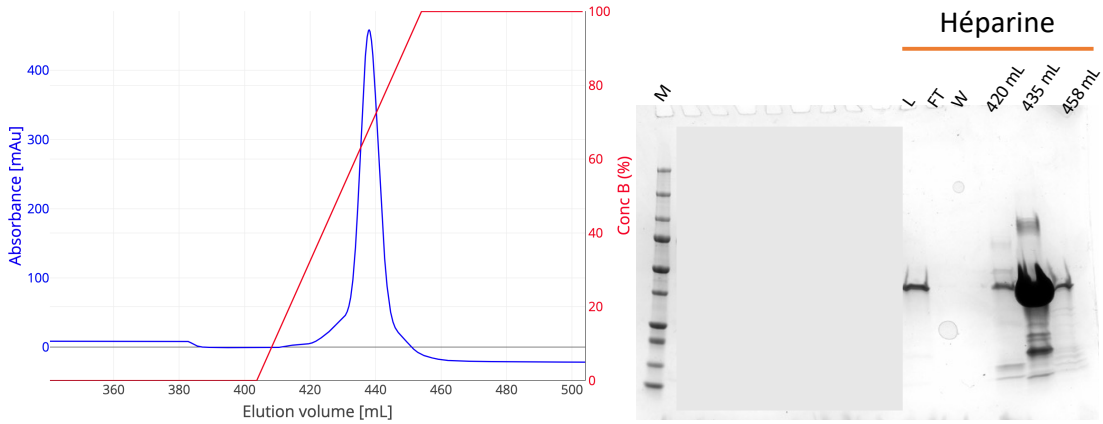
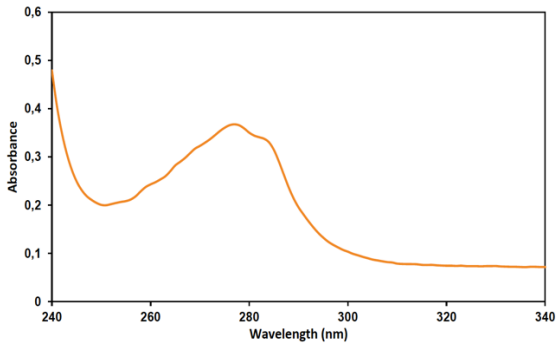


Figure 128 : Deuxième étape (Héparine) de purification de l'ADN polymérase β. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. À droite : SDS-PAGE. M : Marqueur de poids moléculaire; L : load, fraction soluble chargée sur la résine; FT :

Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Après avoir été analysées par SDS-PAGE, les fractions de 420 mL à 458 mL ont été regroupées et concentrées, jusqu'à une concentration de 183,23 µM (A_{280nm} = 4,557). L'échantillon a été dialysé après ajout de protéase TEV concentrée à 10 mg/mL. Le lendemain, le dialysat a été injecté sur une colonne HisTrap, et les fractions contenant les protéines non fixées à la résine ont été mélangées et concentrées jusqu'à 3 mL à 2,4 mg/mL. Afin de contrôler que l'étiquette avait bien été retirée, un contrôle qualité a été réalisé sur l'échantillon. La masse moléculaire a été mesurée par spectrométrie de masse MALDI-TOF, le spectre d'absorption a été obtenu, et la monodispersité de l'échantillon a été estimée par DLS.



Abs 260 nm	Abs 280 nm	Abs 340 nm
0,243	0,350	0,072

Figure 129 : Spectre d'absorption de l'échantillon de l'ADN polymérase β humaine entre 240 et 340 nm (les valeurs d'absorbance à 260 nm, 280 nm et 340 nm sont indiquées en dessous)

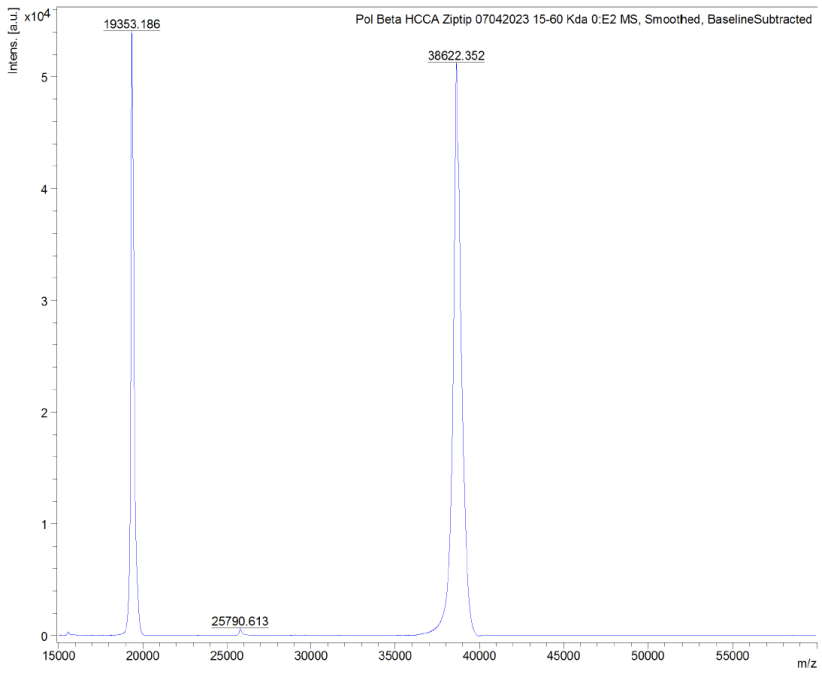
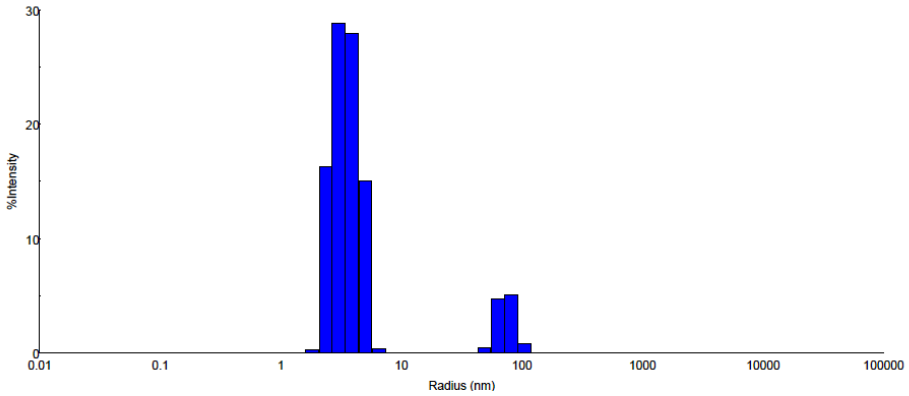


Figure 130 : Spectre de masse obtenu par MALDI-TOF de l'échantillon de l'ADN polymérase β humaine



	Rayon (nm)	*Poids Moléculaire (kDa)*	*% d'intensité*	*% de masse*
Pic 1	3,5	64,3	88,8	100
Pic 2	73,5	78409,4	11,2	0

Figure 131 : Mesure de la dispersité de l'échantillon de l'ADN polymérase β humaine. Le tableau en dessous indique les valeurs obtenues de rayon hydrodynamique, de poids moléculaire (approximatif), d'intensité, et de masse pour chaque espèce.

Le spectre d'absorption, réalisé avec un échantillon dilué au 1/10^{e}, montre bien l'absence de contamination aux acides nucléiques (A_{260nm}/A_{280nm} = 0,69) ainsi que l'absence d'agrégats (A_{340} ≈ 0,07). La mesure de spectrométrie de masse montre deux types d'ions : monochargés, donnant un m/z de 38622,352 (soit un poids moléculaire de 38,622 kDa); et doublement chargés, donnant un m/z de 19353,186 (soit un poids moléculaire de 38,706 kDa). La masse attendue étant de 38,652 Da, la protéine obtenue a bien la masse attendue, avec l'incertitude des mesures en particulier pour les ions doublement chargés, le MALDI-TOF n'étant pas adapté pour ces ions. Enfin, la mesure en DLS de l'échantillon montre une population très majoritaire (100%), ce qui indique que la protéine est bien homogène en solution dans le tampon choisi, et a un rayon hydrodynamique d'environ 3,5 nm. L'échantillon a été aliquoté et conservé à -20°C.

2.1.1.6 L'ADN polymérase λ humaine

Comme pour l'ADN polymérase β, le plasmide d'expression a été synthétisé par GenScript. Une fois le plasmide obtenu et sa séquence confirmée, la protéine a été produite en bactéries *E. coli* BL21star(DE3).

La purification a ensuite été réalisée en suivant le protocole global énoncé en Matériel et Méthodes, à partir d'une production de 2L de culture bactérienne. La lyse bactérienne a été réalisée au sonicateur, et après centrifugation, la fraction soluble a été chargée sur une colonne His-Trap, et les protéines fixées ont été éluées avec un gradient d'imidazole.

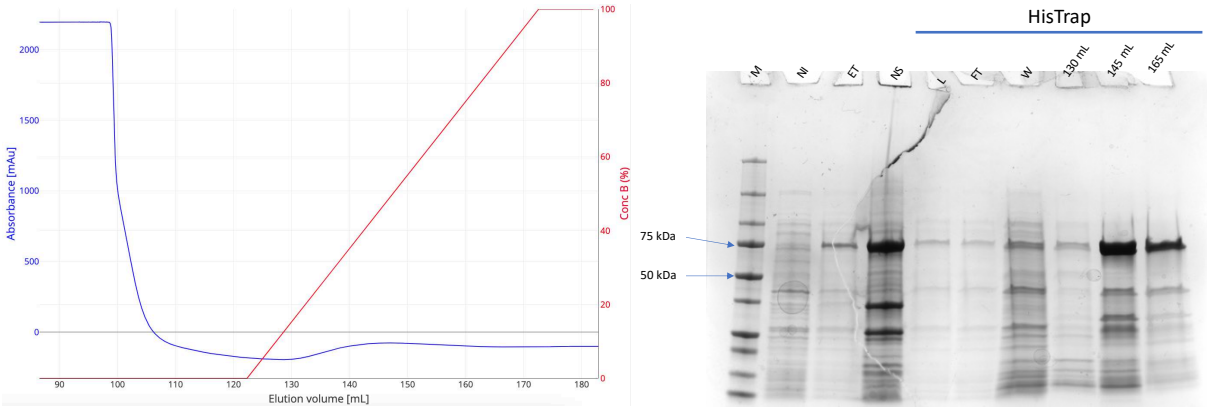


Figure 132 : Première étape (HisTrap) de purification de l'ADN polymérase λ. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. À droite : SDS-PAGE M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées) ; NI : bactéries avant induction de la production de protéines à l'IPTG ;ET : extrait total, ensemble des protéines solubles et insolubles obtenues après la lyse bactérienne ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Certaines fractions ont été analysées par SDS-PAGE. Une bande correspondant au poids moléculaire attendu (65 kDa environ) était bien présente, donc les fractions allant de 130 mL à 171 mL ont été rassemblées. La protéine était également présente dans la fraction non soluble, indiquant qu'une partie de la protéine produite avait été perdue lors de la lyse bactérienne, et que cette étape pourrait être optimisée. L'ensemble des fractions rassemblées avait une A_{280nm} de 0,57 avec un rapport A_{260nm}/A_{280nm} de 0,75, n'indiquant pas de contamination aux acides nucléiques. Ces fractions ont été diluées et injectées sur une colonne Héparine.

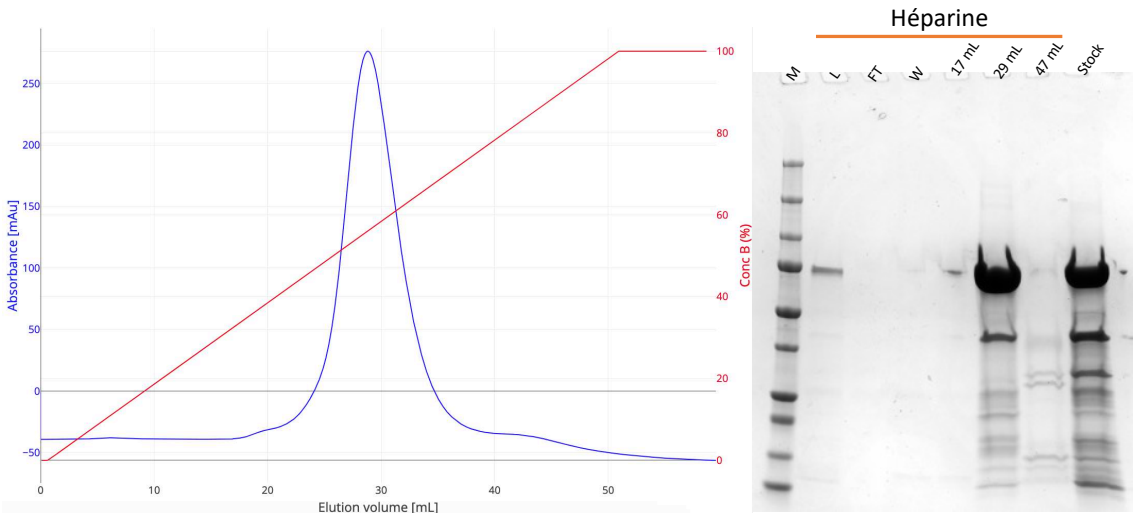


Figure 133 : Deuxième étape (Héparine) de purification de l'ADN polymérase λ. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. À droite : SDS-PAGE. M : Marqueur de poids moléculaire; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée ; Stock : fractions de 17 mL à 47 mL mélangées.

L'analyse SDS-PAGE ayant montré que la protéine d'intérêt y était présente en large majorité, les fractions allant de 17 mL à 47 mL ont été rassemblées et concentrées jusqu'à une absorbance à 280 nm de 1,447 (24,4 µM, ce qui est suffisant en vue de tests d'activité). Des aliquots de 50 µL ont été congelés et stockés à -20°C.

2.1.1.7 Récapitulatif

Les expériences décrites ici ont permis d'obtenir sous forme soluble toutes les protéines étudiées, à différents niveaux de pureté. Concernant les purifications des ADN polymérases X de *Paramecium*, les rendements ont été variables. En particulier pour PolXdFL, les purifications réalisées avant le re clonage (indiquées en gris dans le tableau 9) ont donné des rendements très faibles. L'ADN polymérase µ humaine a été purifiée au laboratoire par le Dr Sophia Missoury.

Tableau 9 : Résumé des purifications réalisées pour les PolX de Paramecium tetraurelia (celles indiquées en gris ont été réalisées avant de re générer des plasmides d'expression). La colonne Volume de culture indique le volume de culture bactérienne utilisé dans la purification associée. La colonne Tag indique la présence du tag 14 histidines (Oui) ou son absence (Non) sur les protéines produites. Le niveau de purification indique à quelle étape la purification a été arrêtée. La colonne QC indique si un QC a été réalisé et s'il a garanti la qualité de la protéine (ok).

Construction	Volume de culture utilisé (L)	Aliquots			Tag	Niveau de purification	QC	Utilisation	Rendement (mg/L de culture)
		Nombre	Volume (µL)	Concentration (mg/mL)					
PolXaΔBRCT	8	19	50	10,21	Non	GF (chromatographie d'exclusion stérique)		Essais de cristallisation	1,21
	4	18	50	2,17	Oui	GF	ok	Tests d'activité	0,488
	8	42	50	8,27	Non	GF	ok	Essais de cristallisation	2,17
PolXbΔBRCT	4	60	50	2,617	Oui	Héparine		Tests d'activité	1,96
	8	40	50	0,882	Oui	Héparine		Tests d'activité	0,220
PolXdΔBRCT	6	16	100	24,92	Non	GF		Essais de cristallisation	7,06
		25	20	4,984	Non	GF		Tests d'activité	
	10	37	50	10,22	Non	GF		Essais de cristallisation	0,91
		1	20	20,3	Non	GF		Essais de cristallisation	
PolXdFL	4	40	50	7,5	Non	GF	ok	Tests d'activité	3,75
	5	50	100	2,17	Non	Héparine + TEV		Tests d'activité	2,17
	5	2	50	17,19	Oui	GF		Tests d'activité	1,15
		4	50	20,06	Oui	Héparine		Tests d'activité	
	5	16	50	2,05	Oui	GF		Tests d'activité	0,32
	8	1	130	0,82	Oui	GF		Tests d'activité	0,013
ADN polymérase β	2	37	50	7,73	Oui	Héparine		Tests d'activité	10,77
		60	50	2,421	Non	Héparine+TEV	ok	Tests d'activité	
ADN polymérase λ	2	111	50	1,58	Oui	Héparine		Tests d'activité	4,38

2.1.2 Caractérisation enzymatique des ADN polymérases X de *Paramecium tetraurelia* dans différents contextes

Une fois produites et purifiées, ces ADN polymérases ont été testées dans différents contextes, de façon à caractériser leur activité enzymatique, et les situer parmi les ADN polymérases X déjà caractérisées.

2.1.2.1 Extension d'amorce

Dans les conditions d'extension d'amorce, deux tests ont été réalisés, à haute (1 µM) et basse (1 nM) concentration pour les ADN polymérases X de *P. tetraurelia*.

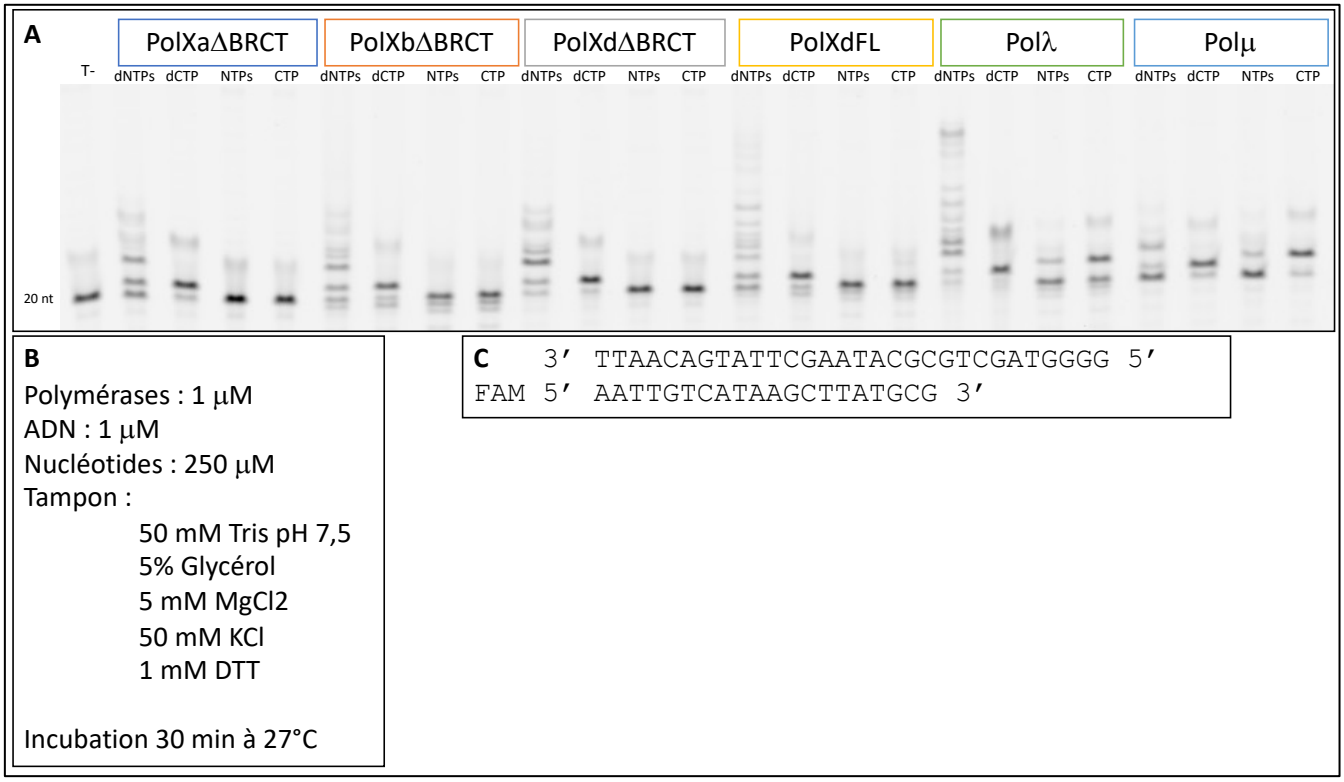


Figure 134 : : Test d'activité - Extension d'amorce à haute concentration. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

À haute concentration, les ADN polymérases X (y compris humaines) semblent être actives dans ce contexte, mais ne parviennent pas à étendre entièrement l'amorce : l'ADN polymérase λ ajoute 10 nt, mais cette activité reste partielle ; l'ADN polymérase µ ajoute 2 à 3 dNTPs ; et les ADN polymérases X de *Paramecium* ajoutent au maximum 6 dNTPs. En présence du premier dNTP seul, toutes les enzymes testées semblent fidèles puisqu'elles n'ajoutent qu'un nucléotide. Les deux ADN polymérases X humaines se montrent cependant capables d'utiliser des ribonucléotides avec une activité faible, mais pas les ADN polymérases X de *Paramecium*.

Figure 135 : Test d'activité - Extension d'amorce à basse concentration. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

À basse concentration, les ADN polymérases X de *P. tetraurelia* ne semblent pas actives : elles ajoutent très faiblement une molécule de dCTP, sauf PolXbΔBRCT qui présente une activité plus élevée en présence de ce nucléotide. La reproductibilité de ce résultat semble indiquer une activité particulière, qui pourrait reposer sur une affinité particulièrement élevée de cette enzyme pour ce nucléotide.

2.1.2.2 NHEJ

Le substrat utilisé ici est particulièrement adapté à l'ADN polymérase µ. En effet, parmi les ADN polymérases testées ici, elle seule peut accommoder des micro-homologies inférieures à 2 pb, alors que l'ADN polymérase λ peut réparer des substrats présentant des micro-homologies plus longues.

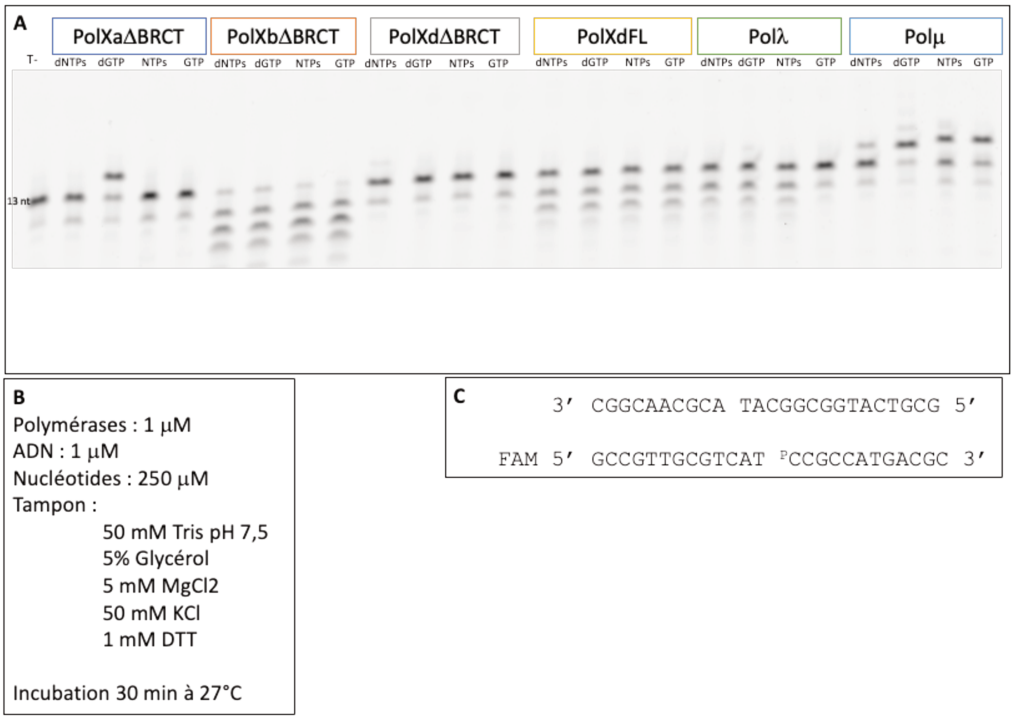


Figure 136 : Test d'activité – NHEJ. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 13 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Comme attendu, ici seule l'ADN polymérase µ est active : elle ajoute un nucléotide à ce substrat avec deux nucléotides de micro-homologie. Elle est également active en présence de ribonucléotides (Martin *et al.*, 2013). Dans ce contexte, l'ADN polymérase λ n'est donc pas active, comme attendu, à cause de la microhomologie qui est trop courte (Nick McElhinny *et al.*, 2005). De la même manière, les ADN polymérases X de *P. tetraurelia* ne sont pas actives, et dégradent même pour certaines (PolXbΔBRCT et PolXdFL) l'amorce avec une activité de pyrophosphorolyse (l'activité inverse de la polymérisation, caractérisée en détails pour l'ADN polymérase β). Une activité est cependant observée pour PolXaΔBRCT en présence de dGTP, de façon reproductible. Là encore cela pourrait être dû à une affinité particulière pour ce nucléotide. En effet, la caractérisation cinétique de ces enzymes (présentée en partie 1.4) indique que les ADN polymérases X de *P. tetraurelia* présentent des efficacités catalytiques similaires, mais des affinités différentes pour les nucléotides.

Ces résultats semblent indiquer que les ADN polymérases X de *Paramecium* sont plus proches (en termes d'activité) de l'ADN polymérase λ que de l'ADN polymérase µ, puisqu'elles ne sont pas actives avec cette micro-homologie de 2 nt. C'est une conclusion logique, puisque ces ADN polymérases n'ont pas de boucle 1 d'après les analyses de séquences présentées dans le chapitre 1, contrairement à l'ADN polymérase µ (or cette boucle permet à

l'ADN polymérase μ d'être actives dans ce contexte ce substrat (Freudenthal *et al.*, 2013; Loc'h *et al.*, 2019; Shock *et al.*, 2017).

2.1.2.3 MMEJ

Les ADN polymérases X de *Paramecium tetraurelia* ont été testées dans un contexte proche appelé ici MMEJ, pour *Microhomology Mediated End Joining*. Il s'agit d'un substrat proche du substrat NHEJ, mais avec une micro-homologie bien plus longue, de 10 nt.

Figure 137 : Test d'activité – MMEJ. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 35 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Les quatre constructions d'ADN polymérases X testées ici semblent actives dans ce contexte. À concentration équivalente, cette activité est plus forte lorsque le nucléotide correct à incorporer (dGTP) est seul qu'en présence des trois autres dNTPs. Cela indique que *in vitro* l'activité de l'enzyme est ralentie par les nucléotides incorrects. De plus, aucune des ADN polymérases X de *P. tetraurelia* n'incorpore de ribonucléotides malgré leur haute concentration (proche des concentrations cellulaires) et un temps de 30 min.

2.1.2.4 Terminal transférase

L'activité de terminal transférase est une particularité de la Tdt et de l'ADN polymérase μ (pour cette dernière, cette activité est dépendante de la présence d'ions Mn^{2+}). Ce sont les seules ADN polymérases de la famille X à avoir cette activité, de façon dépendante de leur

boucle 1 (Juárez *et al.*, 2006). Les ADN polymérases X de *Paramecium,* n'ayant pas cette boucle (comme l'ADN polymérase λ), ne devraient logiquement pas présenter cette activité.

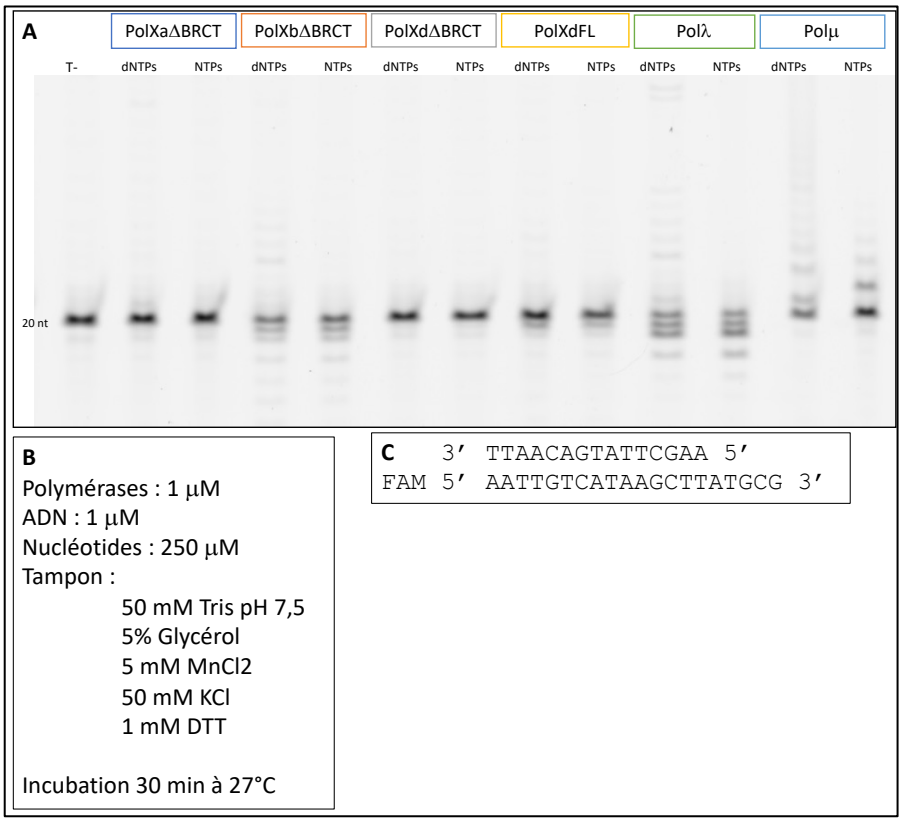


Figure 138 : Test d'activité – Terminal Transférase. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Tout d'abord, on observe que l'ADN polymérase μ semble active en présence de manganèse, que ce soit avec les dNTPs ou les NTPs, comme attendu (Domínguez *et al.*, 2000). L'ADN polymérase λ semble avoir une très faible activité polymérase, mais montre une activité de pyrophosphorolyse. Les ADN polymérases X de *Paramecium* n'ont pas non plus d'activité dans ce contexte, et pour certaines elles dégradent aussi l'amorce (PolXbΔBRCT et PolXdFL), et ce malgré la présence de manganèse.

Ce résultat semble indiquer que les ADN polymérases X de *Paramecium* sont plus proches de l'ADN polymérase λ que de l'ADN polymérase μ.

2.1.2.5 Gap-filling

J'ai également testé les ADN polymérases X de *P. tetraurelia* en contexte de gap-filling, dans lequel un nucléotide doit être incorporé dans un brin d'ADN discontinu, hybridé à un brin continu. Ici, leur rôle est d'ajouter un nucléotide, sans déplacer l'amorce en aval du dommage. C'est le substrat retrouvé dans la réparation BER, qui implique l'ADN polymérase β, mais c'est aussi un contexte utilisé pour caractériser l'ADN polymérase λ.

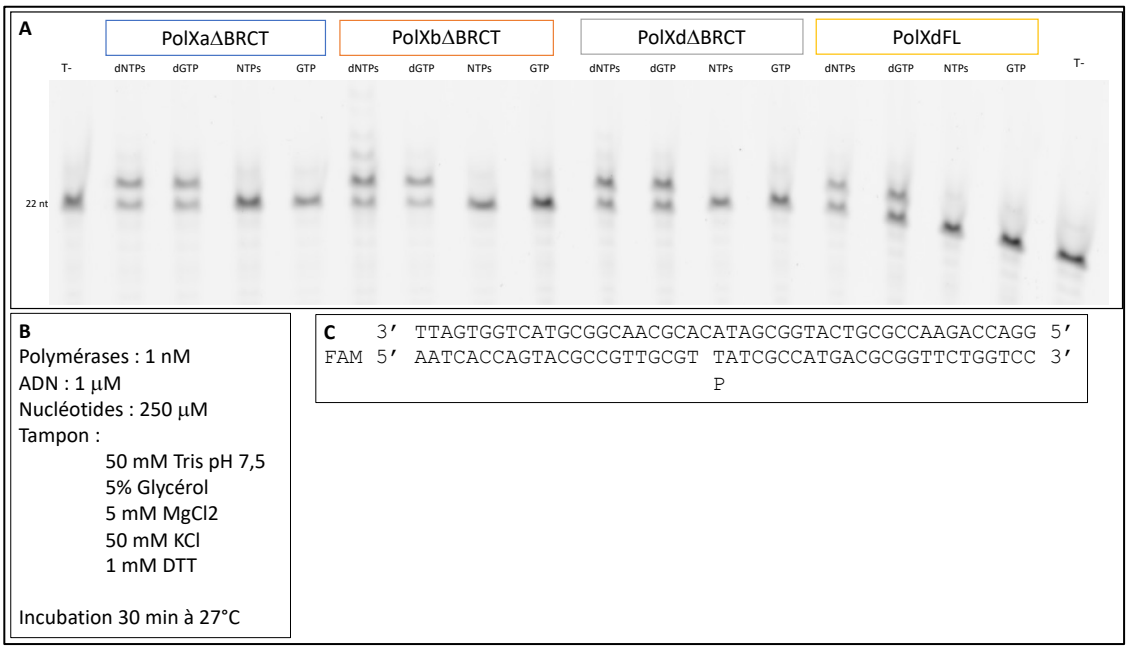


Figure 139 : Test d'activité – Gap-filling. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Ici, l'ADN polymérase λ a été testée dans des conditions légèrement différentes, décrites dans la littérature (Garcia-Diaz *et al.*, 2005a).

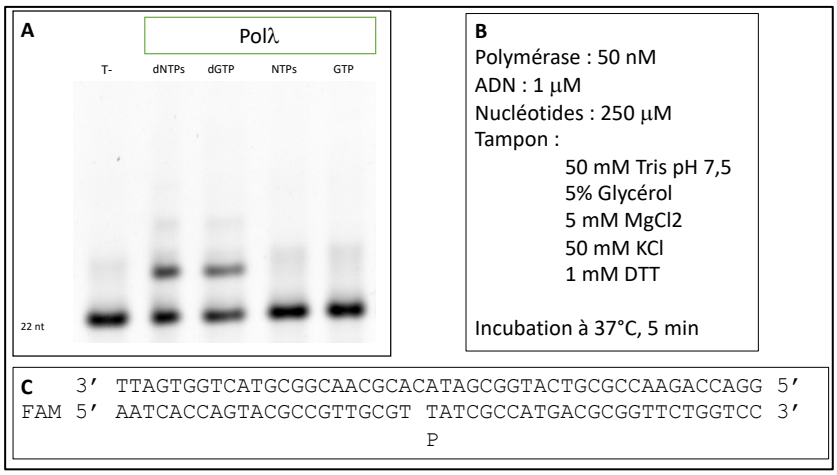


Figure 140 : Test d'activité – Gap-filling (ADN polymérase λ). A : Résultat du test. Les nucléotides testés sont indiqués (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Comme attendu, l'ADN polymérase λ a bien une activité de gap-filling, et n'incorpore pas de NTPs. Le résultat semble le même pour les ADN polymérases X de *Paramecium* : elles incorporent un désoxynucléotide, sans déplacer l'amorce en aval et sans incorporer de NTPs.

2.1.2.6 NHEJ-cis

Le substrat utilisé ici est le plus proche de celui rencontré à la suite de l'élimination des IES chez *P. tetraurelia*. Il s'agit d'une cassure double brins, mais ici la base instructive est sur le duplexe d'ADN que la polymérase doit étendre (en *cis*). Il s'agit d'un substrat proche d'une extension d'amorce, mais l'ADN polymérase doit maintenir la micro-homologie avec le duplexe d'ADN en *trans*, sans déplacer le brin amorce en aval, comme en gap-filling.

Le test a d'abord été réalisé avec de basses concentrations d'ADN polymérases.

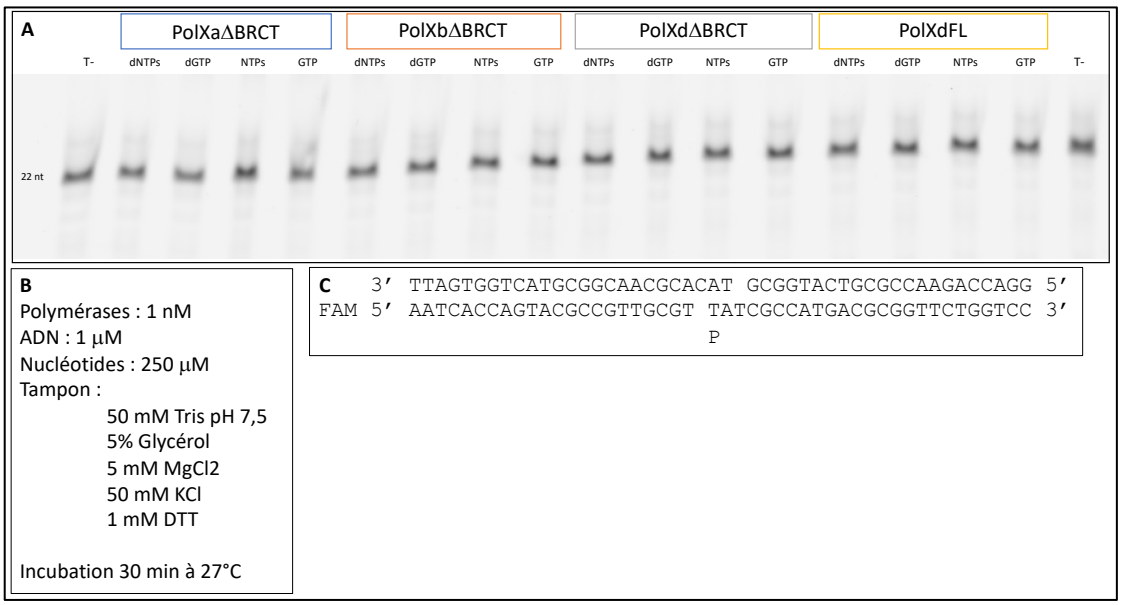


Figure 141 : Test d'activité – NHEJ-cis à basse concentration. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Dans ces conditions, les ADN polymérases testées ne sont pas actives. Cependant, ces enzymes étant surexprimées lors des réarrangements du génome de *P. tetraurelia*, nous avons choisi de tester des concentrations croissantes d'enzyme (uniquement avec l'une d'entre elles : PolXaΔBRCT).

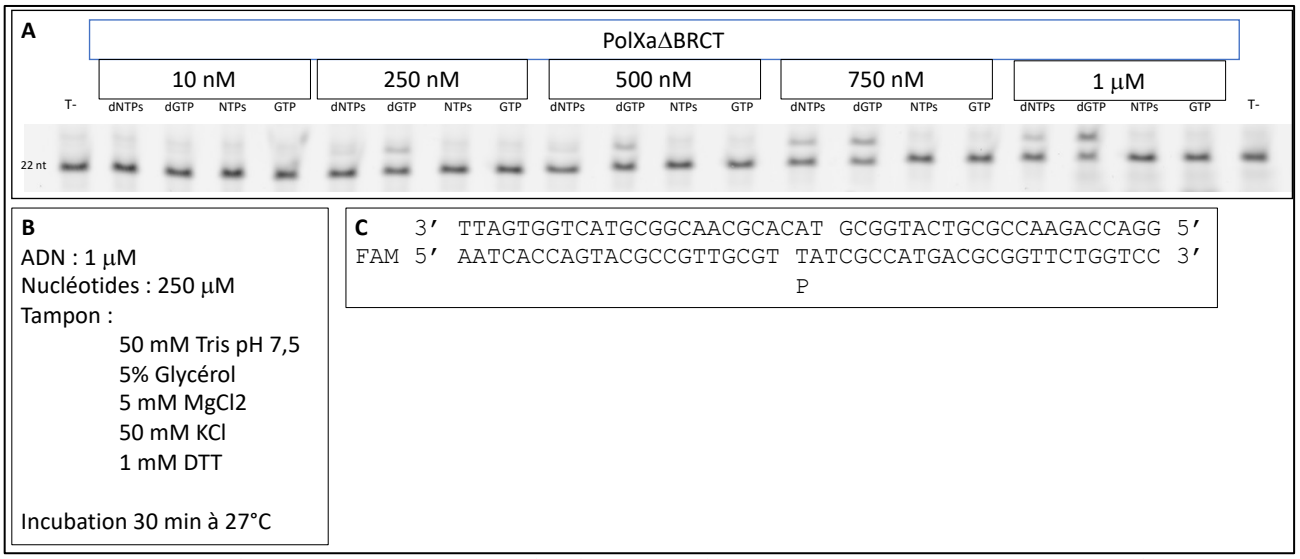


Figure 142 : Test d'activité – NHEJ-cis. A : Résultat du test. La polymérase testée et ses concentrations sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

En augmentant la concentration de l'enzyme, celle-ci est capable d'ajouter un nucléotide (dès 250 nM). Par ailleurs, on remarque que même à haute concentration elle n'incorpore pas de NTPs, et n'ajoute pas plus d'un nucléotide, signe qu'elle maintient la micro-homologie TA, puisqu'elle pourrait étendre une amorce sur plus de 1 nt dans ce contexte (Chapitre 2, 2.1.2.1, page 182).

Ce résultat semble cohérent : les ADN polymérases X de *Paramecium* ne sont actives avec leur substrat « naturel » que lorsqu'elles sont à relativement haute concentration et conservent leur fidélité. Cette activité spécifique à haute concentration peut être mise en lien avec la surexpression de l'ADN polymérase X a chez *P. tetraurelia* lors des réarrangements programmés du génome, ou avec le recrutement de ces ADN polymérases par leurs partenaires protéiques comme Ku70/80, qui peut augmenter leur concentration locale (d'autant plus dans un contexte épigénétique où les régions endommagées sont regroupées au sein du noyau (Arnould *et al.*, 2021; Marnef and Legube, 2017)).

2.1.2.7 Conclusion

L'ensemble de ces tests ont permis de définir dans quels contextes les ADN polymérases X de *Paramecium* sont actives.

Tableau 10 : Récapitulatif des contextes dans lesquels les ADN polymérases X de Paramecium tetraurelia sont actives (+ : activité présente ; - : pas d'activité)

Contexte	Activité	Fidélité	Note
Extension d'amorce	+	+	Activité dépendante de la concentration
NHEJ (2pb MH)	-		
MMEJ (10pb MH)	+	+	
Terminal transférase	-		
Gap-filling	+	+	
NHEJ-cis (2pb MH)	+	+	Activité dépendante de la concentration

Les ADN polymérases X de *P. tetraurelia* semblent être actives dans les mêmes contextes que ceux connus pour l'ADN polymérase λ, alors que dans les contextes spécifiques de l'ADN polymérase μ, elles ne sont pas actives. De plus, elles ont montré dans tous ces tests une bonne discrimination contre les ribonucléotides, contrairement à la l'ADN polymérase μ ; et dans le cadre du test d'extension d'amorce à haute concentration, l'ADN polymérase λ incorpore des NTPs, mais pas les ADN polymérases X de *P. tetraurelia*. Celles-ci semblent donc avoir une meilleure capacité à discriminer le type de nucléotides qu'elles incorporent. Elles ont également montré une activité correcte, dépendante de leur concentration dans le test avec le substrat NHEJ-cis. Cela peut être mis en lien avec leur surexpression lors des réarrangements du génome chez *P. tetraurelia*, ou avec leur recrutement par leurs partenaires protéiques.

2.1.3 Test de l'activité dRP lyase d'une ADN polymérase X de *Paramecium tetraurelia*

L'activité dRP lyase a pour rôle de retirer un groupement désoxyribose-phosphate d'une extrémité 5' d'ADN. Cette activité est essentielle pour les ADN polymérases λ et β, qui ont besoin de reconnaitre un groupement 5'P pour synthétiser les bases manquantes.

Cette activité a une importance particulière au sein du système de réparation BER (Allinson *et al.*, 2001). En effet, ce système de réparation produit après ses premières étapes une cassure simple brin suivie en aval d'un groupement 5'dRP, créé par l'activité d'une ADN glycosylase. Or, dans le cadre de la réparation BER *Short-range*, l'étape suivante est la synthèse par l'ADN polymérase β du nucléotide manquant (Figure 47). Le groupement dRP doit donc être remplacé par un groupement phosphate par l'activité dRP lyase de l'ADN polymérase β.

Cette activité est portée par quelques résidus situés dans le domaine de 8 kDa des ADN polymérases X, les mêmes que ceux impliqués dans la reconnaissance des groupements 5'P. Des équivalents de ces résidus sont trouvés chez l'ADN polymérase λ, mais aussi chez les ADN polymérases X de *P. tetraurelia*, comme indiqué dans le chapitre 1 (page 103). L'existence de résidus équivalents aux ADN polymérases X humaines chez *Paramecium* m'a poussé à chercher l'existence de l'activité dRP lyase chez ces ADN polymérases. Pour cela, j'ai réalisé une expérience visant à reconstituer *in vitro* un équivalent du système de réparation BER (Chapitre 2, 1.1.7.1, page 144). Un duplexe d'ADN portant un dU a été mis en contact avec une uracile glycosylase et une endonucléase de façon à éliminer le dU, en laissant un espace de 1 nt suivi d'un groupement dRP en 5' de l'ADN en aval. Les ADN polymérases X a de *Paramecium tetraurelia* et β humaine ont été incubées avec cet ADN, des nucléotides et une ADN ligase. Les ADN polymérases ne peuvent ajouter un nucléotide (comme en gap-filling) que si le groupement dRP est supprimé par leur activité dRP lyase. Par conséquent, si elles portent bien cette activité, elles peuvent ajouter un nucléotide, puis l'ADN ligase peut rétablir la liaison phosphodiester de l'ADN. La taille de l'ADN est suivie pendant l'expérience par fluorescence : sa mise en contact avec l'ADN glycosylase et l'endonucléase réduit sa taille visualisée sur gel, mais la réparation par les ADN polymérases et ligase doit lui rendre sa taille d'origine.

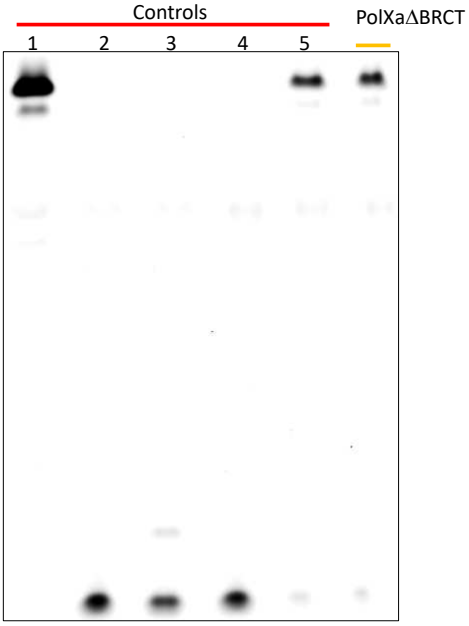


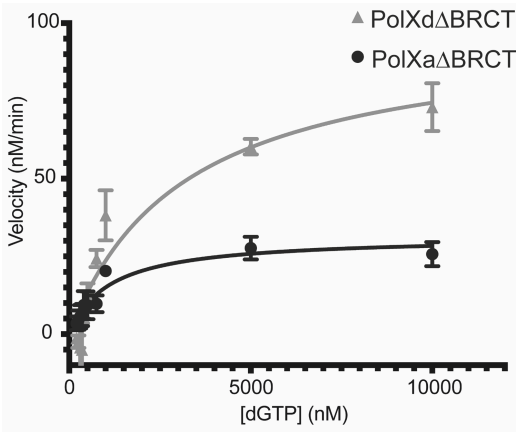
Figure 143 : Résultat du test d'activité dRP lyase de PolXaΔBRCT. Contrôles : 1 : ADN initial non traité par les enzymes du kit USER3 ; 2 : ADN traité par les enzymes USER3 ; 3 : réaction sans ADN ligase ; 4 : réaction sans ADN polymérase ; 5 : contrôle positif réalisé avec l'ADN polymérase β.

Les contrôles 1 et 2 permettent de vérifier que l'ADN initial est de plus grande taille qu'après son traitement par le kit USER3 : c'est bien le cas ici, après utilisation du kit USER3, la taille de l'ADN baisse. Dans le troisième contrôle négatif, en absence d'ADN ligase l'ADN

reste de petite taille malgré un possible ajout de nucléotide par l'ADN polymérase. Le quatrième contrôle montre que sans ADN polymérase, la ligation et le retour de l'ADN a sa taille initiale ne sont pas possibles. Le cinquième contrôle est quant à lui un contrôle positif réalisé avec l'ADN polymérase β humaine : celle-ci a une activité dRP lyase, comme attendu (Allinson *et al.*, 2001). L'ensemble de ces contrôles indiquent que le protocole de test est correct. Le test réalisé avec la construction PolXaΔBRCT montre un résultat similaire à celui obtenu avec l'ADN polymérase β humaine. Cela indique que l'ADN polymérase X a de *Paramecium tetraurelia* porte bien une activité dRP lyase. Enfin, puisque les quatre ADN polymérases X de *P. tetraurelia* présentent les mêmes résidus impliqués dans cette activité d'après les analyses de séquences, il est très probable que ces quatre enzymes aient cette même activité.

2.1.4 Caractérisation cinétique des ADN polymérases X de *Paramecium tetraurelia*

Deux constructions d'ADN polymérases X de *Paramecium tetraurelia* (PolXaΔBRCT et PolXdΔBRCT) ont été caractérisées du point de vue cinétique en contexte de gap-filling, de façon à les comparer aux ADN polymérases λ et β humaines. Les deux enzymes ont été testées en présence de quantités croissantes de dGTP, le nucléotide à incorporer sur le substrat d'ADN présent dans l'expérience. Ces expériences ont permis d'obtenir un profil cinétique en accord avec le modèle de Michaelis-Menten. Les valeurs de k_{obs} et de K_m ont donc pu être obtenues, ainsi que les valeurs bornes des intervalles de confiance à 95%.



	PolXaΔBRCT	PolXdΔBRCT	hPolλΔBRCT (Garcia-Diaz *et al.*, 2004)	hPolβ (Chagovetz *et al.*, 1997)
k_{cat} (min^{-1})	6,30 (5,779 - 6,843)	19,57 (15,59 - 26,25)	2,1 ± 0,546	120 ± 24
K_m (nM)	1095 (886,5 – 1357)	3151 (1857 – 6198)	1500 ± 44	200 ± 100
Efficacité catalytique ($min^{-1}.\mu M^{-1}$)	5,75	6,21	1,4	600

Figure 144 : Analyse cinétique de l'incorporation de dGTP dans un substrat de gap-filling par PolXaΔBRCT et PolXdΔBRCT. En haut : représentation de Michaelis de la vélocité des enzymes en fonction de la concentration en dGTP. En bas : Valeurs de k_{obs}, K_m et obtenues par approximation du modèle de Michaelis Menten avec les données expérimentales (et valeurs limites de l'intervalle de confiance à 95%), valeurs de k_{cat} et K_m connues dans la littérature pour la PolλΔBRCT humaine et pour la Polβ humaine en gap-filling, et valeurs d'efficacité catalytique calculées.

Les valeurs de k_{obs} et K_m obtenues pour PolXdΔBRCT, bien qu'elles soient du même ordre de grandeur que pour PolXaΔBRCT, en diffèrent car PolXdΔBRCT semble plus rapide que PolXaΔBRCT. Cependant, PolXaΔBRCT a un K_m de 1095 nM, trois fois plus faible que PolXdΔBRCT (3151 nM), qui semble indiquer une meilleure affinité pour le substrat testé (ici le dGTP). Leurs efficacités catalytiques sont donc comparables malgré des paramètres cinétiques différents.

Ces valeurs k_{obs} et K_m semblent similaires à celles connues pour l'ADN polymérase λ (sans domaine BRCT et linker) humaine, mais diffèrent drastiquement de celles de l'ADN polymérase β humaine, qui semble beaucoup plus rapide avec un k_{cat} 10 à 20 fois plus élevé, et un K_m 5 à 15 fois plus faible, suggérant une forte affinité pour le nucléotide entrant. Au vu des similitudes entre les ADN polymérases X de *P. tetraurelia* et l'ADN polymérase λ humaine, cela semble logique : contrairement à l'ADN polymérase β, elles ne sont pas spécialisées dans le gap-filling, et sont donc moins efficaces dans ce contexte. Il est également possible qu'une différence de mécanisme soit à l'origine de ces différences. L'ADN polymérase β est connue pour être la seule ADN polymérase X de métazoaire à alterner des formes ouvertes et fermées

au cours de son cycle catalytique. Cette alternance facilite l'entrée des substrats (ADN et nucléotide entrant) dans le site actif, ce qui augmente son affinité pour ces molécules. Les informations cinétiques obtenues ici semblent indiquer que les ADN polymérases X de *Paramecium tetraurelia* sont plus proches de l'ADN polymérase λ du point de vue cinétique. Il est donc possible que comme l'ADN polymérase λ, leur domaine catalytique reste sous forme fermée tout au long de la catalyse.

2.1.5 Comparaison de la fidélité des ADN polymérases X de *Paramecium tetraurelia* et de l'ADN polymérase λ humaine

Comme indiqué en introduction de ces travaux, mon hypothèse pour expliquer le très faible taux d'erreurs du système NHEJ spécialisé de *Paramecium tetraurelia* repose sur les ADN polymérases X impliquées, qui pourraient être très hautement fidèles. J'ai donc cherché à observer cette fidélité en les comparant directement à l'ADN polymérase λ, avec laquelle elles partagent de nombreuses similitudes (contextes d'activité, activité dRP lyase, constantes cinétiques, présence d'un domaine BRCT, etc.). La caractérisation initiale de l'ADN polymérase λ humaine indiquait que sa fidélité dépend en partie de son extension N-terminale (Fiala *et al.*, 2006). J'ai donc comparé deux constructions de l'ADN polymérase λ (entière (FL) et sans extension N-terminale (construction λmut, étudiée dans la seconde partie de ces travaux)) à deux constructions équivalentes d'ADN polymérases X de *P. tetraurelia* (PolXaΔBRCT et PolXdFL).

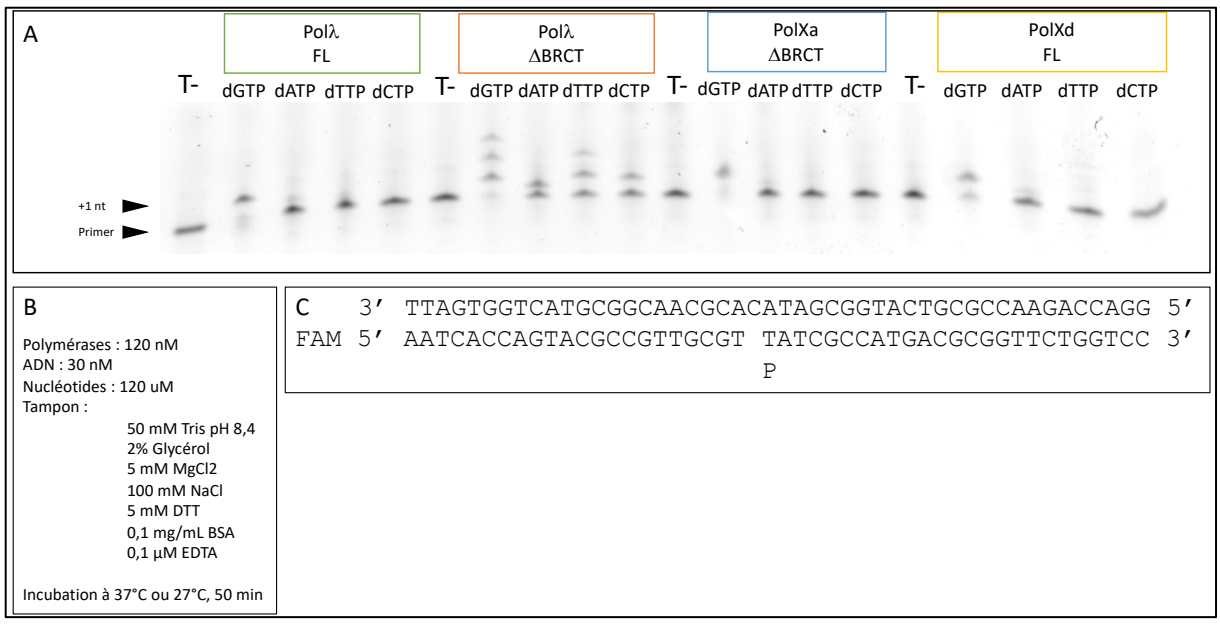


Figure 145 : Test d'incorporation de nucléotides incorrects – Gap-filling. A : Résultat du test. Les polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Toutes les constructions testées incorporent du dGTP, ce qui indique qu'elles sont bien actives. La seule construction testée ici qui réalise des incorporations incorrectes semble être la version tronquée de l'ADN polymérase λ humaine, qui incorpore des dATP, dTTP et dCTP après 50 min. Dans le cas du dTTP, on remarque même une double incorporation incorrecte. Cela peut être dû à une activité de déplacement de brin. Cette même construction incorpore également plusieurs molécules de dGTP, de façon incorrecte. La version sauvage de cette même polymérase λ humaine ne semble ici pas faire autant d'erreurs, même si elle incorpore faiblement un dATP. Les deux constructions d'ADN polymérases X de *P. tetraurelia* testées ici ont cependant un comportement similaire dans ce test : ces enzymes ne semblent pas incorporer de nucléotides incorrects, et ce indépendamment de la présence ou absence des éléments présents en N-terminal (linker et domaine BRCT).

La différence entre les deux constructions testées pour l'ADN polymérase λ a déjà été étudiée (Fiala *et al.*, 2006), dans des travaux qui ont montré une différence de fidélité entre les versions complètes et tronquées de cette ADN polymérase. Cette différence n'est pas justifiée par la présence ou l'absence du domaine BRCT, qui n'a aucune fonction catalytique, mais peut être expliquée par la présence ou l'absence d'une séquence riche en sérines et prolines présente dans le linker de l'ADN polymérase λ. Cependant, le rôle exact de cette séquence est encore mal compris. Les résultats observés pour la version complète de cette polymérase sont cohérents avec les premières publications sur cette polymérase (García-Díaz *et al.*, 2002), qui indiquaient que ses erreurs sont plus fréquemment des transitions (intégration d'une purine à la place de l'autre (A=G) ou d'une pyrimidine à la place de l'autre (C=T)) que des transversions (intégrations de purines à la place de pyrimidines et inversement(A/G=C/T)).

Les résultats obtenus pour PolXaΔBRCT et PolXdFL indiquent que les ADN polymérases X de *P. tetraurelia* sont au moins aussi fidèles que la forme sauvage de l'ADN polymérase λ humaine, qui est presque aussi fidèle que l'ADN polymérase β humaine avec des taux d'erreurs du même ordre, à savoir 1 erreur toutes les 10^4 à 10^5 bases (Bebenek *et al.*, 2003; Fiala *et al.*, 2006; García-Díaz *et al.*, 2002; Ramadan *et al.*, 2002; Yamtich and Sweasy, 2010). Une autre information d'importance est que cette fidélité des ADN polymérases X de *Paramecium* n'est pas influencée par un élément situé en N-terminal comme pour l'ADN polymérase λ, mais qu'elle a pour origine le domaine catalytique lui-même, puisque la suppression de toute la partie située en N-terminal ne modifie pas la fidélité observée dans ces expériences.

2.1.6 Essais de cristallogenèse des ADN polymérases X de *Paramecium tetraurelia*

Après avoir caractérisé l'activité des ADN polymérases X de *Paramecium tetraurelia*, l'objectif était d'obtenir leur structure cristallographique à haute résolution pour comprendre les bases structurales de la fidélité de ces ADN polymérases. Plusieurs campagnes de criblage de conditions de cristallisation des construction PolXaΔBRCT et PolXdΔBRCT ont été réalisées, dans des conditions inspirées par les articles ayant permis d'obtenir les structures des ADN polymérases λ et β. Plus de 23000 conditions de cristallisation (dans différentes conditions physico-chimiques, avec différents oligonucléotides et nucléotides entrants) ont été testées avec l'aide de la plateforme de cristallographie de l'Institut Pasteur, mais aucune n'a permis d'obtenir de cristaux avec ces protéines.

2.1.7 Conclusion

Les expériences décrites dans cette première partie ont permis de produire et purifier quatre constructions protéiques basées sur les ADN polymérases X sauvages de *Paramecium tetraurelia*.

Ces constructions ont pu être caractérisées d'un point de vue enzymatique. Les ADN polymérases X de *P. tetraurelia* semblent partager des points communs avec l'ADN polymérase λ humaine. Elles sont en effet inactives dans les mêmes contextes (terminal transférase, NHEJ avec une courte microhomologie) et actives dans les mêmes contextes (gap-filling, « MMEJ »). Leurs caractéristiques cinétiques sont du même ordre que l'ADN polymérase λ. Les ADN polymérases X de *Paramecium* ont également une activité dRP lyase, comme les ADN polymérases λ et β. Il semble également que leur activité dans le contexte le plus proche de leur activité physiologique (NHEJ-cis) soit dépendante de leur concentration, ce qui pourrait être mis en lien avec leur surexpression lors des réarrangements du génome de *P. tetraurelia*, ou avec leur recrutement au sein du système de réparation.

Dans tous les contextes testés, les enzymes de *P. tetraurelia* ont montré une meilleure capacité que l'ADN polymérase λ humaine à bloquer l'incorporation de NTPs. Elles semblent être au moins aussi fidèles que l'ADN polymérase λ, et il semble que leur fidélité soit liée à leur domaine catalytique, tandis que la fidélité de l'ADN polymérase λ repose au moins en partie sur une séquence en N-terminal. L'objectif était de comprendre l'origine de la fidélité du système de réparation des cassures doubles brin programmées chez *P. tetraurelia*. Il semble que

cette fidélité repose au moins en partie sur celle des ADN polymérases impliquées. Pour comprendre le fonctionnement de la fidélité de ces enzymes, nous avons essayé d'en obtenir des cristaux pour les étudier d'un point de vue structural. Cela n'a pas fonctionné, et nous avons donc opté pour une approche indirecte.

2.2 Étude indirecte d'un mécanisme de fidélité des ADN polymérases X de *Paramecium tetraurelia* reposant sur l'activation du site catalytique

Les criblages de conditions de cristallogenèse réalisés sur les constructions d'ADN polymérases X de *Paramecium tetraurelia* n'ont pas permis d'obtenir de cristaux et de structure cristallographique permettant d'expliquer la fidélité observée chez ces enzymes. Une autre approche a donc été employée pour étudier un des possibles mécanismes de fidélité porté par ces enzymes : le mécanisme d'activation / inactivation (ou fermeture / ouverture) du site catalytique. S'il existe, ce mécanisme doit être proche de celui rencontré chez l'ADN polymérase β humaine lui permettant d'activer ou non son site catalytique, mécanisme qui semble dépendre essentiellement d'un pont salin entre la position 5 du second motif catalytique et le résidu central du motif SD2. Pour rappel, la fidélité de l'ADN polymérase β humaine repose sur un mécanisme d'ouverture-fermeture du site catalytique (voir page 122) : sous forme ouverte, le site catalytique est inactif, et les substrats (ADN et dNTP correct) peuvent se placer, et lorsque leur géométrie permet la catalyse, le domaine catalytique se referme, et la catalyse peut avoir lieu. Ces mécanismes reposent sur quelques résidus : les aspartates catalytiques, séparés en deux motifs (**D**M**D** et RI**D**IR), un résidu arginine situé à la fin du motif RIDI**R**, la phénylalanine du *steric gate* Y**F**TGS, et le motif SD2, qui chez l'ADN polymérase β est un motif N**E**Y. Comme nous l'avons vu dans le chapitre 1, les ADN polymérases X de *P. tetraurelia* portent quant à elles de façon unique parmi les ADN polymérases X des résidus équivalents : **D**M**D**, RI**D**L**K**, Y**F**TGS et S**D**H (motif SD2). Nous pensons donc qu'elles pourraient disposer d'un mécanisme similaire.

Nous avons cherché à conférer ce possible mécanisme (et son homologue de l'ADN polymérase β) à l'ADN polymérase λ humaine par mutagenèse dirigée. Les expériences réalisées sur les constructions similaires à l'ADN polymérase β devaient servir de témoin : s'il est possible de donner à l'ADN polymérase λ le mécanisme de l'ADN polymérase β (incluant les changements au niveau des résidus du site catalytique et les changements conformationnels globaux), il est normalement possible d'observer un mécanisme similaire présent chez les ADN

polymérases X de *P. tetraurelia* en réalisant des expériences équivalentes. La structure de l'ADN polymérase λ a été étudiée en 2022 (Jamsen *et al.*, 2022) par cristallographie, en utilisant un protocole permettant d'obtenir de façon reproductible des cristaux de bonne qualité. Nous avons appliqué ce protocole à plusieurs constructions mutantes de cette enzyme, et obtenu des cristaux qui ont permis d'étudier l'impact des mutations réalisées sur l'activité de l'ADN polymérase λ humaine.

2.2.1 Rappel des constructions étudiées

Cinq constructions mutées basées sur l'ADN polymérase λ humaine ont été produites. Toutes présentaient des mutations permettant de faciliter la cristallisation ou l'obtention d'une meilleure résolution lors des expériences de diffraction des rayons X : délétion de l'extension N-terminale (domaine BRCT et linker, résidus 0-241), modification de la boucle 1 (remplacement des résidus 464-472 par l'équivalent chez l'ADN polymérase β : KGET) et mutation C544A). Les mutants λmutR et Polβ-like portaient de plus les résidus impliqués dans le mécanisme de fidélité de l'ADN polymérase β, de façon incomplète pour λmutR (I492R) et complète pour Polβ-like (I492R et résidus 529 à 531 (motif SD2) mutés en NEY). Les mutants λmutK et Ptet-like portaient les équivalents retrouvés chez les ADN polymérases X de *P. tetraurelia*, de façon incomplète pour λmutK (I492K) et complète pour Ptet-like (I492K et E530D).

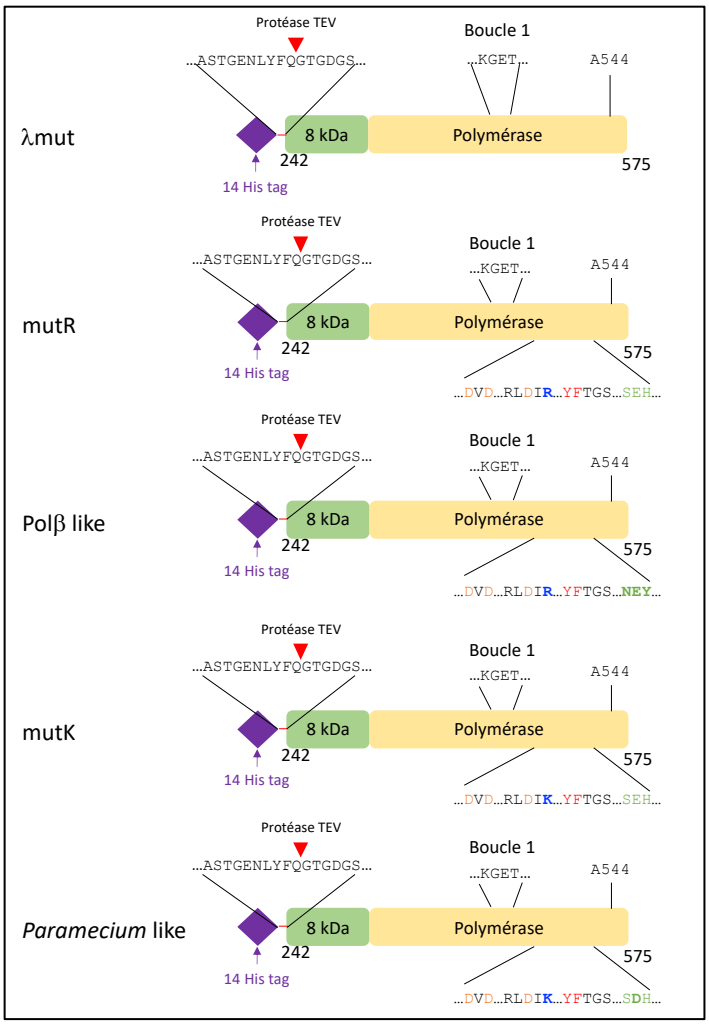


Figure 146 : Schéma des constructions basées sur l'ADN polymérase λ produites et étudiées. Le losange violet indique l'étiquette de 14 histidines. La séquence indiquée au-dessus haut est celle liant cette étiquette à la séquence de la protéine produite. Le site de clivage de la protéase TEV est indiqué avec un triangle rouge. Les domaines 8 kDa et polymérase sont indiqués respectivement dans des rectangles verts et jaunes. Les résidus inclus dans chaque construction sont indiqués sous la construction associée. Les mutations communes à toutes les constructions sont indiquées au-dessus de celles-ci, et leurs équivalents dans la protéine sauvage sont indiqués sur celle-ci. Les mutations spécifiques de chaque mutant sont indiquées en gras sous ce mutant, et leurs équivalents dans la protéine sauvage sont indiqués sur celle-ci. Les résidus d'intérêt sont indiqués : résidus catalytiques en orange, résidus impliqués dans l'activation/inactivation du site catalytique en bleu, résidus du steric gate en rouge, et résidus du motif SD2 en vert.

2.2.2 Expression et purification de constructions mutantes de l'ADN polymérase λ humaine

Les cinq constructions mutantes de l'ADN polymérase λ étudiées ici ont été obtenues par des mutagenèses dirigées successives réalisées à partir du plasmide LS05 ayant permis de produire l'ADN polymérase λ humaine. Une fois les plasmides permettant l'expression des mutants obtenus et leur séquence confirmée, il a été possible de les produire en système bactérien (*E.coli* BL21star(DE3)) par induction avec 1 mM d'IPTG et culture à 20°C. Les purifications ont pu être réalisées, de façon très reproductible entre toutes les constructions. Un exemple de purification, réalisée sur la construction λmutR, est présenté ici.

2.2.2.1 Exemple : la purification de la construction λmutR

La purification a été réalisée en suivant le protocole global énoncé en Matériel et Méthodes. Je présente ici les résultats d'une purification réalisée à partir d'une production de 2L de culture bactérienne. La lyse bactérienne a été réalisée au sonicateur, et la fraction contenant les protéines solubles a été injectée sur une colonne de chromatographie His-Trap. Les fractions présentant la plus haute concentration en protéine (d'après le chromatogramme, les mesures d'absorbance et l'analyse SDS-PAGE) ont été diluées et chargées sur une colonne Héparine. Les fractions contenant la protéine d'intérêt ont été récupérées, puis dialysées une nuit en présence de protéase TEV, de façon à éliminer le tag de 14 histidines. Le dialysat a été injecté sur la colonne His-Trap, et les fractions contenant les protéines non fixées à la résine ont été analysées. Celles contenant la protéine d'intérêt ont été concentrées jusqu'à 10 mg/mL et un volume de 5 mL, et injectées sur une colonne de chromatographie par exclusion stérique (GF) S200 16/60 PG. Les fractions présentant la protéine d'intérêt ont été concentrées jusqu'à 16 mg/mL, et conservées à -20 °C.

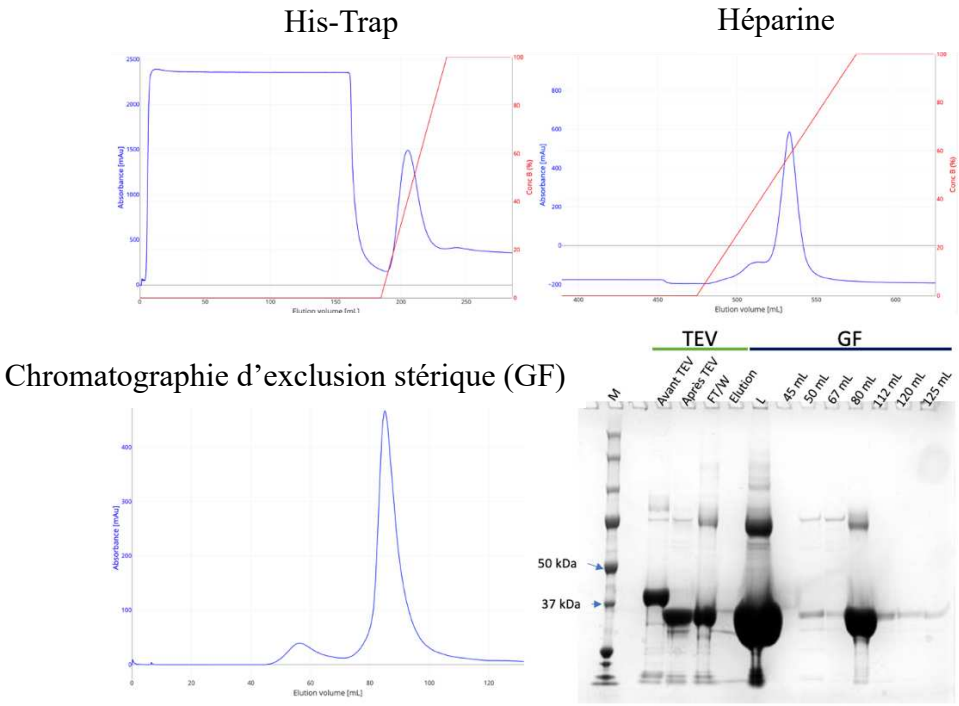


Figure 147 : Résumé d'une purification de Polλ-mutR. En haut à gauche : Chromatogramme de l'étape HisTrap. L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. En haut à droite : Chromatogramme de l'étape Héparine (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. En bas à gauche : Chromatogramme de l'étape de chromatographie d'exclusion stérique (GF). L'absorbance à 280 nm (en mUA) est indiquée en bleu. En bas à droite : SDS-PAGE des étapes de clivage du tag et de gel filtration. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; Les fractions avant et après l'incubation avec la protéase TEV sont indiquées en vert ; FT/W : mélange des fractions FT et W issues de la chromatographie HisTrap réalisée après le clivage par la protéase TEV ; Élution : fraction éluée avec 500 mM d'imidazole ; Les fractions issues de l'étape de gel filtration sont indiquées en noir ; L : load, fraction soluble chargée sur la résine; Les volumes indiquent des fractions choisies d'après le chromatogramme.

2.2.2.2 Récapitulatif

Les purifications de toutes ces constructions ont pu être réalisées de la même manière de façon reproductible. Seule une purification réalisée pour le mutant *Ptet-like* a donné un rendement plus faible, en raison d'un souci rencontré sur l'appareil de purification.

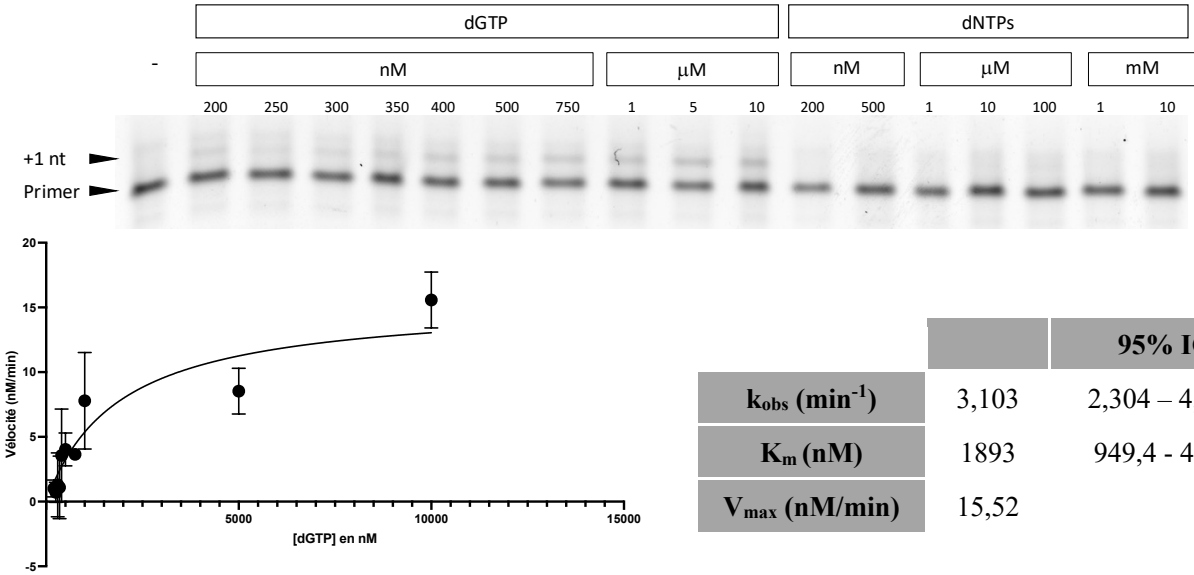
Tableau 11 : Résumé des purifications réalisées pour les constructions basées sur l'ADN polymérases λ humaine. La colonne Volume de culture indique le volume de culture bactérienne utilisé dans la purification associée. La colonne Tag indique la présence du tag 14 histidines (Oui) ou son absence (Non) sur les protéines produites. Le niveau de purification indique à quelle étape la purification a été arrêtée.

Construction	**Volume de culture (L)**	**Aliquots**			**Tag**	**Niveau de purification**	**Utilisation**	**Rendement (mg/L de culture)**
		Nombre	**Volume (µL)**	**Concentration (mg/mL)**				
λmut	4	18	200	16	Non	Chromatographie d'exclusion stérique	Essais de cristallisation	14,4
λmutR	2	6	200	16				9,6
	4	40	100	14,5				14,5
λmutK	2	12	50	27,11				8,13
	4	40	100	24,37				24,37
Ptet-like	4	25	100	1,29				0,8
	4	15	100	20,79				7,79
Polβ–like	2	14	100	10,27				7,19

2.2.3 Caractérisation enzymatique des constructions mutantes de l'ADN polymérase λ humaine: cinétique et fidélité

Une fois les constructions mutantes produites, elles ont été caractérisées d'un point de vue enzymatique. Elles ont été testées de la même façon que les ADN polymérases X de *P. tetraurelia* de façon à obtenir des informations sur leur cinétique d'incorporation d'un nucléotide correct dans un substrat de *gap-filling*, et des informations qualitatives sur leur fidélité.

2.2.3.1 Construction λmut

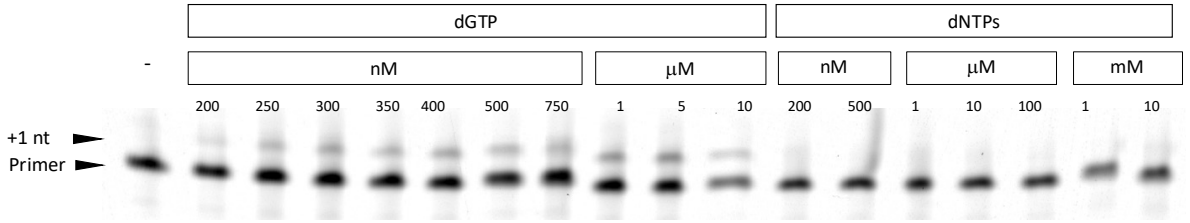


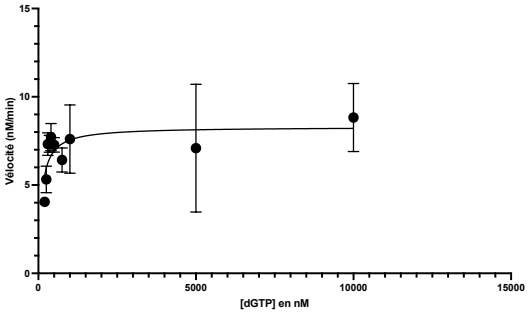
		95% IC
k_{obs} (min^{-1})	3,103	2,304 – 4,447
K_m (nM)	1893	949,4 - 4602
V_{max} (nM/min)	15,52	

Figure 148 : Analyse cinétique de l'incorporation de dGTP dans un substrat de gap-filling par la construction λmut. En haut : Gel urée-PAGE du test d'incorporation de dGTP et de dNTPs incorrects. Les concentrations de chaque dNTP sont indiquées. - : témoin négatif (20 nt). En bas à gauche : représentation de Michaelis de la vélocité de l'enzyme en fonction de la concentration en dGTP. En bas à droite : Valeurs de k_{obs}, K_m et V_{max} obtenues par approximation du modèle de Michaelis Menten avec les données expérimentales (et valeurs limites de l'intervalle de confiance à 95% pour les valeurs de k_{obs} et K_m).

Les expériences réalisées avec du dGTP ont permis d'obtenir la vélocité de la polymérase en fonction de la concentration en dGTP, et ces informations ont été analysées pour approximer le modèle de Michaelis-Menten et obtenir le k_{obs}, le K_m et la V_{max}. Les valeurs obtenues (k_{obs} = 3,103 min-1 ; K_m = 1,893 µM ; V_{max} = 15,52 nM/min) correspondent à celles connues dans la littérature (Garcia-Diaz *et al.*, 2004) pour cette version tronquée de l'ADN polymérase λ (l'efficacité catalytique est ici de 1,64 $nM^{-1}.min^{-1}$) Dans les conditions de ce test, cette enzyme n'a cependant pas montré d'incorporations incorrectes, malgré un long temps d'incubation et de hautes concentrations de dNTPs, contrairement au test en conditions *single turnover* réalisé précédemment (Chapitre 2, 2.1.5, page 194).

2.2.3.2 *Construction λmutR*





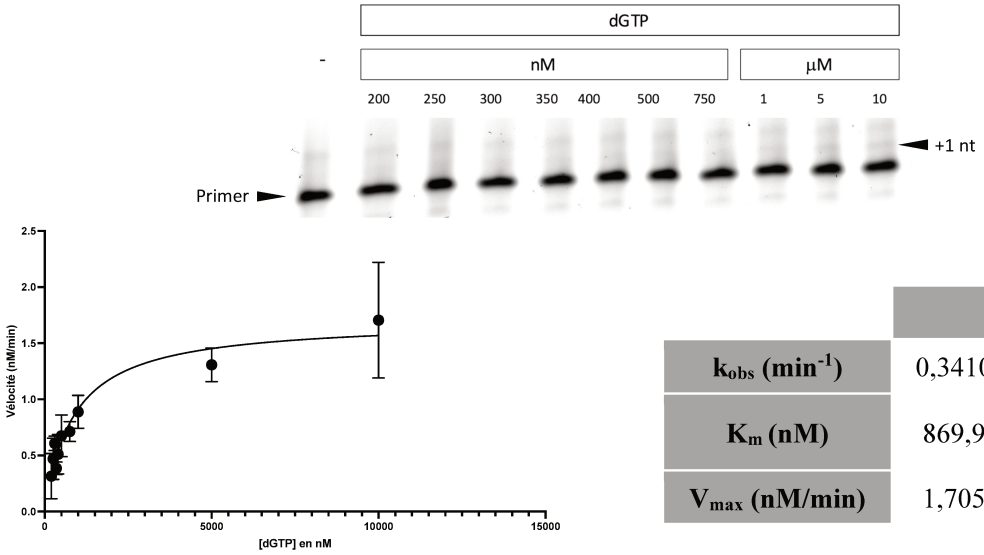
		95% IC
k_{obs} (min^{-1})	1,658	1,396 – 1,946
K_m (nM)	98,22	19,40 – 205,7
V_{max} (nM/min)	8,291	

Figure 149 : Analyse cinétique de l'incorporation de dGTP dans un substrat de gap-filling par la construction λmutR. En haut : Gel urée-PAGE du test d'incorporation de dGTP et de dNTPs incorrects. Les concentrations de chaque dNTP sont indiquées. - : témoin négatif (20 nt). En bas à gauche : représentation de Michaelis de la vélocité de l'enzyme en fonction de la concentration en dGTP. En bas à droite : Valeurs de k_{obs}, K_m et V_{max} obtenues par approximation du modèle de Michaelis Menten avec les données expérimentales (et valeurs limites de l'intervalle de confiance à 95% pour les valeurs de k_{obs} et K_m).

Les expériences réalisées avec ce mutant ont montré que sa vélocité maximale était atteinte avec des concentrations de dGTP plus faibles, et même au maximum cette activité est elle aussi plus faible. Cela se traduit respectivement par un K_m plus faible (92,22 nM), et par une V_{max} et un k_{obs} beaucoup plus faibles, de 8,291 nM/min et 1,658 min^{-1} respectivement. L'efficacité catalytique de ce mutant est de 16,91 $nM^{-1}.min^{-1}$, bien supérieure à celle de la construction « sauvage », λmut. Là encore, aucune incorporation incorrecte n'a pu être observée.

2.2.3.3 Construction Polβ-like

La caractérisation cinétique de ce mutant avec le dGTP a montré une quasi-absence d'activité enzymatique, c'est pourquoi aucun test avec les dNTPs incorrects n'a été réalisé.

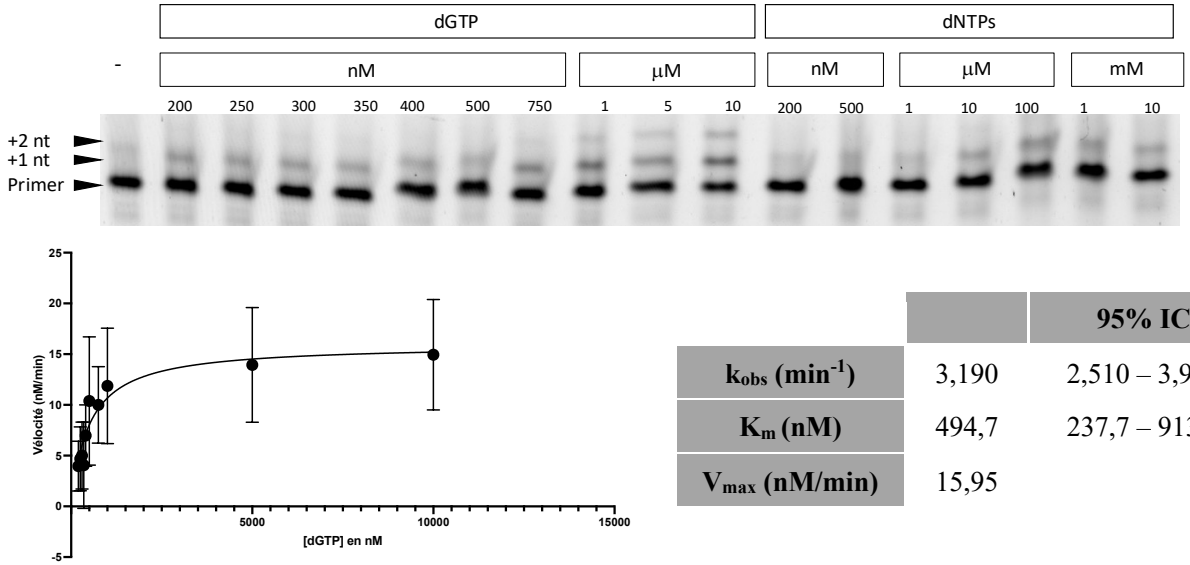


		95% IC
k_{obs} (min^{-1})	0,3410	0,2935– 0,3932
K_m (nM)	869,9	587,8 – 1277
V_{max} (nM/min)	1,705	

Figure 150 : Analyse cinétique de l'incorporation de dGTP dans un substrat de gap-filling par la construction Polβ–like. En haut : Gel urée-PAGE du test d'incorporation de dGTP. Les concentrations de dGTP sont indiquées. - : témoin négatif (20 nt). En bas à gauche : représentation de Michaelis de la vélocité de l'enzyme en fonction de la concentration en dGTP. En bas à droite : Valeurs de k_{obs}, K_m et V_{max} obtenues par approximation du modèle de Michaelis Menten avec les données expérimentales (et valeurs limites de l'intervalle de confiance à 95% pour les valeurs de k_{obs} et K_m).

Ce mutant porte l'ensemble des résidus qui confèrent à l'ADN polymérase β sa fidélité. Cependant, on observe ici une activité enzymatique très faible, avec des valeurs de k_{obs} et V_{max} très faibles en comparaison avec tous les autres mutants, ainsi que l'ADN polymérase β (Chagovetz *et al.*, 1997). Il semble donc que l'introduction du mécanisme « ouvert-fermé » de l'ADN polymérase β chez l'ADN polymérase λ ne fait que réduire drastiquement son activité. Cependant, le K_M obtenu (869,9 nM) est intermédiaire entre celui du WT (1,8 µM) et ceux des mutants λmutK et Ptet-like (respectivement de 494,7 nM et 399,3 nM). L'efficacité catalytique de cette construction est de 3.10^{-4} $nM^{-1}.min^{-1}$.

2.2.3.4 Construction λmutK

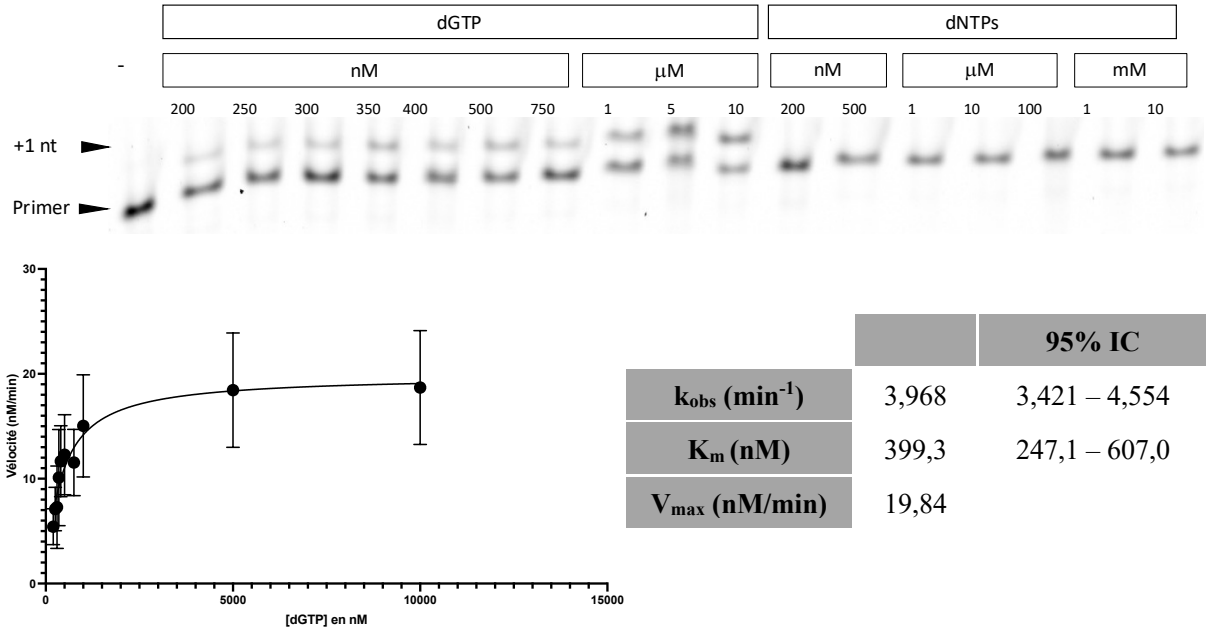


		95% IC
k_{obs} (min^{-1})	3,190	2,510 – 3,952
K_m (nM)	494,7	237,7 – 913,3
V_{max} (nM/min)	15,95	

Figure 151 : Analyse cinétique de l'incorporation de dGTP dans un substrat de gap-filling par la construction λmutK. En haut : Gel urée-PAGE du test d'incorporation de dGTP et de dNTPs incorrects. Les concentrations de chaque dNTP sont indiquées. - : témoin négatif (20 nt). En bas à gauche : représentation de Michaelis de la vélocité de l'enzyme en fonction de la concentration en dGTP. En bas à droite : Valeurs de k_{obs}, K_m et V_{max} obtenues par approximation du modèle de Michaelis Menten avec les données expérimentales (et valeurs limites de l'intervalle de confiance à 95% pour les valeurs de k_{obs} et K_m).

Ce mutant, comme la construction λmutR, ne présente qu'une partie des mutations supposées conférer le mécanisme de transition « actif-inactif » au site catalytique de l'ADN polymérase λ. L'expérience réalisée ici montre des erreurs d'incorporation, y compris avec le dGTP seul. En effet, ce mutant semble ajouter du dGTP, mais au lieu d'ajouter un nucléotide, il en ajoute deux. En plus d'être surnuméraire dans le contexte de gap-filling, le second nucléotide ajouté est incorrect, puisqu'il est supposé être hybridé à un dA. En présence des nucléotides incorrects, ce mutant semble également pouvoir étendre l'amorce d'une base, de façon peu efficace. L'incorporation de 2 dGTP a permis d'obtenir les paramètres cinétiques de ce mutant, qui sont proches de ceux du mutant λmut, sauf le K_m qui est 3,8 fois plus faible. Son efficacité catalytique est donc de 6,46 nM^{-1}.min^{-1}.

2.2.3.5 Construction Paramecium Ptet-like



		95% IC
k_{obs} (min^{-1})	3,968	3,421 – 4,554
K_m (nM)	399,3	247,1 – 607,0
V_{max} (nM/min)	19,84	

Figure 152 : Analyse cinétique de l'incorporation de dGTP dans un substrat de gap-filling par Polλ Ptet-like. En haut : Gel urée-PAGE du test d'incorporation de dGTP et de dNTPs incorrects. Les concentrations de chaque dNTP sont indiquées. - : témoin négatif (20 nt). En bas à gauche : représentation de Michaelis de la vélocité de l'enzyme en fonction de la concentration en dGTP. En bas à droite : Valeurs de k_{obs}, K_m et V_{max} obtenues par approximation du modèle de Michaelis Menten avec les données expérimentales (et valeurs limites de l'intervalle de confiance à 95% pour les valeurs de k_{obs} et K_m).

Ce mutant porte l'ensemble des résidus supposés permettre aux ADN polymérases X de *Paramecium* d'activer ou non le site catalytique. Comme pour les constructions λmut et λmutR, on observe ici une incorporation unique du dGTP, qui a permis d'obtenir les paramètres cinétiques de l'enzyme. Ceux-ci sont du même ordre que pour la construction sauvage, sauf concernant le K_m qui est 4,7 fois plus faible, ce qui donne une efficacité catalytique supérieure, de 9,94 nM^{-1}.min^{-1}. Ici, aucune incorporation incorrecte n'est observée, alors que la seule différence entre ce mutant et le mutant λmutK est une mutation d'un glutamate vers un aspartate dans le motif SD2.

2.2.3.6 Discussion

D'un point de vue qualitatif, seul le mutant λmutK a une activité particulière puisque c'est le seul à faire des incorporations erronées dans ces tests. Deux erreurs d'incorporation sont observées : l'incorporation de nucléotides incorrects dans le substrat gap-filling ; et l'incorporation de 2 dGTP au lieu d'un, pouvant supposer que ce mutant a une activité de déplacement de brin augmentée, et que sa capacité à incorporer des nucléotides incorrects entraine l'incorporation d'un dGTP au lieu d'un dTTP face à un dA. Cependant, le mutant Ptet-

like ne fait aucune de ces erreurs, ce qui laisse penser que la mutation du motif SD2 de SEH vers SDH suffit à empêcher les erreurs observées.

D'un point de vue quantitatif, plusieurs éléments sont à souligner, et permettent d'établir des hypothèses sur les effets des mutations testées.

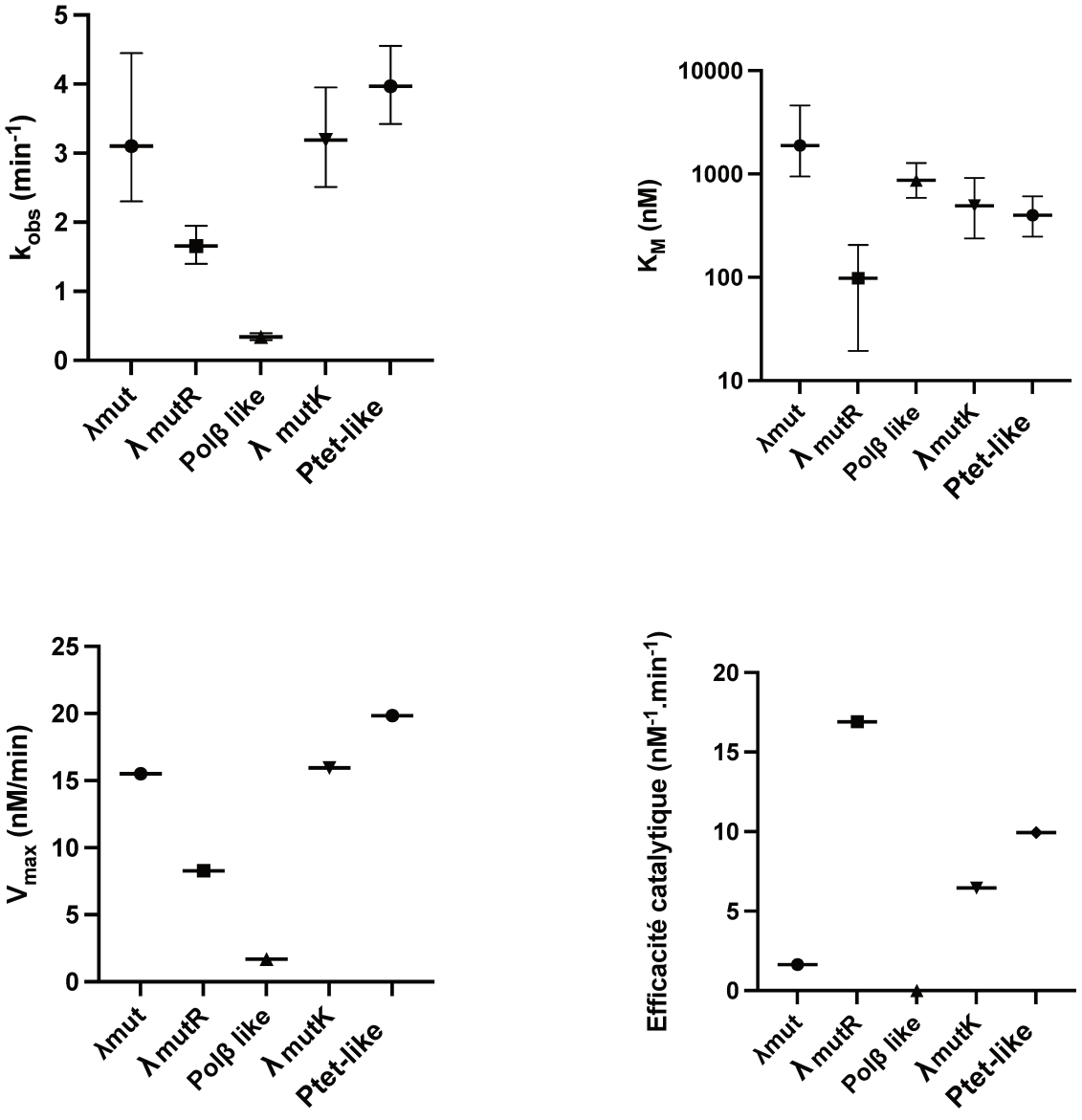


Figure 153 Comparaison des paramètres cinétiques obtenus pour les 5 constructions de l'ADN polymérase λ testées. En haut à gauche : comparaison des k_{obs} ; En haut à droite : comparaison des K_m, En bas à gauche : comparaison des V_{max} ; En bas à droite : comparaison des efficacités catalytiques. Les barres d'erreurs correspondent aux intervalles de confiance à 95% des valeurs de k_{obs} et K_m.

Le mutant λmutR a des valeurs de k_{obs}, de K_m et de V_{max} plus faibles que la construction λmut. Celle-ci mime correctement l'ADN polymérase λ, ce qui indique que les mutations réalisées pour faciliter l'étude cristallographique n'ont pas d'impact sur la catalyse. Les valeurs

de k_{obs} et de V_{max} de la construction λmutR indiquent que l'enzyme réalise sa catalyse lentement, et sa valeur de K_m peut indiquer une grande affinité de l'enzyme pour le dGTP. Cela peut être un signe d'une très forte stabilité du complexe enzyme-substrat, qui entraine un ralentissement de la catalyse pour former le produit et le séparer de l'enzyme. Ce mutant est cependant celui qui présente la meilleure efficacité catalytique.

Le mutant Polβ-like, supposé permettre une activité normale (voire meilleure, en s'approchant du fonctionnement de l'ADN polymérase β) en présence du nucléotide entrant correct, présente les k_{obs} et V_{max} les plus bas, ce qui indique que son activité est très fortement réduite. Sa faible activité se traduit par son efficacité catalytique quasi nulle et peut avoir plusieurs origines : soit le mécanisme introduit ne fonctionne pas comme attendu, et un des résidus catalytiques pourrait être bloqué par l'arginine du motif RLDIR ; soit les substrats ne peuvent pas se placer correctement dans le site actif, ce qui peut limiter l'activité.

Le mutant λmutK présente des paramètres cinétiques proches de ceux de la construction λmut, en particulier concernant le k_{obs} et la V_{max}. Cela suppose que la mutation réalisée n'affecte pas directement la vitesse de catalyse. Cependant, le K_m de ce mutant est plus faible que celui de la construction λmut, ce qui suppose une plus forte affinité pour le dGTP. Cela peut être le signe que ce substrat peut plus facilement entrer dans le site catalytique, en particulier si celui-ci est sous une forme active. Ce mutant a donc une meilleure efficacité catalytique que la construction λmut, mais il fait des incorporations erronées. On peut alors supposer que la simple mutation du motif RLDII vers RLDIK permet d'obtenir une forme active du site catalytique, mais que celui-ci contrôle mal les nucléotides entrants.

Enfin, le mutant Ptet-like présente les valeurs de k_{obs} et de V_{max} les plus élevées, ce qui suppose que c'est la construction testée la plus rapide pour réaliser sa catalyse en présence de dGTP. Son K_m est cependant très proche de celui du mutant λmutK, ce qui peut être le signe de l'existence d'une forme du site actif permettant une fixation facilitée du substrat, qui se traduit par un K_m plus faible que pour la construction λmut. Le fait que ce mutant soit parmi ceux testés le plus rapide tout en conservant une bonne fidélité peut être le signe que la mutation additionnelle du motif SD2 permet de garantir une bonne fidélité. On peut supposer que ces mutations permettent bien une alternance de formes du site actif : une forme inactive facilitant la fixation du dGTP, et une forme active facilitant la catalyse, et prenant fin après un cycle catalytique complet et correct.

2.2.4 Étude structurale de l'impact des mutations réalisées sur la catalyse des constructions mutantes de l'ADN polymérase λ

Afin de mieux comprendre les résultats obtenus lors des expériences de caractérisation enzymatique, nous avons résolu les structures cristallographiques des quatre constructions mutées de l'ADN polymérase λ.

2.2.4.1 Conditions d'obtention des cristaux

Pour chacun des mutants étudiés, des cristaux ont pu être obtenus dans une des matrices préparées (voir Chapitre 2, 1.2.5.1, page 156). Ils ont ensuite été pêchés et congelés après trempage dans une solution cryoprotectrice.

Tableau 12 : Conditions d'obtention des cristaux des différentes constructions de l'ADN polymérase λ dont la structure a été obtenue

Mutant	**Préparation**	**Condition**	**Cryoprotectant**
λmutR	16 mg/mL + ADN + dTTP + $CaCl_2$	20 mM bicine pH 7,5, 300 mM Na-K tartrate, 22,5% PEG Smear high	Ethylène glycol
Polβ-like	16 mg/mL + ADN + dTTP + $CaCl_2$	20 mM bicine pH 7,5, 300 mM Na-K tartrate, 20% PEG 1000	Ethylène glycol
λmutK	16 mg/mL + ADN + dTTP + $CaCl_2$	20 mM bicine pH 7,5, 300 mM Na-K tartrate, 17,5% PEG 20K 10 mM acétate de praséodyme	Glycérol
Ptet-like	16 mg/mL + ADN + dTTP + $CaCl_2$	20 mM bicine pH 7,5, 300 mM Na-K tartrate, 20% PEG 20K	Ethylène glycol
	16 mg/mL + ADN + dCTP (incorrect) + $CaCl_2$	20 mM bicine pH 7,5, 300 mM Na-K tartrate, 14% PEG 10K	Ethylène glycol

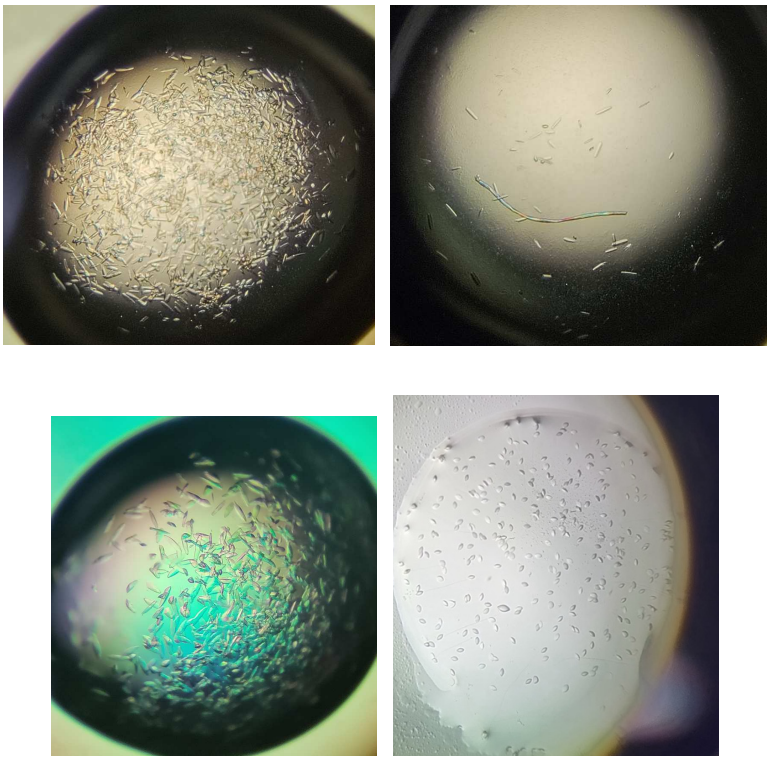


Figure 154 : Exemples de cristaux obtenus dans différentes conditions testées pour les différents mutants : En haut à gauche : λmutR (condition B2, matrice 1) ; En haut à droite : λmutR, (condition D3, matrice 1) ; En bas à gauche : λmutK, (condition D4, matrice 1) ; En bas à droite : Ptet-like (dCTP) (condition C1, matrice 1).

2.2.4.2 Construction des modèles

Pour tous les mutants testés, un remplacement moléculaire a été réalisé en utilisant le modèle PDB 7M43, qui correspond à la structure d'une version sauvage de l'ADN polymérase λ obtenue dans les mêmes conditions que les mutants étudiés.

Les structures des 4 mutants ont ainsi pu être obtenues (dans deux conditions pour le mutant Ptet-like). Les statistiques des jeux de données et des affinements sont présentées dans le tableau 13. Le groupe d'espace obtenu a été le même pour la majorité des complexes étudiés : P 21 21 21 pour les mutants λmutR (+dTTP), λmutK (+dTTP), Ptet-like (+dTTP et +dCTP). Cependant, pour le mutant Polβ-like, un groupe d'espace différent a été obtenu (R 3 2), à partir d'un jeu de données anisotrope, traité par le Dr Pierre Legrand sur le serveur Staraniso.

Les données ont pu être obtenues à haute résolution pour tous les mutants, avec une résolution maximale de 2,12 Å pour le mutant Ptet-like (en présence de dTTP). Les données ont pu être affinées dans tous les cas jusqu'à des valeurs satisfaisantes de R_{work} et R_{free}.

Tableau 13 : Statistiques de collecte des données de diffraction et d'affinement des modèles

	λmutR (dTTP)	λSD2β	λmutK d(TTP)	λSD2Ptet (dTTP)	λSD2Ptet (dCTP)
Collecte des données					
Groupe d'espace	P 21 21 21	H32	P 21 21 21	P 21 21 21	P 21 21 21
a. b. c (Å)	56.28 62.51 140.22	149.904 149.904 272.154	56.03 62.49 141.37	56.4 62.47 139.57	56.35 62.76 139.72
α. β. γ (°)	90 90 90	90 90 120	90 90 90	90 90 90	90 90 90
Longueur d'onde (Å)	0.9801	0.9801	0.9801	0.9801	0.9801
Résolution (Å)	46.74 - 2.32 (2.38 - 2.32)	117.173 - 3.543 (4.018 - 3.543)	24.04 - 1.916 (2.118 - 1.916)	46.55 - 2.12 (2.18 - 2.12)	46.69 - 2.737 (2.987 - 2.737)
*Limite estimée de résolution (Å)**		5.491 5.491 3.344			
R-pim	0.040 (0.629)	1.514 (3.179)	0.052 (0.807)	0.038 (0.69)	0.204 (0.845)
Complétude (%)	99.9 (98.8)	100 (100)	93.3 (63.9)	99.8 (97.5)	86.4 (48.5)
Multiplicité	13.3 (13.7)	18.2 (19.2)	11.9 (11.7)	13.4 (13.7)	9.3 (9.2)
I/σ(I)	17.85 (1.99)	5.00 (0.70)	11.0 (1.70)	16.6 (1.6)	4.6 (1.40)
CC1/2	99.9% (82.0%)	65.7 % (63.9%)	99.6% (50.4%)	99.9% (79.4%)	97.5% (54.8%)
*R-pim**		0.658 (1.192)			
*Complétude (%)**		91.7 (74.2)			
*Multiplicité**		17.7 (15.2)			
*I/σ(I)**		7.80 (1.80)			
*CC1/2**		80.5% (74.7%)			
Affinement					
Nombre de réflections	22111	5991	25444	28561	8572
R_{work} / *Rfree*	0.208 / 0.256	0.263 / 0.292	0.204 / 0.245	0.193 / 0.224	0.209 / 0.259
Nombre d'atomes non hydrogène					
Macromolécules	2459	5036	2312	2514	2497
Ligands	47	0	31	47	28
Solvant	303	21	278	306	76
Géométrie protéique					
RMSD - liaisons (Å)	0.008	0.007	0.009	0.009	0.007
RMSD - angles (°)	0.89	0.81	0.87	0.95	0.83
*Ramachandran favorisés(%)***	97.47	96.54	96.04	94.10	96.54
*Ramachandran aberrants (%)***	0.00	0.31	0.33	0.00	0.31
*Rotamères aberrants (%)***	1.52	5.63	0.44	1.85	5.64
Clashscore	4	10	4	4	8
B-factor (Å²)					
B-factor moyen	57.19	79.25	73.50	57.10	52.90
Macromolécules	60.92	79.81	49.67	60.89	43.27
Ligands	49.62	0	117.7	49.50	115.40
Solvant	61.03	78.68	53.14	60.92	20.66

2.2.4.3 Analyse des modèles structuraux obtenus par cristallographie

L'analyse des structures obtenues a été réalisée en les comparant à la structure de la construction λmut publiée en 2022 (Jamsen *et al.*, 2022) (PDB 7M43). L'objectif ici était de conférer à l'ADN polymérase λ le mécanisme d'alternance entre des formes ouverte/fermée du domaine catalytique de l'ADN polymérase β (et le mécanisme équivalent des ADN polymérases X de *P. tetraurelia*). Théoriquement, l'obtention de ces structures sans nucléotide entrant était supposée permettre d'obtenir une forme inactive et ouverte du domaine catalytique. Malheureusement, en absence de dTTP, aucune structure n'a pu être obtenue pour les mutants λmutR, λmutK et Ptet-like. Seules des structures en présence du nucléotide entrant ont été obtenues pour ces constructions, et toutes étaient sous forme fermée.

2.2.4.3.1 Structure de la construction λmutR en présence de dTTP

Le repliement global du mutant λmutR est très proche de celui observé pour la construction λmut : on obtient sur ChimeraX un RMSD de 0,380 Å. Cette valeur représente l'écart de distance moyen entre les carbones α homologues des deux structures (ici entre chacun des 320 carbones α présents dans chaque structure). Cette valeur est très faible, donc on peut considérer que le repliement global est le même, et qu'aucun domaine ou structure secondaire n'est déplacé chez ce mutant.

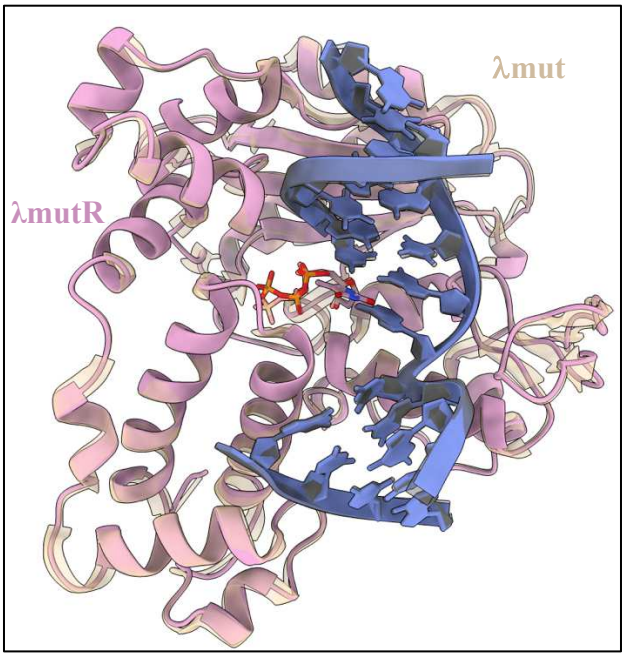


Figure 155 : Superposition des structures de la construction λmut (PDB 7M43, en beige, ADN en gris) et du mutant λmutR (en rose, ADN en bleu).

L'analyse du cœur catalytique de l'enzyme permet de mieux comprendre l'activité enzymatique observée. En effet, le résidu catalytique D490 est détourné du site actif, et lié par un pont salin à R492, l'arginine introduite par la mutation. Le reste des résidus d'intérêt ne semble cependant pas subir de changement conformationnel, tout comme l'ADN et le nucléotide entrant.

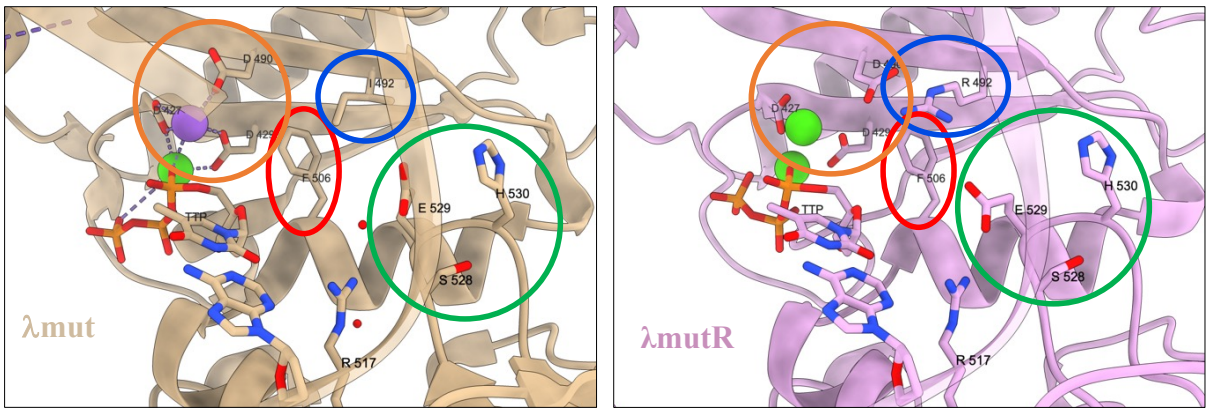


Figure 156 : Comparaison des résidus d'intérêt au sein du site actif de la construction λmut (à gauche, en beige) et du mutant λmutR (à droite, en rose). Le nucléotide template est affiché, le nucléotide entrant est indiqué (TTP), et les résidus sont indiqués et entourés selon leur rôle : résidus catalytiques en orange ; résidu du steric gate en rouge, résidu d'intérêt du motif RxDx(Φ/+) en bleu, et motif SD2 en vert. Les ions placés dans le site catalytique sont indiqués en violet (Na) et vert (Ca).

Ici, il semble donc qu'on observe un phénotype proche de celui de l'ADN polymérase β, avec un résidu aspartate détourné du site actif par l'arginine du motif RxDx(Φ/+). Cependant, chez l'ADN polymérase β, ce phénotype n'est observé qu'en absence d'un nucléotide correct, dans la forme ouverte de l'enzyme, et c'est un autre aspartate qui est lié à l'arginine (D192, équivalent de D429 chez l'ADN polymérase λ). Dans les conditions utilisées ici (en présence d'ADN et de dTTP), on s'attendrait donc à observer une conformation active, avec l'arginine orientée vers le motif SD2.

Ces observations peuvent expliquer le résultat obtenu lors des tests enzymatiques de ce mutant. Pour rappel, le mutant λmutR montrait dans ces essais des paramètres cinétiques (k_{obs}, V_{max} et K_m) plus faibles que la construction λmut mimant l'ADN polymérase λ sauvage. L'observation d'un aspartate catalytique détourné du site actif peut expliquer ces changements : le nucléotide peut entrer dans le site actif et s'y stabiliser parfaitement normalement, mais la catalyse ne peut pas avoir lieu normalement car un des aspartates catalytiques est détourné du site actif. Tout cela se traduit par une baisse d'activité enzymatique (k_{obs} et V_{max} faibles) et une augmentation de l'affinité de la polymérase pour le nucléotide entrant (K_m faible).

L'hypothèse initiale en utilisant ce mutant reposait sur l'idée que le mécanisme de fidélité de l'ADN polymérase β nécessite cette arginine, et que le motif SD2 (SEH) de l'ADN polymérase λ était équivalent à celui de l'ADN polymérase β (NEY) pour se lier à l'arginine dans un contexte de nucléotide correct, ce qui permettrait à l'aspartate catalytique détourné de revenir dans le site actif pour réaliser la catalyse. La structure obtenue indique que dans ce contexte, il n'est pas possible pour l'arginine de quitter la liaison formée avec l'aspartate. J'ai donc émis l'hypothèse que le motif SD2 de l'ADN polymérase λ n'était pas assez électrophile pour attirer l'arginine, ce qui a pu être testé avec le mutant Polβ-like.

2.2.4.3.2 Structure de la construction Polβ-like en présence de dTTP

Comme indiqué précédemment, la structure de ce mutant a pu être obtenue, mais de façon différente des autres mutants. En effet, les données de diffraction obtenues étaient anisotropes et ont dû être traitées avant d'être utilisées pour obtenir une structure à haute résolution. Cette construction montrait deux différences avec les autres mutants : le groupe d'espace obtenu et le nombre de molécules par unité asymétrique. Pour tous les autres mutants, le groupe d'espace obtenu était P 21 21 21, avec une molécule par unité asymétrique. Ici, le groupe d'espace est R 3 2, avec deux molécules par unité asymétrique. Le remplacement moléculaire et l'affinement ont donc été réalisés pour les deux molécules présentes.

L'affinement a montré que le nucléotide entrant, bien que présent dans l'expérience de cristallogenèse, est absent du site actif des deux structures présentes dans l'unité asymétrique. La superposition directe des deux molécules avec la structure de référence (PDB 7M43) donne des RMSD globaux (incluant tous les carbones α des deux protéines) de 1,499 Å et 1,058 Å, respectivement, et pour des alignements optimisés (n'incluant pas les résidus déplacés de plus de 2 Å) ces RMSD sont respectivement de 0,562 Å (pour 308 Cα) et 0,576 Å (pour 311 Cα). Cela indique que l'ADN polymérase λ mutée n'adopte pas une forme ouverte, malgré l'absence de dNTP dans le site actif, mais qu'une partie bien spécifique des résidus diverge : il s'agit essentiellement de la boucle 3, qui dans les deux structures du mutant n'est pas située au contact de l'ADN (contrairement à toutes les autres structures montrées ici), et présente les *B-factors* les plus élevés au sein des deux structures (de 130 à 260 $Å^2$ environ), ce qui indique une grande flexibilité. De plus, l'ADN instructeur des structures Polβ-like semble ne pas se superposer avec celui de la construction λmut au niveau du site actif. En particulier, le nucléotide template dA semble déplacé en amont du site catalytique.

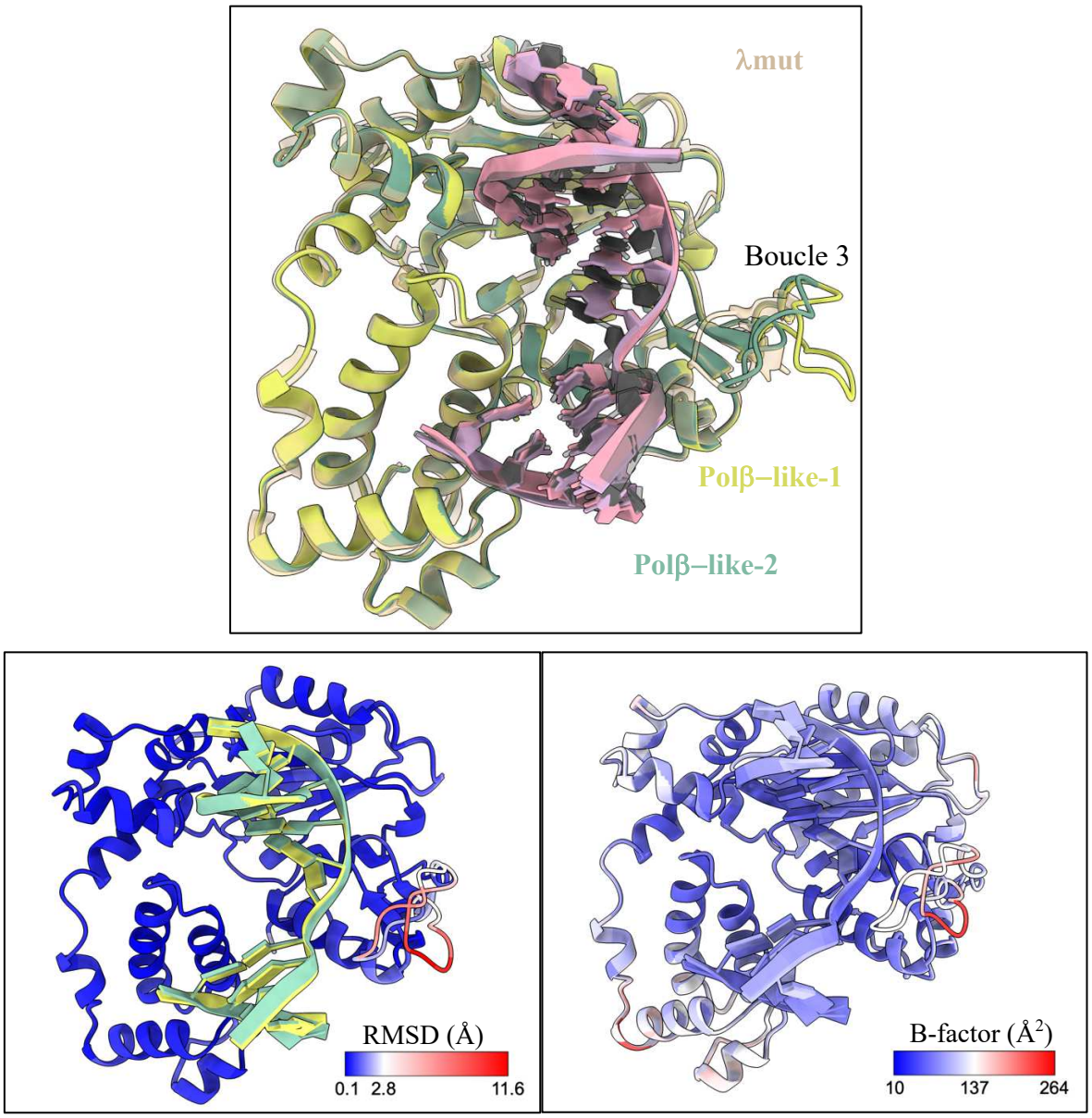


Figure 157 : Comparaison des structures de la construction λmut (PDB 7M43, en beige, ADN en gris) et des deux structures obtenues pour le mutant Polβ-like (en jaune et vert, ADN en rose et mauve en haut). En haut : superposition des trois structures. En bas à gauche : Superposition des deux structures obtenues pour le mutant Polβ–like, colorées selon leur RMSD par rapport à la structure de référence (PDB 7M43). L'échelle est indiquée en bas à gauche, en Å. En bas à droite : superposition des deux structures obtenues pour le mutant Polβ–like, colorées selon leur B-factor. L'échelle est indiquée en bas à droite, en Å².

Enfin, les résidus ciblés par les mutations adoptent des conformations différentes dans les deux molécules de l'unité asymétrique. Dans celle pour laquelle la boucle 3 est la plus proche de l'ADN, (Polβ–like 2, en vert sur les figures), les trois aspartates catalytiques sont orientés vers le site catalytique, alors qu'aucun nucléotide n'est présent. La phénylalanine du *steric gate* adopte quant à elle une orientation différente de celles vues précédemment : ici, la conformation locale s'apparente plutôt à celle observée chez les ADN polymérases λ et β lorsqu'elles sont sous forme inactive (forme ouverte), ce qui (chez l'ADN polymérase β) permet une interaction entre l'arginine du motif RIDIR et un résidu aspartate catalytique. Le

groupement guanidinium de cette arginine se place entre les aspartates catalytiques et le motif SD2, mais un de ses atomes d'azotes η est à une distance du motif SD2 et des aspartates catalytiques qui peut permettre de former des liaisons électrostatiques avec les deux motifs (3,5 Å). Dans la seconde molécule (Polβ–like 1, en jaune sur les figures), pour laquelle la boucle 3 est à distance de l'ADN, un des aspartates catalytiques est détourné du site actif, alors que l'arginine 492 est orientée clairement vers le motif SD2, à une distance suffisamment faible pour permettre la formation de liaisons électrostatiques avec le glutamate (3,5 Å) et la tyrosine (3,8 Å). Là encore, le placement de la phénylalanine 506 du *steric gate* est proche de sa forme inactive et ouverte rencontrée chez les ADN polymérases λ et β. Comme indiqué précédemment, dans ces deux structures le nucléotide entrant est absent, et l'ADN template n'est pas placé correctement, en particulier le nucléotide template dA qui est à distance de son placement optimal et n'est pas orienté correctement vers le site catalytique.

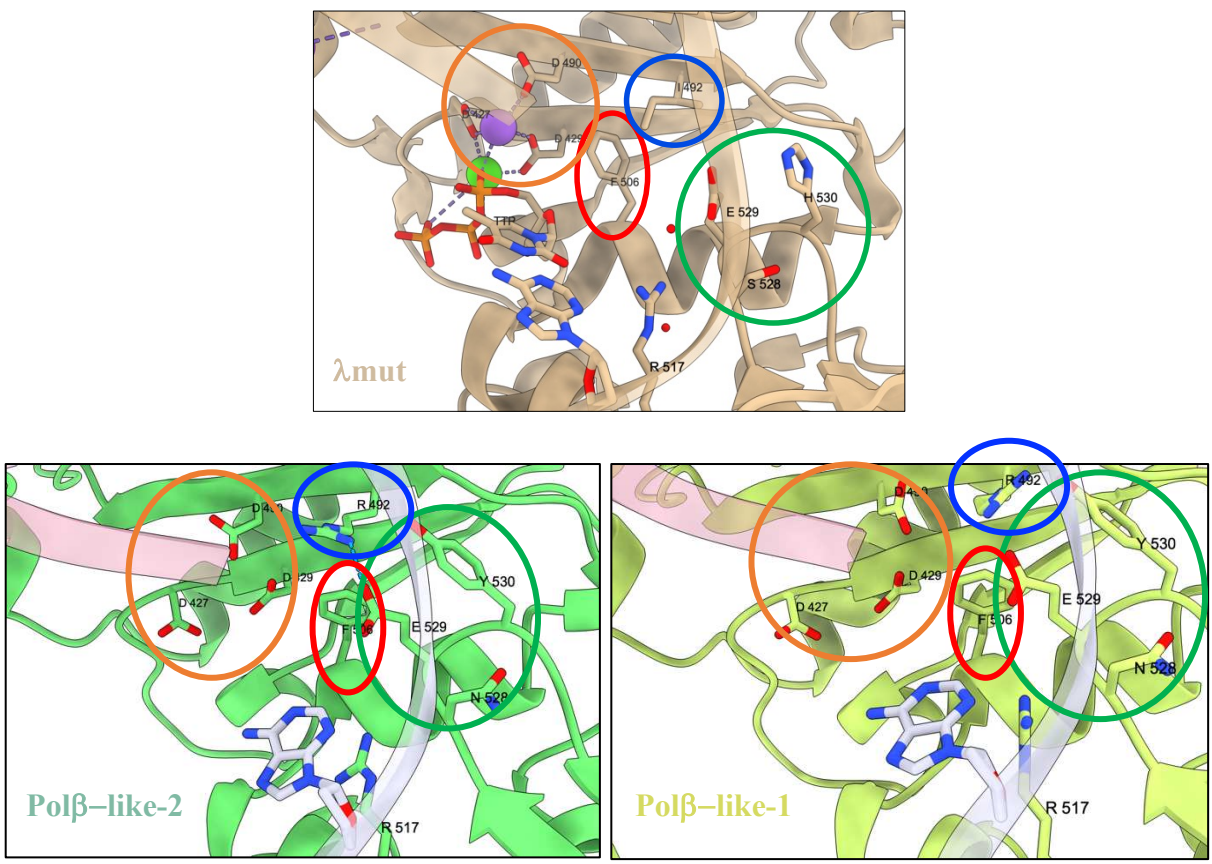


Figure 158 : Comparaison des résidus d'intérêt au sein du site actif de la construction λmut (en haut, en beige) et les structures du mutant Polβ–like (en bas, en vert et en jaune). Le nucléotide template est affiché, le nucléotide entrant est indiqué (TTP) lorsqu'il est présent, et les résidus sont indiqués et entourés selon leur rôle : résidus catalytiques en orange ; résidu du steric gate en rouge, résidu d'intérêt du motif RxDx(Φ/+) en bleu, et motif SD2 en vert. Les ions placés dans le site catalytique sont indiqués en violet (Na) et vert (Ca).

L'ensemble de ces observations peut être mis en lien avec les résultats obtenus lors des tests enzymatiques : pour rappel, la construction Polβ-like a montré une activité quasi nulle. Il

semble que la perte d'activité observée soit due à une difficulté pour obtenir une bonne géométrie entre l'ADN template et le nucléotide entrant, qui ne peut alors pas être stabilisé.

D'après les observations faites ici et les connaissances actuelles sur l'ADN polymérase λ, il est possible que ce mauvais placement de l'ADN soit lié à la grande flexibilité observée pour la boucle 3. Celle-ci n'est pas au contact de l'ADN et ne peut pas stabiliser son placement dans le site actif, ce qui pourrait empêcher la catalyse. Cependant, dans le mutant λmutR, cette boucle était placée de façon normale, et l'ADN aussi. Cela suppose que la mutation supplémentaire réalisée (mutation du motif SD2 vers NEY) a eu un impact sur la forme prise par cette boucle 3. Comme indiqué précédemment, le rôle de cette boucle est encore mal connu, mais elle pourrait être liée à un mécanisme de stabilisation de l'ADN dans le site actif (Jamsen *et al.*, 2022). Il semble donc que ce mécanisme de stabilisation de l'ADN polymérase λ (boucle 3) et le mécanisme de fidélité de l'ADN polymérase β (alternance de formes ouverte/fermée) ne soient pas compatibles dans le contexte du reste de la séquence de l'ADN polymérase λ: l'évolution a optimisé la séquence de l'ADN polymérase λ en vue de stabiliser l'ADN au sein du site actif à l'aide de sa boucle 3 et non pas de permettre l'alternance des formes ouverte et fermée, comme dans l'ADN polymérase β ; la conformation fermée adoptée tout au long de la catalyse par l'ADN polymérase λ « force » l'arginine 492 à se lier au motif SD2 (comme dans la forme fermée et active de l'ADN polymérase β). Il faudrait donc certainement plus de mutations que les deux utilisées ici pour autoriser l'alternance de formes ouvertes et fermées chez l'ADN polymérase λ.

2.2.4.3.3 Structure de la construction λmutK en présence de dTTP

La structure globale de ce mutant est très proche de celle de la construction λmut : la superposition des structures λmut et λmutK indique un RMSD de 0,389 Å.

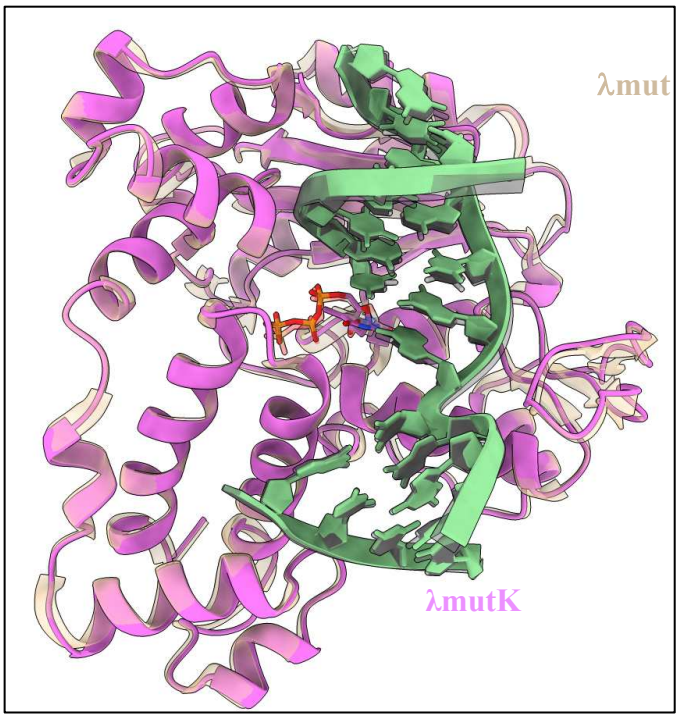


Figure 159 : Superposition des structures de la construction λmut (PDB 7M43, en beige, ADN en gris) et du mutant λmutK (en magenta, ADN en vert).

Il semble que le résidu muté (K492) ait un impact sur la forme du site actif. La phénylalanine du *steric gate* adopte comme pour le mutant Polβ-like une forme « inactive ». Cependant, ici la lysine 492 est orientée vers le motif SD2, et son groupement amine ζ est situé à 3,7 Å du carboxylate du glutamate 529, ce qui permet d'établir une liaison électrostatique. Par ailleurs, ce résidu E529 est orienté vers cette lysine.

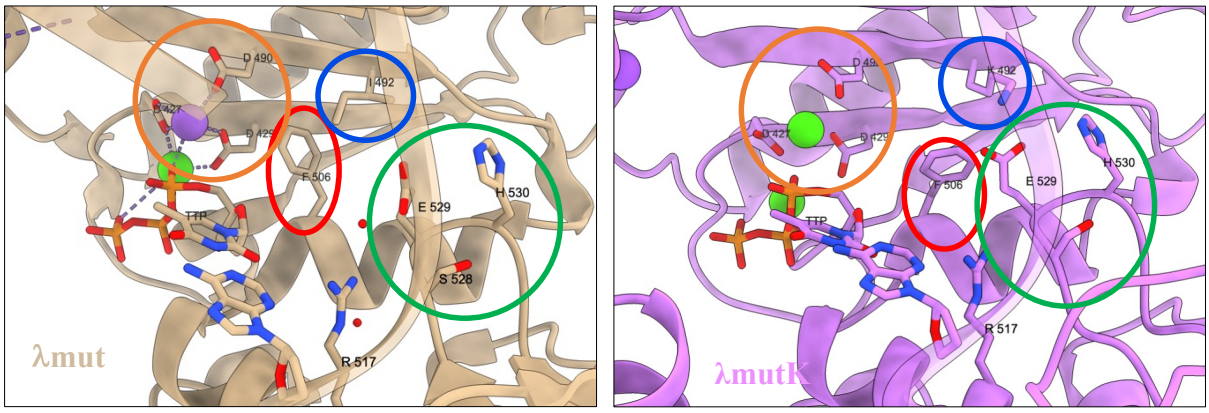


Figure 160 : Comparaison des résidus d'intérêt au sein du site actif de la construction λmut (à gauche, en beige) et du mutant λmutK (à droite, en magenta). Le nucléotide template est affiché, le nucléotide entrant est indiqué (TTP), et les résidus sont indiqués et entourés selon leur rôle : résidus catalytiques en orange ; résidu du steric gate en rouge, résidu d'intérêt du motif RxDx(Φ/+) en bleu, et motif SD2 en vert. Les ions placés dans le site catalytique sont indiqués en violet (Na) et vert (Ca).

Ces observations peuvent expliquer les résultats obtenus dans les tests enzymatiques réalisés avec ce mutant. Pour rappel, cette construction était la seule à faire des incorporations erronées, et montrait des caractéristiques cinétiques similaires à λmut, sauf concernant le K_m,

qui était plus faible. Cela indique que le nucléotide se fixe plus facilement dans le site actif, mais que la catalyse n'est pas plus rapide.

Cette fixation facilitée et ces incorporations erronées peuvent avoir une origine commune : le mauvais placement des résidus du *steric gate*. En effet, en situation normale, la phénylalanine de l'ADN polymérase λ (ou β) est en conformation active et fermée lors de la fixation du nucléotide dans le site actif. Mais lorsque la phénylalanine du *steric gate* est sous forme ouverte (comme ici pour le mutant λmutK), l'ADN polymérase est inactive. Ici, la situation est une sorte d'entre deux : la phénylalanine est sous forme ouverte, mais aucun résidu catalytique n'est détourné du site actif, donc la catalyse peut avoir lieu normalement. La mauvaise orientation de cette phénylalanine peut permettre à des nucléotides incorrects d'entrer dans le site catalytique de façon mal contrôlée, et les résidus catalytiques peuvent alors facilement incorporer ces nucléotides. Il semble donc que la simple mutation du motif RLDII en RLDIK non seulement ne permet pas à la polymérase de profiter d'un mécanisme de fidélité, mais empire la situation en permettant des incorporations erronées.

2.2.4.3.4 Structure de Polλ Ptet-like

Le dernier mutant étudié ici est supposé conférer à l'ADN polymérase λ humaine le mécanisme d'alternance de formes actives (fermée) et inactives (ouverte) des ADN polymérases X de *P. tetraurelia*, ou du moins la possibilité de former un pont salin entre les motif 2 et SD2 ou entre les deux motifs catalytiques. Sa structure a pu être obtenue en présence du nucléotide entrant correct et en présence d'un nucléotide incorrect.

2.2.4.3.4.1 En présence de dTTP, nucléotide correct

Là encore, la structure globale de ce mutant est très proche de celle de λmut lorsque le nucléotide entrant correct est présent, avec un RMSD de seulement 0,499 Å. La forme obtenue est donc une forme fermée.

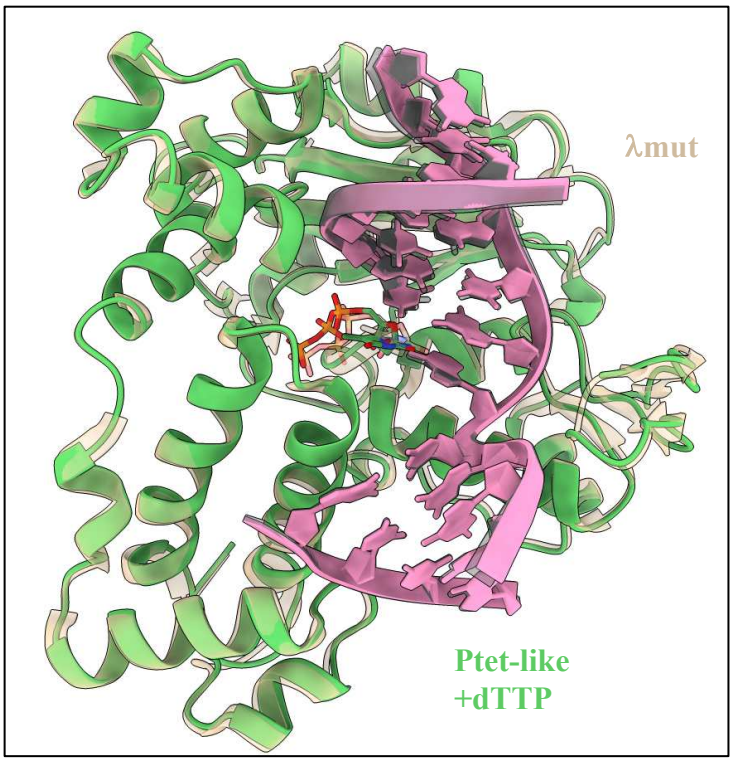


Figure 161 : Superposition des structures de la construction λmut (PDB 7M43, en beige, ADN en gris) et du mutant Ptet-like en présence de dTTP (en vert, ADN en rose).

Lors de l'affinement de cette structure, il est apparu que deux formes locales du site actif coexistent dans le cristal : pour l'une la lysine introduite est orientée vers le motif SD2 et les aspartates catalytiques sont libres pour réaliser une catalyse (en vert dans les figures 162 et 163); et dans l'autre la lysine est orientée vers un des aspartates catalytiques, qui est alors détourné du site actif (en rouge dans les figures 162 et 164).

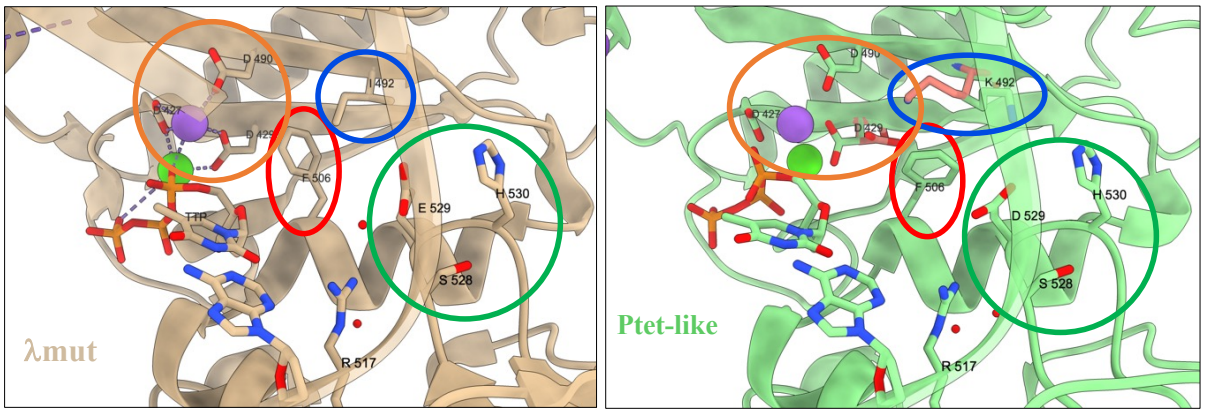


Figure 162 : Comparaison des résidus d'intérêt au sein du site actif de la construction λmut (à gauche, en beige) et du mutant Ptet-like en présence de dTTP (à droite, en vert ou rouge selon les formes possibles de certains résidus). Le nucléotide template est affiché, le nucléotide entrant est indiqué (TTP), et les résidus sont indiqués et entourés selon leur rôle : résidus catalytiques en orange ; résidu du steric gate en rouge, résidu d'intérêt du motif RxDx(Φ/+) en bleu, et motif SD2 en vert. Les ions placés dans le site catalytique sont indiqués en violet (Na) et vert (Ca).

Ces deux états locaux du site actif coexistent au sein d'une forme fermée du domaine catalytique. Dans la forme indiquée en vert, nommée forme active, les trois résidus aspartates catalytiques sont orientés vers le site catalytique et le nucléotide entrant. Leur position les rend donc disponibles pour la réaction enzymatique. La phénylalanine du *steric gate* est quant à elle encore dans un état proche de celui rencontré dans la structure de l'ADN polymérase λ (ou β) inactive. Enfin, la lysine 492 semble ici s'orienter vers le motif SD2, à 5 Å du groupement carboxylate de l'aspartate 529. Cette distance ne permet qu'une interaction électrostatique plus faible que dans le mutant λmutK, dans lequel le glutamate, plus long que l'aspartate, permettait la formation d'un pont salin. Cependant, il semble qu'ici, un réseau de molécules d'eau permet de lier la lysine à l'aspartate *via* sa chaine latérale et son groupement carboxyle sur la chaine principale.

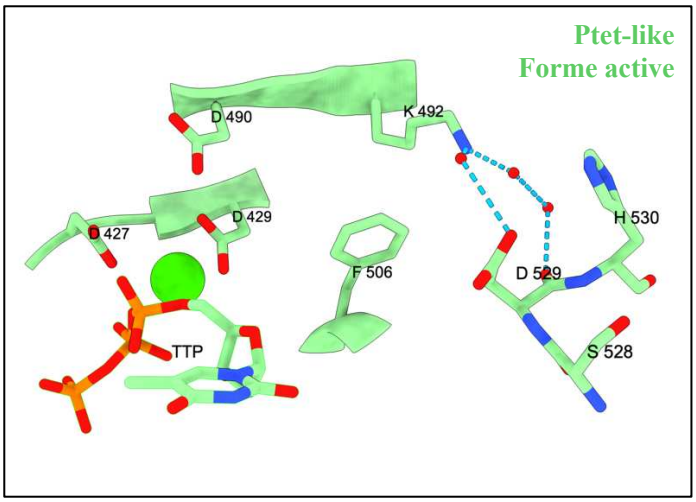


Figure 163 : État du site actif dans la forme « active » du mutant Ptet-like en présence de dTTP. Les résidus d'intérêt sont indiqués, ainsi que les molécules d'eau (en sphères rouge), l'atome de calcium placé dans le site actif (en sphère verte), et le nucléotide entrant (TTP). Les liaisons hydrogène sont indiquées par des traits bleus en pointillés.

Dans la seconde forme, indiquée en rouge et appelée inactive, le motif SD2 est placé de façon identique, tout comme les résidus F506, D490 et D427. Cependant, ici la lysine 492 et l'aspartate 429 sont orientés l'un vers l'autre, et leurs groupements amine et carboxylate sont à seulement 3 Å l'un de l'autre, une distance permettant la mise en place d'un pont salin.

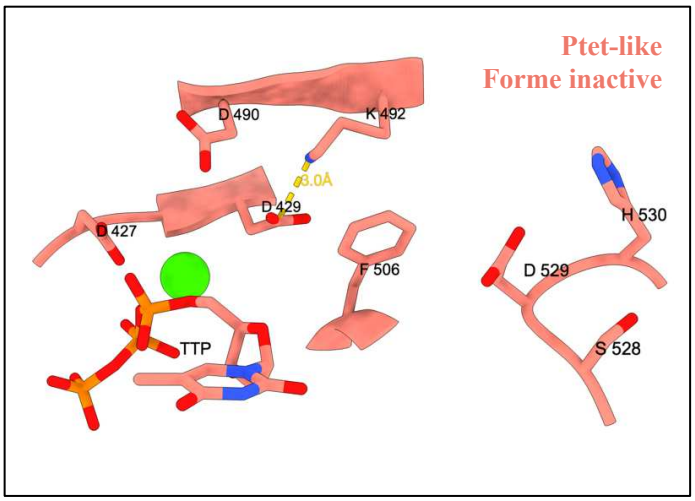


Figure 164 : État du site actif dans la forme « inactive » du mutant Ptet-like en présence de dTTP. Les résidus d'intérêt sont indiqués, ainsi que l'atome de calcium placé dans le site actif (en sphère verte), et le nucléotide entrant (TTP). La distance entre le groupement amine de la lysine 492 et le groupement carboxylate du glutamate 429 est indiquée en jaune (3 Å).

2.2.4.3.4.2 En présence de dCTP (nucléotide incorrect)

L'ajout d'un nucléotide incorrect ne semble pas altérer la structure globale de la protéine : l'écart moyen entre les carbones α homologues dans cette nouvelle structure par rapport à la structure de référence (PDB 7M43) est de seulement 0,422 Å. Un changement important serait pourtant attendu en présence d'un nucléotide incorrect si le passage vers une forme ouverte avait pu être induit par les mutations apportées, comme c'est le cas pour l'ADN polymérase β.

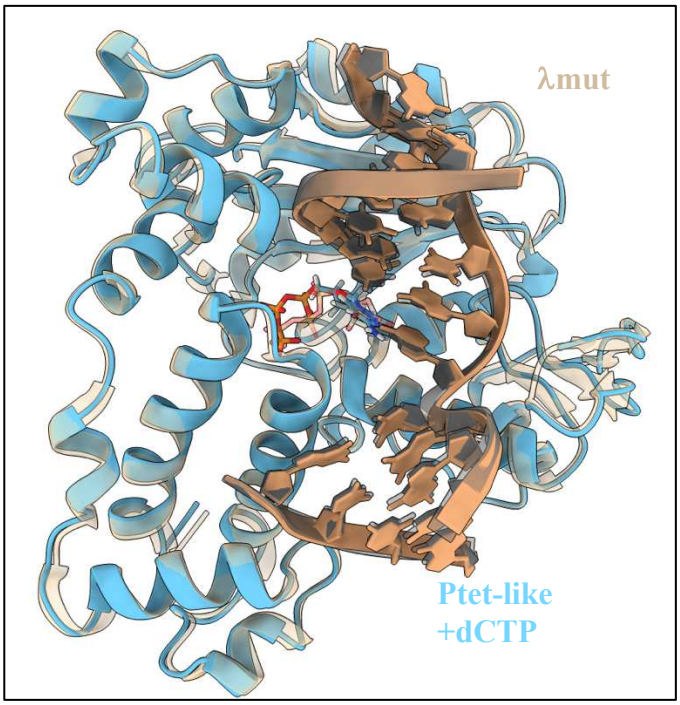


Figure 165 : Superposition des structures de la construction λmut (PDB 7M43, en beige) et du mutant Ptet-like en présence de dCTP (en bleu).

Cependant, on remarque dans cette structure que le mutant Ptet-like semble moins bien stabiliser l'ADN : cette seconde structure présente des *B-factors* plus élevés que dans la structure avec le nucléotide correct, en particulier au niveau de la boucle 3, de l'ADN, et même du domaine de 8 kDa impliqué dans la stabilisation de groupement 5'P sur l'amorce en aval du dommage. Le dCTP présent dans le site actif a, quant à lui, les *B-factors* les plus élevés de cette structure, ce qui indique qu'il est présent mais mal stabilisé au sein du site actif. Cela n'était observé que pour les phosphates du nucléotide entrant dans les autres structures (qui sont d'ailleurs indiqués sous 2 conformations alternatives dans la structure de référence 7M43).

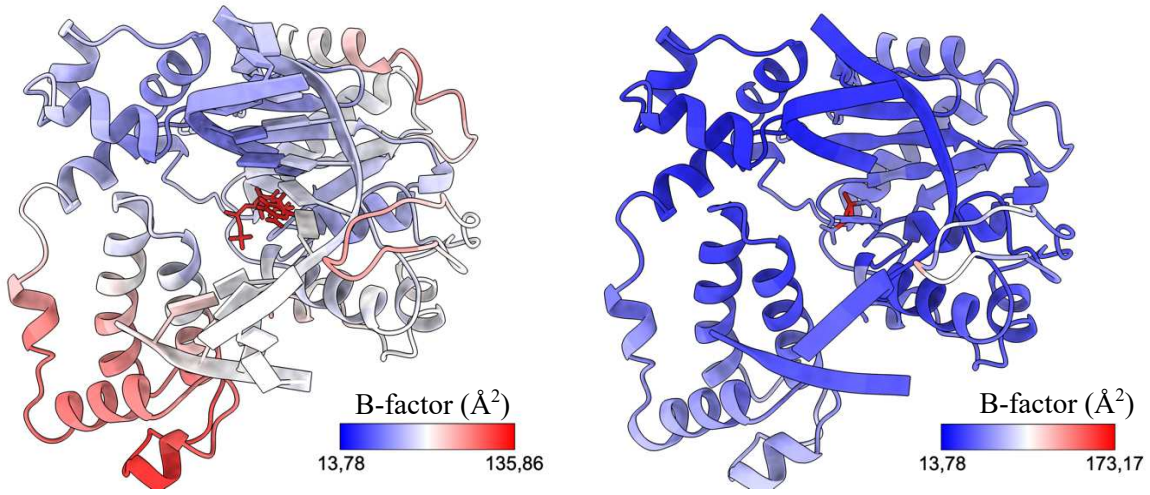


Figure 166 : Comparaison des deux structures obtenues pour le mutant Ptet-like en présence de dTTP (à droite) et de dCTP (à gauche). L'ADN et le nucléotide entrant sont indiqués. Les deux structures sont colorées selon leur B-factor. L'échelle est indiquée en bas à droite, en Å².

Comme dans la structure précédente, la phénylalanine du *steric gate* (F506) est en position « ouverte ». Le motif SD2 n'est pas lié à la lysine 492. Celle-ci est orientée vers l'aspartate 429, et son groupement amine est à seulement 2,7 Å du groupement carboxylate de l'aspartate, qui est donc détourné du site catalytique par un pont salin. Les deux autres aspartates catalytiques sont, quant à eux, dans une conformation normale. Concernant le nucléotide, qui est ici un dCTP, il n'est pas hybridé à l'adénine de l'ADN template. Il se place au fond du site actif, ce qui déplace les groupements phosphate à distance des aspartates catalytiques et ne permet plus l'entrée des ions catalytiques dans le site actif. Enfin, il n'est pas sur le même plan que la base template, mais presque 2 Å en aval.

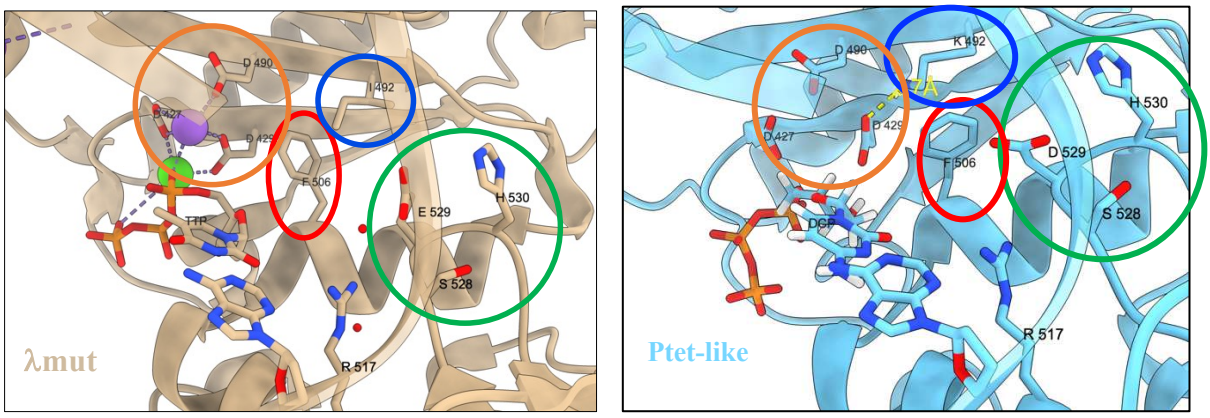


Figure 167 : Comparaison des résidus d'intérêt au sein du site actif de la construction λmut (à gauche, en beige) et du mutant Ptet-like en présence de dCTP (à droite, en bleu). Le nucléotide template est affiché, le nucléotide entrant est indiqué (DCP), et les résidus sont indiqués et entourés selon leur rôle : résidus catalytiques en orange ; résidu du steric gate en rouge, résidu d'intérêt du motif RxDx(Φ/+) en bleu, et motif SD2 en vert. Les ions placés dans le site catalytique sont indiqués en violet (Na) et vert (Ca). La distance entre le groupement amine de la lysine 492 et le groupement carboxylate du glutamate 429 est indiquée en jaune (2,7 Å).

L'ensemble des résultats structuraux obtenus pour le mutant Ptet-like peut être mis en lien avec les résultats obtenus lors de sa caractérisation cinétique. Celui-ci a montré en présence du nucléotide correct des caractéristiques cinétiques légèrement meilleures que λmut et le mutant λmutK concernant la vitesse de catalyse, et un K_m du même ordre que celui de λmutK, meilleur que pour λmut. Ce mutant semble donc plus rapide que les autres constructions, et avec une meilleure affinité pour le dGTP, ce qui traduit une meilleure efficacité catalytique, De plus, ce mutant présente une meilleure fidélité que le mutant λmutK. Cela suppose que la mutation du motif SD2 (SEH vers SDH) empêche l'incorporation de nucléotides erronés.

Dans le mutant λmutK, la lysine 492 partage une liaison électrostatique avec le glutamate du motif SD2, et ne peut donc pas changer de conformation pour empêcher la catalyse. Cependant, la phénylalanine du *steric gate* se positionne dans une conformation ouverte, qui peut limiter le contrôle du nucléotide entrant. Ce mutant semble donc être sous une forme active permanente mais modifiée, qui favorise les incorporations erronées. Le motif SD2 du mutant Ptet-like présente un résidu aspartate et non un glutamate. Il semble que pour ce mutant deux formes puissent exister : une forme « inactive » dans laquelle la lysine 492 est liée à un aspartate catalytique (forme rencontrée en présence d'un nucléotide incorrect) ; et une forme « active » dans laquelle les trois aspartates catalytiques sont disponibles pour catalyser la réaction enzymatique, et où la lysine 492 est liée à l'aspartate du motif SD2, de façon moins forte que dans le mutant λmutK puisque cette liaison se fait via un réseau de molécules d'eau. L'existence de ces différentes formes suggère qu'il y a un passage d'une forme inactive à une forme active en présence d'un dNTP correct, pour permettre son insertion (la structure obtenue

ici avec le dTTP est probablement un état intermédiaire, d'où l'existence des deux formes dans le cristal) ; mais que cette transition n'a pas lieu en présence d'un nucléotide incorrect, possiblement car le placement du nucléotide entrant ne permet pas d'atteindre la géométrie adéquate pour réaliser la catalyse. Ces transitions favorisent le bon placement du nucléotide correct au sein du site actif, ce qui se traduit par une augmentation de son affinité pour la polymérase (donc un K_m plus faible), et facilitent aussi la catalyse, ce qui se traduit par une réaction plus rapide (k_{obs} et V_{max} augmentés).

Ces résultats suggèrent une alternance entre deux formes active/inactive locales au sein du site catalytique mais sans mouvement global de domaines chez les ADN polymérases X de *Paramecium tetraurelia*, qui pourrait participer à leur grande fidélité en empêchant l'insertion de nucléotides incorrects.

2.2.5 Discussion

L'insertion d'une arginine en position 5 du motif RxDx(Φ/+) n'a pas suffi à conférer à l'ADN polymérase λ le mécanisme de transition d'une forme ouverte à une forme fermée, observé chez l'ADN polymérase β. Le résultat est le blocage d'un résidu aspartate catalytique par cette arginine, comme dans la forme inactive de l'ADN polymérase β mais sans mouvement de domaines. Cette mutation a des effets sur l'activité enzymatique, avec une réduction drastique de la vitesse de catalyse et une rétention du nucléotide entrant dans le site catalytique de la polymérase. L'ajout de la mutation du motif SD2 en NEY, qui confère l'ensemble des résidus impliqués dans ce mécanisme chez l'ADN polymérase β, ne permet pas au nucléotide d'entrer dans le site catalytique, tout en conservant la forme fermée de l'ADN polymérase λ (alors qu'en absence du nucléotide, l'ADN polymérase β est sous forme ouverte). Cela est probablement dû au mauvais placement de la boucle 3, qui ne peut pas stabiliser l'ADN instructeur au sein du site actif. Cet ADN étant mal placé, le nucléotide ne peut pas être stabilisé dans le site catalytique, et la catalyse ne peut pas avoir lieu. Cela se traduit par une perte quasi-totale d'activité enzymatique. Cela semble indiquer que le mécanisme de fidélité de l'ADN polymérase β n'est pas compatible avec l'ADN polymérase λ, qui a besoin du mouvement de sa boucle 3 pour être active.

L'insertion d'une lysine en position 5 du motif RxDx(Φ/+) entraine la formation d'un état actif permanent de l'ADN polymérase λ, sans contrôle des nucléotides entrants par le *steric gate* : cette lysine forme un pont salin avec le glutamate du motif SD2, et la phénylalanine du

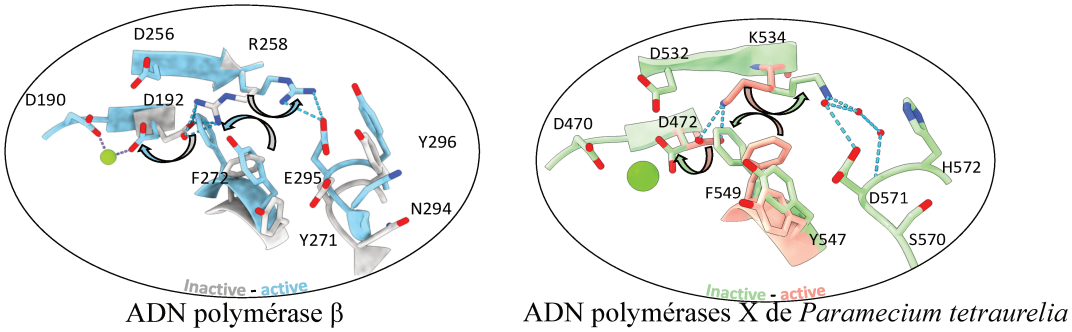
steric gate est déplacée. Cela se traduit par une fixation facilitée des nucléotides dans le site actif, une vitesse catalytique normale, mais surtout par des insertions erronées. La mutation additionnelle du motif SD2 de SEH vers SDH mime chez l'ADN polymérase λ la possibilité de former un pont salin comme chez les ADN polymérases X de *Paramecium*. Plusieurs états du site actif peuvent être observés, et indiquent qu'en présence d'un nucléotide incorrect, un résidu aspartate catalytique est détourné du site actif par la lysine ; alors qu'en présence du nucléotide correct, cette lysine se lie au motif SD2 via un réseau de molécules d'eau, et la catalyse peut avoir lieu normalement. Cela se traduit par une fixation facilitée du nucléotide entrant correct, et une catalyse légèrement plus rapide.

L'absence totale d'activité enzymatique du mutant Polβ-like suggère que les mécanismes de fidélité des ADN polymérases λ (impliquant la boucle 3, étudié ci-après en partie 3, à partir de la page 229) et β ne sont pas directement compatibles. Comme discuté précédemment, il est probable que les ADN polymérases X de *P. tetraurelia*, contrairement à l'ADN polymérase β, ne subissent pas les changements conformationnels alternant entre les formes ouverte et fermée de leur domaine catalytique. Le mutant Polβ-like a montré une incapacité à passer à un état ouvert, malgré l'absence de nucléotide rentrant dans le site actif, et le mutant Ptet-like a montré une conformation fermée en présence d'un nucléotide incorrect : ces observations diffèrent du résultat attendu pour un mécanisme similaire à celui de l'ADN polymérase β, qui est sous forme ouverte dans ces deux cas. Ici, c'est probablement dû à l'ADN polymérase λ utilisée pour réaliser ces mutations, qui n'est pas en mesure d'alterner des formes ouvertes et fermées de son domaine catalytique. Cependant, malgré l'absence de ces grands changements conformationnels, le mutant Ptet-like a montré une capacité à alterner les ponts salins K492-SD2 et K492-D429, selon le type de nucléotide entrant dans le site catalytique : si le nucléotide est correct, la lysine forme un pont salin avec le motif SD2 et la catalyse peut avoir lieu ; si le nucléotide est incorrect, la lysine se lie à l'aspartate catalytique 429, ce qui empêche la catalyse. Sans grand changement conformationnel, il semble donc que l'ensemble de résidus K492, D429 et SDH permette quand même une alternance de formes actives et inactives de façon locale selon le nucléotide entrant chez le mutant Ptet-like.

Ces résultats soulèvent de plus la question de l'absence des grands mouvements de domaines chez les ADN polymérases X de *P. tetraurelia*. Deux hypothèses pourraient expliquer cette absence. Tout d'abord, l'interaction entre la lysine du second motif catalytique et le motif SD2 (sous forme active) n'est peut-être pas assez forte pour provoquer un grand changement conformationnel. En effet chez l'ADN polymérase β cette interaction se fait entre une arginine

et le motif SD2, en particulier avec un glutamate et une tyrosine : l'interaction semble donc plus forte, et pourrait théoriquement provoquer un réarrangement global. La seconde explication implique la boucle 3 : comme cela est détaillé dans la partie suivante, elle a pour rôle de stabiliser l'ADN au sein du site catalytique. Cette interaction est indispensable pour avoir une activité enzymatique correcte, comme l'a montré le mutant Polβ-like, pour lequel la boucle 3 ne stabilise pas l'ADN, ce qui bloque l'activité enzymatique. Il est possible que le mouvement de cette boucle ne soit pas compatible avec l'existence d'une forme ouverte du domaine catalytique : sous forme ouverte, la boucle 3 serait trop éloignée de l'ADN pour le stabiliser au sein du site catalytique

En résumé, il semble possible que les ADN polymérases X de *Paramecium* puissent partager un mécanisme d'alternance de ponts salins avec l'ADN polymérase β, mais sans grand changement conformationnel, peut-être parce que le mouvement de la boucle 3 nécessaire à l'activité enzymatique l'empêche.



ADN polymérase β ADN polymérases X de *Paramecium tetraurelia*

Figure 168 : Mécanisme de fidélité de l'ADN polymérase β (à gauche) et des ADN polymérases X de Paramecium tetraurelia (à droite), basés sur des alternances de formes globales (ADN polymérase β) ou locales (ADN polymérases X de P. tetraurelia). En absence d'ADN et du nucléotide correct dans le site actif, les polymérases sont sous forme inactive. Sous cette forme, un des aspartates catalytiques est détourné du site actif par un résidu chargé positivement en position 5 du motif RxDx(Φ/+). Lorsqu'un nucléotide correct se place dans le site actif, avec une géométrie favorable à son incorporation dans l'ADN vis-à-vis du nucléotide template, l'ADN polymérase passe sous forme active. Chez l'ADN polymérase β, cette transition s'accompagne d'un mouvement global de fermeture du domaine catalytique. Au sein du site catalytique, plusieurs transitions conformationnelles ont lieu : les résidus du steric gate changent de place, et la phénylalanine rompt la liaison entre l'aspartate catalytique et le résidu chargé positivement. L'aspartate est alors libre pour s'engager dans la réaction enzymatique, et le résidu chargé positivement est stabilisé par une liaison avec les résidus du motif SD2. Chez l'ADN polymérase β cette interaction se fait directement, et chez les ADN polymérases X de Paramecium cette liaison est médiée par un réseau de molécules d'eau.

Des expériences devront cependant être réalisées pour tester cette hypothèse, par exemple en mutant les résidus impliqués dans ce mécanisme chez l'ADN polymérase Xa de *P. tetraurelia* : si la lysine 534 est impliquée dans ce mécanisme de fidélité, la muter en alanine ou en isoleucine (comme chez l'ADN polymérase λ) pourrait provoquer des erreurs d'incorporation, ou bloquer l'activité enzymatique. Sa mutation en arginine donnerait probablement le même résultat que le mutant λmutK. De la même façon, muter le glutamate du motif SD2 en alanine pourrait affecter ce mécanisme ainsi que la fidélité de l'enzyme et son activité.

2.3 L'implication de la boucle 3 dans la fidélité des ADN polymérases X de *Paramecium tetraurelia* et de l'ADN polymérase λ humaine

2.3.1 Introduction

Comme indiqué à la fin du chapitre 1 (page 127), deux hypothèses ont pu être émises pour expliquer la fidélité des ADN polymérases X de *Paramecium tetraurelia* : la première concerne le mécanisme d'activation du site catalytique par alternance entre deux formes (ouverte et fermée) et a été étudiée dans les expériences décrites précédemment ; et la seconde concerne la boucle 3. Cette boucle n'est présente que chez les ADN polymérases X proches des ADN polymérases λ, et jusqu'en 2022 son rôle n'était pas connu. Des travaux publiés en 2022 (Jamsen *et al.*, 2022) étudient la fidélité de l'ADN polymérase λ, et constatent un mouvement de la boucle 3 lors de la fixation de l'ADN et du nucléotide entrant dans le site actif. Ces travaux indiquent que cette boucle peut jouer un rôle important dans la catalyse réalisée par l'enzyme, et potentiellement dans sa fidélité.

L'objectif des travaux décrits dans cette partie a donc été de déterminer si la boucle 3 présente chez l'ADN polymérase λ et chez les ADN polymérases X de *Paramecium tetraurelia* a un impact sur la fidélité de ces enzymes. Pour cela, des constructions mutantes de ces enzymes ont été produites et purifiées, et leur fidélité a été testée. Enfin, les structures des mutants étudiés précédemment ont également permis d'étudier le fonctionnement de la stabilisation de l'ADN par cette boucle.

2.3.2 Expression et purification des constructions mutantes de l'ADN polymérase λ humaine

2.3.2.1 Construction PolXaΔBRCT-Loop3β

Cette construction mutante de l'ADN polymérase X a pour laquelle la boucle 3 a été remplacée par les résidus équivalents chez l'ADN polymérase β (Chapitre 2, 1.3.1.1, page 163) a été obtenue par mutagenèse dirigée du plasmide LS05- PolXaΔBRCT. Une fois le plasmide permettant l'expression du mutant obtenu et sa séquence confirmée, il a été possible de produire cette construction en système bactérien (*E.coli* BL21star(DE3)) par induction avec 1 mM d'IPTG et culture à 20°C.

Une purification a été réalisée à partir d'une production de 8L de culture bactérienne. La purification a suivi le protocole global indiqué en Matériel et Méthodes (les tampons utilisés sont indiqués en Annexe 1.6, page X). Après la lyse des bactéries par sonication et une centrifugation, la fraction soluble a été chargée sur une colonne His-Trap, et les protéines fixées ont été éluées avec un gradient d'imidazole.

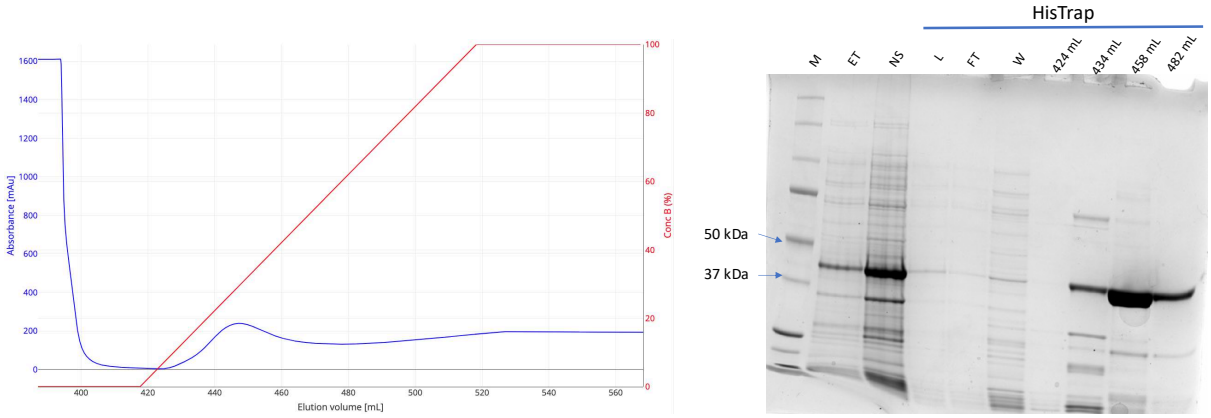


Figure 169 : Première étape (HisTrap) de purification de PolXaΔBRCT-Loop3β. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. À droite : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; ET : extrait total, obtenu après lyse bactérienne et avant séparation des fractions solubles et insolubles ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Une analyse SDS-PAGE a été réalisée avec les fractions issues de cette chromatographie, qui ont montré une bande correspondant à la masse attendue (45 kDa environ). Les fractions allant de 424 mL à 482 mL ont été rassemblées et diluées avant d'être chargées sur une colonne Héparine. L'élution a été réalisée avec un gradient de NaCl.

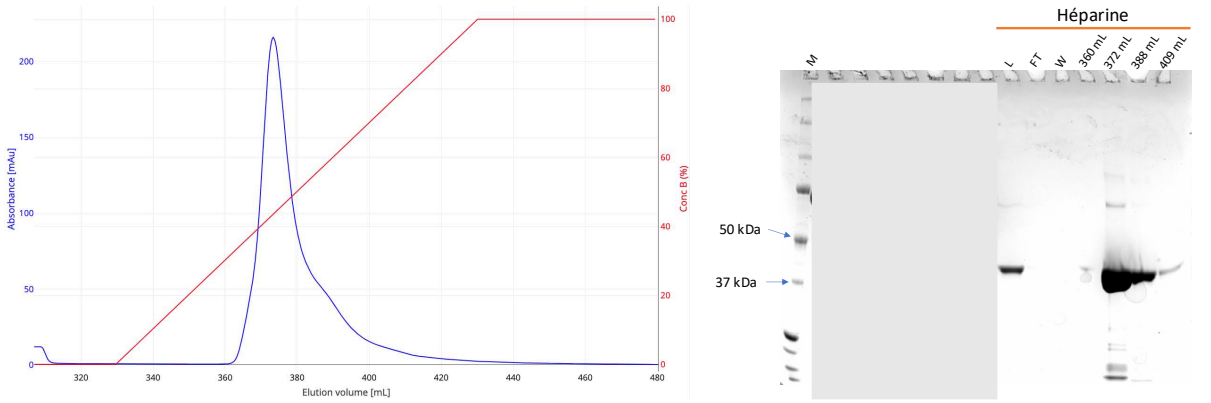


Figure 170 : Deuxième étape (Héparine) de purification de PolXaΔBRCT-loop3β. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. À droite : : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Une analyse SDS-PAGE a été réalisée, et a montré une bande majoritaire correspondant à environ 45 kDa. Les fractions de 360 mL à 409 mL ont été rassemblées, et l'ensemble avait une A_{280nm} de 0,242, avec un rapport A_{260nm}/A_{280nm} de 0,60. Une partie des fractions rassemblées a été concentrée, jusqu'à un volume final de 10 mL à une absorbance de 0,854 à 280 nm, soit 0,9 mg/mL. 10 mL ont été conservés à une concentration d'environ 0,25 mg/mL. Tous ces échantillons ont été aliquotés stockés à -20°C.

2.3.2.2 Construction λLoop3β

Cette construction mutante de l'ADN polymérase λ pour laquelle la boucle 3 a été remplacée par les résidus équivalents chez l'ADN polymérase β (détaillée en Chapitre 2,1.3.1.2, page 164) a été obtenue par mutagenèse dirigée du plasmide LS05- ADN polymérase λ. Une fois le plasmide permettant l'expression du mutant obtenu et sa séquence confirmée, il a été possible de produire cette construction en système bactérien (*E.coli* BL21star(DE3)) par induction avec 1 mM d'IPTG et culture à 20°C.

Une purification a été réalisée à partir d'une production de 2L de culture. La purification a suivi le protocole global indiqué en Matériel et Méthodes (les tampons utilisés sont indiqués en Annexe 1.6, page X). Après la lyse des bactéries par sonication et une centrifugation, la fraction soluble a été chargée sur une colonne His-Trap, et les protéines fixées ont été éluées avec un gradient d'imidazole.

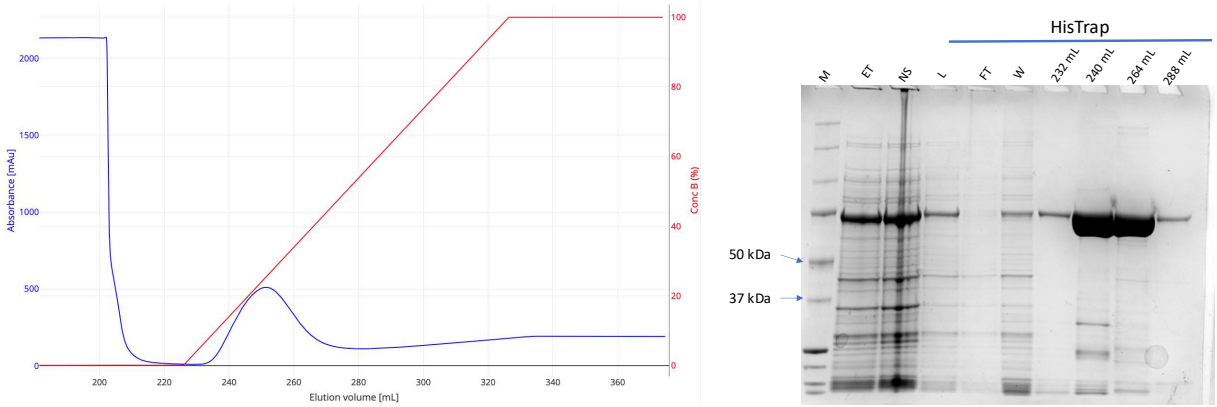


Figure 171 : Première étape (HisTrap) de purification la construction λLoop3β. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon B (en %) est indiquée en rouge. À droite : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; ET : extrait total, obtenu après lyse bactérienne et avant séparation des fractions solubles et insolubles ; NS : fraction non soluble ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Une analyse SDS-PAGE a été réalisée avec les fractions issues de cette chromatographie, qui ont montré une bande correspondant à la masse attendue (75 kDa environ). Les fractions allant de 232 mL à 288 mL ont été rassemblées et diluées avant d'être chargées sur une colonne Héparine. L'élution a été réalisée avec un gradient de NaCl.

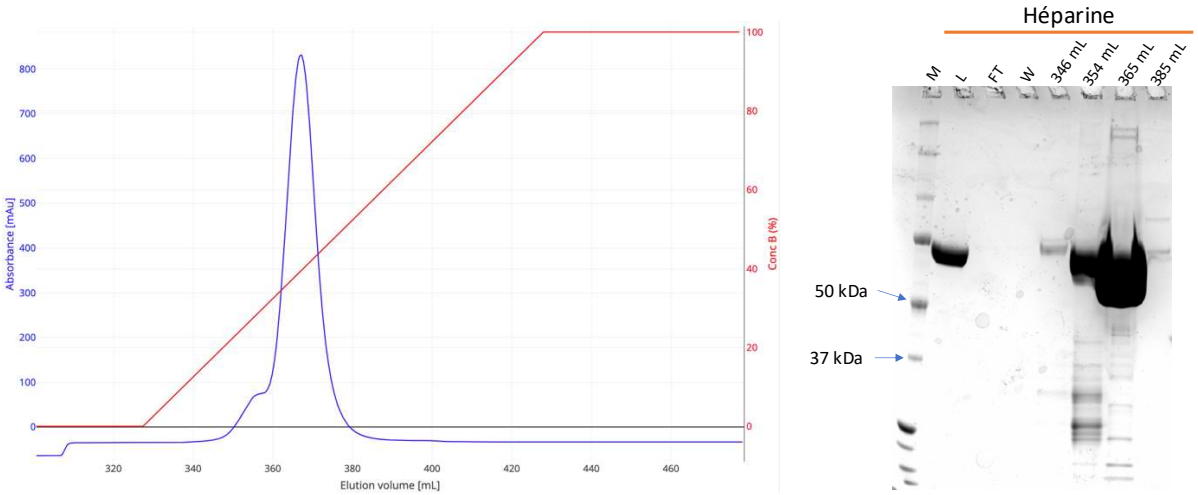


Figure 172 : Deuxième étape (Héparine) de purification de la construction λLoop3β. À gauche : Chromatogramme (centré sur l'étape d'élution). L'absorbance à 280 nm (en mUA) est indiquée en bleu ; la concentration du tampon E (en %) est indiquée en rouge. À droite : : SDS-PAGE. M : Marqueur de poids moléculaire (les bandes correspondant aux masses moléculaires de 50 et 37 kDa sont indiquées ; L : load, fraction soluble chargée sur la résine; FT : Flow-Through, protéines non fixées sur la résine ; W : wash, fraction correspondant au lavage de la colonne ; Les volumes indiquent des fractions choisies d'après le chromatogramme de l'étape associée.

Une analyse SDS-PAGE a été réalisée, et a montré une bande majoritaire correspondant à environ 75 kDa. Les fractions de 364 mL à 385 mL ont été rassemblées, et l'ensemble avait une A_{280nm} de 1,163, avec un rapport A_{260nm}/A_{280nm} de 0,66. Les fractions rassemblées ont été concentrées, jusqu'à un volume final de 9,5 mL à une absorbance de 3,998 à 280 nm, soit 4,54 mg/mL, puis aliquotées et stockées à -20°C.

2.3.3 Comparaison de la fidélité d'une ADN polymérase X de *P. tetraurelia* et de l'ADN polymérase λ avec et sans leurs boucles 3

Après avoir produit et purifié les deux constructions ne portant plus de boucle 3, leur fidélité a été testée, de la même façon que pour les ADN polymérases X a et d de *P. tetraurelia* (Chapitre 2, 2.1.5, page 194). Elles ont été comparées à leurs homologues portant la boucle 3.

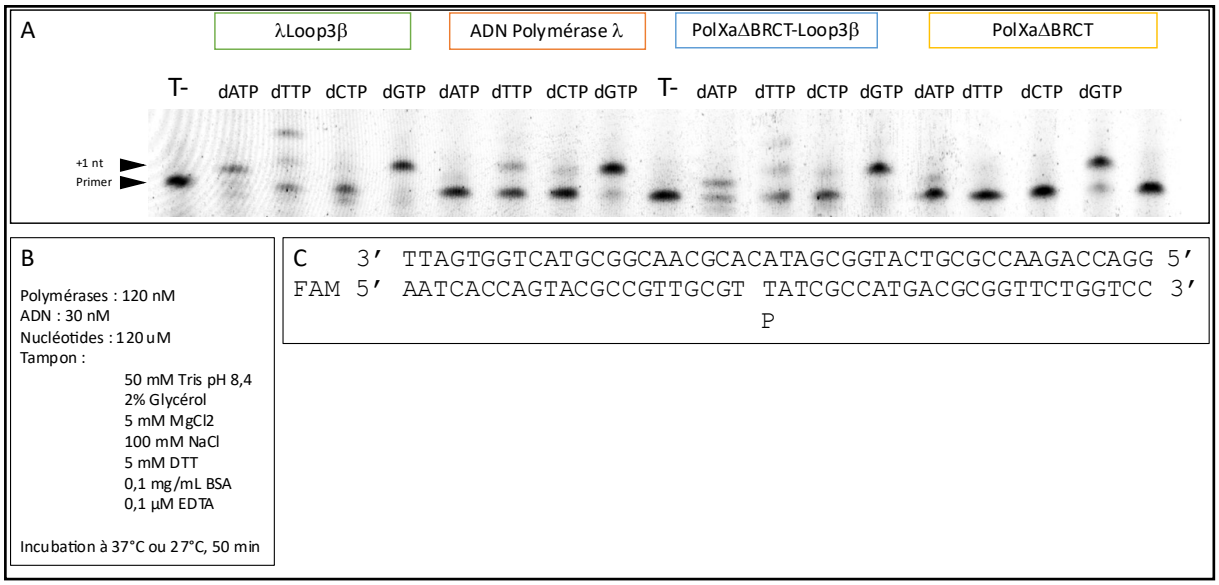


Figure 173 : Test d'incorporation de nucléotides incorrects par des constructions de l'ADN polymérase λ humaine et de l'ADN polymérase Xa de Paramecium tetraurelia portant ou non leur boucle 3 – Gap-filling. A : Résultat du test. Les ADN polymérases testées sont indiquées, ainsi que les nucléotides testés. (T- : témoin négatif, 20 nt). B : conditions expérimentales. C : Duplex d'ADN testé. Le marquage FAM en 5' de l'amorce est indiqué.

Les ADN polymérases testées portant leur boucle 3 montrent des résultats similaires à ceux déjà connus : l'ADN polymérase X a de *P. tetraurelia* ne fait pas d'incorporations erronées, mais étend la quasi-totalité de l'ADN présent dans le test en présence de dGTP. L'ADN polymérase λ humaine semble faire quelques incorporations erronées en présence de dTTP et de dCTP, et incorpore très efficacement le dGTP sur l'ADN. Les constructions ne portant pas de boucle 3 semblent en revanche faire plus d'erreurs d'incorporation. La construction λLoop3β incorpore du dATP et du dTTP au lieu du dGTP, et on observe même deux incorporations en présence de dTTP, probablement parce que la base template après le gap est un dA. La construction PolXaΔBRCT-Loop3β incorpore elle aussi du dATP, du dTTP (deux nucléotides incorporés) et du dCTP. Les deux constructions sans boucle 3 incorporent normalement le dGTP. La délétion de la boucle 3 chez ces ADN polymérases semble donc réduire leur fidélité.

2.3.4 Apport de l'étude structurale des mutants de l'ADN polymérase λ visant à lui conférer le mécanisme d'activation du site catalytique

Dans les structures obtenues et présentées précédemment, pour le mutant « Polβ-like » la boucle 3 est éloignée de l'ADN et présente une flexibilité accrue, contrairement aux autres structures où elle interagit avec l'ADN. Dans ces configurations, la boucle 3 s'engage avec l'ADN par le biais de liaisons impliquant ses résidus chargés positivement (K544, R538, H541), ainsi que le groupe carboxyle de G542, qui peut établir des liaisons électrostatiques avec un

groupement phosphate *via* K521. Sans ces interactions, l'ADN semble être déplacé, et le nucléotide instructeur n'est pas correctement positionné pour la catalyse, ce qui conduit vraisemblablement à une incapacité à stabiliser un nucléotide entrant.

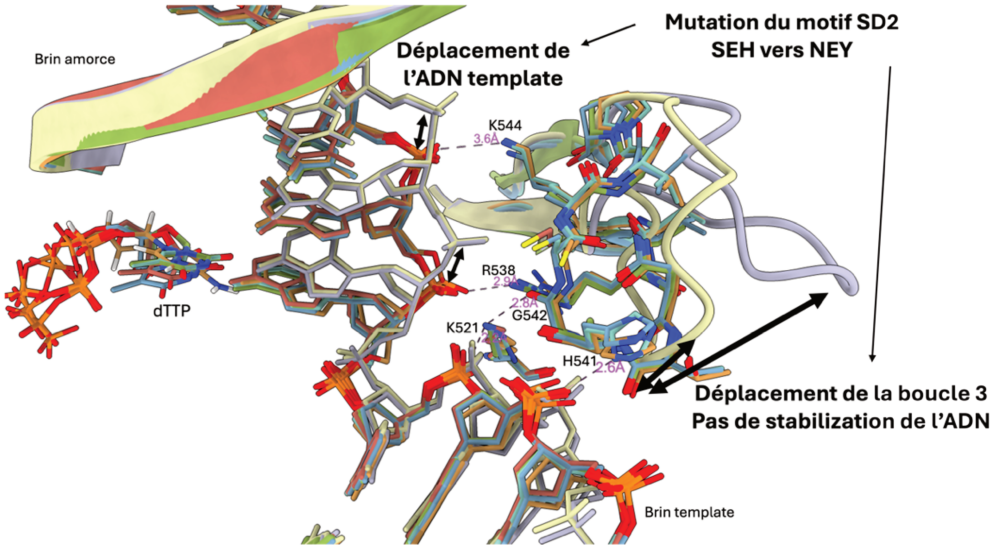


Figure 174 : Comparaison de l'interaction de la boucle 3 avec le brin d'ADN modèle dans les six structures obtenues. La construction λmut (PDB 7M43) est représentée en vert, λmutR en bleu clair, λmutK en rouge, le mutant Ptet-like (avec dTTP) en bleu, la version avec dCTP en orange, et les deux structures obtenues du mutant Polβ-like en violet et jaune. Les boucles 3 déplacées sont indiquées en « cartoons », tandis que pour les autres, la représentation est atomique. Le déplacement de la boucle 3 et de l'ADN modèle est indiqué avec des flèches en gras. Les distances des liaisons possibles entre la boucle 3 et l'ADN modèle sont indiquées en violet.

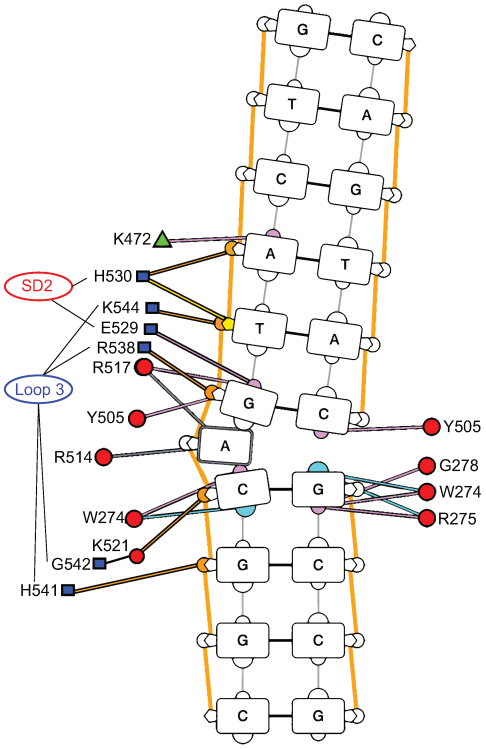


Figure 175 : Interactions de l'ADN polymérase λ avec l'ADN dans les structures des mutants actifs (λmut, λmutR, λmutK et Ptet-like). Tous les résidus impliqués dans l'interaction avec l'ADN sont indiqués. Les résidus indiqués par des cercles rouges sont situés dans des hélices, ceux par des triangles verts se trouvent dans des brins β, et ceux par des carrés bleus sont dans des boucles. Les résidus de la boucle 3 et du motif SD2 sont indiqués. L'interaction des résidus avec les nucléotides est indiquée en fonction de leur couleur : les interactions en cyan sont situées dans le sillon mineur, celles en rose dans le sillon majeur, celles en jaune sont des interactions impliquant le désoxyribose, les interactions avec les bases sont en gris, et les interactions avec les groupements phosphate sont indiquées en orange.

La seule différence entre ce mutant et la construction λmutR, dans laquelle la boucle 3 peut stabiliser l'ADN, réside dans le motif SD2, un NEY (comme pour l'ADN polymérase β) et un SEH pour la construction λmutR. Dans toutes les constructions actives, le résidu H530 du motif SD2 (présent uniquement dans les constructions λmut, λmutR, λmutK, et *Ptet-like*) interagit avec les nucléotides +2 et +3 sur le brin template. Cette interaction est perturbée dans le mutant Polβ-like en raison de la substitution de H530 par une tyrosine, et c'est probablement la raison pour laquelle la boucle 3 ne stabilise pas l'ADN dans cette structure.

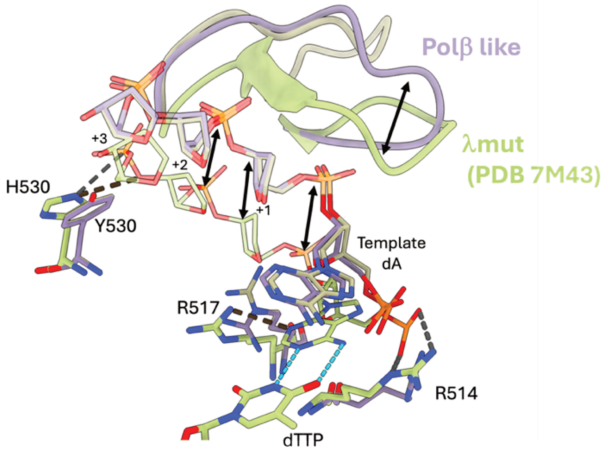


Figure 176 : Comparaison des interactions de l'ADN template dans le site actif entre la construction λmut (PDB 7M43) active et le mutant Polβ-like inactif. Les interactions des résidus avec le nucléotide dA instructeur et les groupements phosphate sont indiquées en gris. Les liaisons hydrogènes entre le dA et le nucléotide entrant sont en cyan. Le déplacement de la boucle 3 et de l'ADN modèle entre les deux structures est indiqué avec des flèches en gras.

Il est probable que dans la construction λmut (et les autres présentant l'interaction boucle 3 - ADN), l'interaction H530-ADN précède celle de la boucle 3 avec l'ADN modèle : en présence d'un nucléotide correct, l'ADN bénéficie de l'hybridation avec le dNTP entrant (stabilisé dans le site actif) et d'autres interactions (comme celles avec Y505, R514, et R517) pour obtenir un positionnement adéquat mais instable. Ce placement facilite l'interaction de H530 avec l'ADN, initiant ainsi des interactions en cascade de la boucle 3 avec l'ADN afin d'assurer son positionnement optimal pour la catalyse. Le premier résidu à interagir avec l'ADN est la lysine 544 qui interagit avec le nucléotide +2 (comme H530), suivie des autres résidus chargés positivement de la boucle 3. En l'absence d'un dNTP correct ou en présence d'un dNTP incorrect, l'ADN ne peut pas être stabilisé dans le site actif, empêchant ainsi l'interaction de H530 et de la boucle 3. Pour la construction Polβ-like, l'interaction du résidu 530 avec l'ADN est impossible, donc la boucle 3 ne peut pas aider à stabiliser l'ADN, qui reste mal positionné dans le site actif même en présence du nucléotide entrant correct (qui ne peut pas non plus être stabilisé, probablement en raison de l'instabilité de l'ADN). Cependant, le mutant Ptet-like semble se comporter différemment, car même en présence d'un nucléotide incorrect, l'ADN et

la boucle 3 sont correctement positionnés. Cette différence peut être attribuée à son motif SD2 (SDH), qui facilite potentiellement l'interaction de H530 avec l'ADN, stabilisant ce dernier même en présence d'un nucléotide incorrect.

2.3.5 Discussion

Ces résultats indiquent un rôle de la boucle 3 dans la fidélité de ces ADN polymérases. D'après les travaux publiés en 2022 (Jamsen *et al.*, 2022), lors de la fixation d'un nucléotide entrant correct, la boucle 3 est proche du brin matrice de l'ADN et le stabilise dans une conformation optimale pour la catalyse. En absence d'un nucléotide entrant correct, cette boucle est à distance de l'ADN, qui est alors mal stabilisé. Les comparaisons des structures des constructions mutantes de l'ADN polymérase λ suggèrent que le mécanisme par lequel la boucle 3 stabilise l'ADN template dépend de l'interaction de l'ADN avec H530, le troisième résidu du motif SD2. L'interaction de H530 avec l'ADN semble être facilitée par la stabilisation du nucléotide entrant correct et par son hybridation avec le nucléotide modèle dans le site actif. Il a été démontré que l'hybridation du dNTP entrant déclenche le repositionnement de l'ADN modèle (Bebenek *et al.*, 2008). Cela suggère un mécanisme selon lequel, lors de l'entrée d'un nucléotide entrant correct dans le site actif, il est stabilisé et s'hybride avec le nucléotide modèle, qui est également stabilisé par R514 et R517 (Jamsen *et al.*, 2022). Cela repositionne l'ADN et le rapproche de H530, permettant l'interaction entre les deux, et déclenchant une interaction en cascade des résidus de la boucle 3 avec l'ADN. Lorsque la boucle 3 interagit avec l'ADN, elle achève sa stabilisation dans le domaine catalytique, facilitant ainsi l'incorporation du nucléotide correct. Des études antérieures (Bebenek *et al.*, 2008; Foley *et al.*, 2010) ont montré que lorsque R517 est muté, l'ADN ne se repositionne pas, pas plus que la boucle 3. Cela est probablement dû à la participation de R517 dans la pré-stabilisation de l'ADN et de la paire de base en formation. Étant donnée la conservation de la boucle 3 et sa forte charge positive chez les ADN polymérases X de *P. tetraurelia* (pI de 10,3 ; 8,23 chez l'ADN polymérase λ humaine), il est probable que son interaction avec l'ADN soit conservée et encore plus forte.

Toutes ces observations montrent l'importance de cette boucle dans la fidélité des ADN polymérases λ et des autres ADN polymérases X ayant une boucle équivalente, comme celles de *Paramecium tetraurelia*.

Enfin, la comparaison des mouvements de la boucle 3 avec les mouvements de domaines de l'ADN polymérase β semble indiquer que ces deux mécanismes donnent le même résultat : ils permettent de mieux stabiliser l'ADN instructeur dans le site catalytique de l'enzyme. Chez l'ADN polymérase β, la fixation d'un nucléotide entrant correct entraine la rupture du pont salin entre l'arginine du motif RIDIR et un aspartate catalytique. L'arginine forme ensuite un pont salin avec le motif SD2, ce qui provoque le mouvement du domaine pouce, et la fermeture du domaine catalytique. Cette fermeture permet l'interaction des résidus R283 et K280 avec la paire de base en formation, et sa stabilisation. Le fonctionnement de l'ADN polymérase λ est différent : c'est la fixation du nucléotide entrant dans le site actif qui permet avec les résidus R514 et R517 de pré stabiliser la paire de base en formation et l'ADN instructeur. Cette interaction est permise par la forme fermée du domaine catalytique de l'ADN polymérase λ (sous forme ouverte, les résidus équivalent de l'ADN polymérase β ne sont pas à proximité de l'ADN). Une fois l'ADN pré stabilisé, il est rapproché du motif SD2, donc H530 peut interagir avec lui, et provoquer la cascade d'interactions de la boucle 3 avec l'ADN. Dans les deux cas, l'ADN est stabilisé en présence d'un nucléotide correct, mais les deux mécanismes sont différents : seule une boucle change de position chez l'ADN polymérase λ, alors que tout le domaine pouce est déplacé chez l'ADN polymérase β.

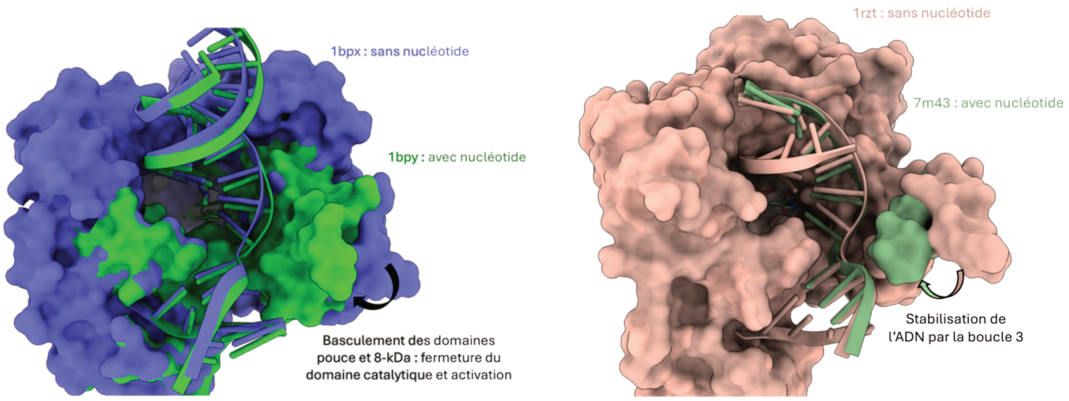


Figure 177 : Comparaison des mécanismes de stabilisation de l'ADN des ADN polymérases β et λ. Pour l'ADN polymérase β (à gauche), ce mécanisme repose sur le mouvement du domaine pouce lors de la fixation d'un nucléotide entrant (passage de la forme bleue à la forme verte). Chez l'ADN polymérase λ (à droite), ce mécanisme repose sur le déplacement de la boucle 3, qui stabilise l'ADN au sein du site catalytique lors de la fixation d'un nucléotide correct (passage de la forme rose à la forme verte).

Il semble donc que l'ADN polymérase λ (et les ADN polymérases portant une boucle 3) aient évolué pour porter un mécanisme de fidélité similaire à celui de l'ADN polymérase β. D'après les hypothèses évolutives actuelles (Bienstock *et al.* 2015), les ADN polymérases λ et β seraient descendantes des ADN polymérases bactériennes. Celles-ci portent les résidus impliqués dans le mécanisme *induced fit* de l'ADN polymérase β, donc elles utilisent

possiblement le même mécanisme. Il semble donc que les ADN polymérases λ (et les ADN polymérases X proches) aient divergé jusqu'à perdre ce mécanisme, adopter une forme fermée permanente du domaine catalytique, et qu'un autre mécanisme soit apparu, impliquant la boucle 3.

Des expériences complémentaires seront nécessaires pour comprendre pleinement ce mécanisme de fidélité. Pour cela, l'étape suivante sera l'obtention de la structure du mutant λLoop3β. En absence de la boucle 3, il est possible en théorie que l'ADN ne parvienne pas à se stabiliser correctement au sein du site catalytique, et que dans la structure sa position soit incorrecte, voire qu'il soit absent. D'autres expériences pourraient apporter davantage d'informations : l'importance de l'histidine du motif SD2 pourrait être prouvée, en la mutant en alanine. Dans ce cas, on pourrait s'attendre d'un point de vue structural à ce que la boucle 3 ne s'approche pas de l'ADN pour le stabiliser ; et d'un point de vue enzymatique, cela devrait donc affecter la fidélité de l'enzyme.

Conclusions et perspectives

Les réarrangements programmés du génome de *Paramecium tetraurelia* font intervenir de nombreux mécanismes, dont une forme de NHEJ spécialisée dans la réparation des cassures doubles brins introduites de façon programmée par PiggyMac. Ce mécanisme de réparation spécialisé est très fidèle, et il est assuré *via* l'introduction des cassures doubles brins par PiggyMac, qui est couplée à leur réparation puisque PiggyMac doit être lié (entre autres) avec l'hétérodimère Ku70a/80c pour introduire des CDB. Ce mécanisme de réparation fait également intervenir des ADN polymérases de la famille X. Les enzymes de cette famille sont généralement considérées comme des ADN polymérases mutagènes, mais chez *P. tetraurelia* la réparation des CDB est réalisée sans erreurs. Mon objectif a été de caractériser ces ADN polymérases spécialisées, afin de comprendre les origines de leur fidélité.

Grâce à une approche bio-informatique utilisant la méthode CLANS, j'ai pu classifier 7250 ADN polymérases X de différents organismes, dont *P. tetraurelia*, et ces résultats m'ont permis d'actualiser la classification actuelle de cette famille de polymérases. Plusieurs groupes d'ADN polymérases X étaient jusqu'ici considérés comme une seule sous famille : par exemple les ADN polymérases X de plantes et certaines ADN polymérases X de champignons ont été intégrées parmi les ADN polymérases λ, d'après leurs similitudes avec l'ADN polymérase λ humaine. Mes travaux ont permis de mieux comprendre les différences entre ces enzymes, en se focalisant sur les éléments conservés dans leurs séquences. J'ai ainsi pu définir 12 groupes d'ADN polymérases de la famille X ainsi que les éléments permettant de les reconnaitre.

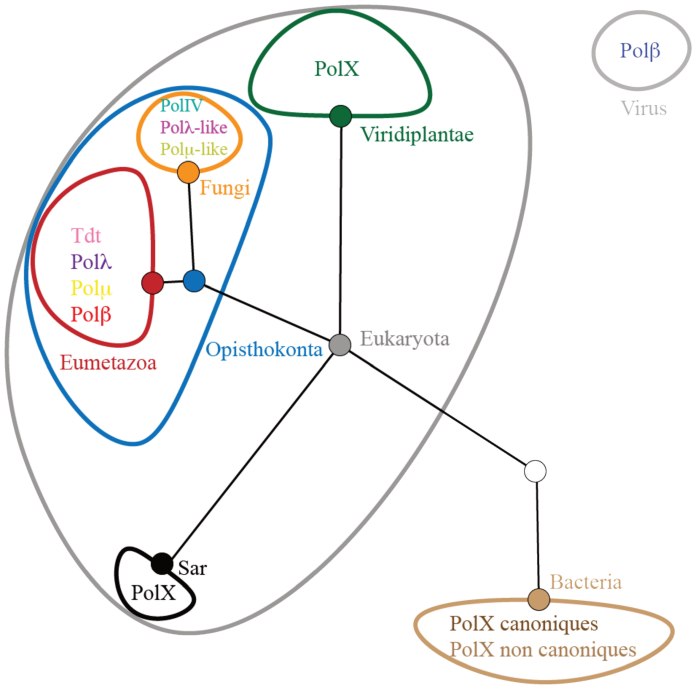


Figure 178 : Vue d'ensemble des 12 groupes de d'ADN polymérases X, séparées selon leur origine au sein du vivant

Cela m'a permis de comparer les ADN polymérases X de *P. tetraurelia* aux autres groupes, en particulier les ADN polymérases β et λ de métazoaires. Des alignements de séquences réalisés avec des ADN polymérases X déjà caractérisées ont permis d'émettre plusieurs hypothèses pour expliquer la fidélité de ces enzymes :

- Un ensemble de résidus localisés dans des motifs connus (DxD, RxDx(Φ/+), *steric gate* et SD2) est similaire à ceux impliqués dans le mécanisme d'ouverture/fermeture di domaine catalytique permettant à l'ADN polymérase β d'être hautement fidèle, suggérant que la fidélité des ADN polymérases X de *P. tetraurelia* et β de métazoaire repose possiblement sur un mécanisme équivalent.
- Les ADN polymérases X partagent avec l'ADN polymérase λ une boucle C-terminale, possiblement impliquée dans un mécanisme de fidélité chez l'ADN polymérase λ. Son rôle exact n'est pas connu, mais si cette boucle est liée à la fidélité de l'ADN polymérase λ, les ADN polymérases X de *Paramecium* pourraient bénéficier du même mécanisme.

Après avoir produit et purifié quatre constructions d'ADN polymérases de la famille X, j'ai pu les caractériser d'un point de vue enzymatique. Cela m'a permis de confirmer leur similarité avec l'ADN polymérase λ, qui partage avec elles plusieurs activités ainsi que des caractéristiques cinétiques proches. J'ai enfin pu montrer que les ADN polymérases X de *Paramecium tetraurelia* sont plus fidèles que l'ADN polymérase λ humaine, et que leur fidélité ne repose que sur leur domaine catalytique et non sur des éléments inclus dans le linker N-terminal reliant ce domaine au domaine BRCT. Il n'a pas été possible d'obtenir la structure de ces ADN polymérases directement par cristallographie, malgré un grand nombre d'essais.

J'ai donc employé une approche indirecte pour tester la première hypothèse sur les bases de la fidélité de ces enzymes, en produisant des constructions mutées de l'ADN polymérase λ humaine. La caractérisation structurale et fonctionnelle de ces mutants a permis de montrer qu'une alternance de formes actives et inactives du site catalytique peut exister chez les ADN polymérases X de *Paramecium tetraurelia*, mais sans grand changement conformationnel par mouvement de domaines, comme chez l'ADN polymérase β. Ce mécanisme peut permettre d'**améliorer la fidélité de ces enzymes**, mais semble moins extensif que celui de l'ADN polymérase β, puisque celle-ci alterne des conformations ouvertes et fermées du domaine

catalytique, ce qui n'est probablement pas le cas chez *Paramecium tetraurelia* au vu des structures cristallographiques présentées ici.

Enfin, j'ai pu produire, purifier et tester des constructions de l'ADN polymérases λ et d'une ADN polymérase X de *P. tetraurelia* ne portant pas de boucle 3. Cela m'a permis de montrer que **la boucle 3 est impliquée dans la fidélité** de ces deux enzymes. L'étude des structures obtenues précédemment a également permis de mieux comprendre ce mécanisme, reposant probablement sur une stabilisation de l'ADN dans le site actif par l'hybridation du nucléotide entrant avec la base instructrice et par la stabilisation de cette conformation par des résidus du domaine « pouce » (R514 et R517), puis sur une interaction du résidu H530 avec l'ADN permettant à la boucle 3 d'interagir à son tour avec l'ADN. Par ailleurs, au vu de son placement vis-à-vis de l'ADN template (au contact des deux nucléotides du brin template en aval du dommage), il est possible que cette boucle joue un rôle prépondérant dans la stabilisation de l'ADN lors de la réparation des CDB induites par PiggyMac chez *P. tetraurelia* : elle pourrait participer à maintenir la micro-homologie TA, avec l'aide du domaine de 8 kDa.

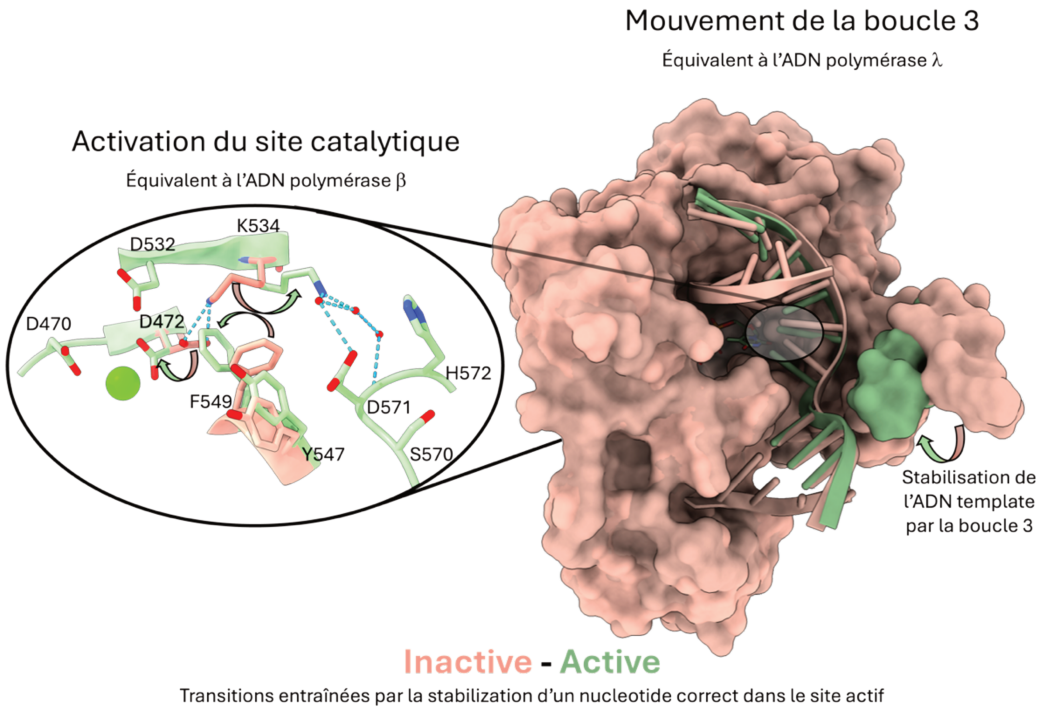


Figure 179 : ***Deux mécanismes de fidélité sont présents dans les ADN polymérases X de Paramecium tetraurelia et sont similaires aux mécanismes de fidélité des ADN polymérases β et λ.*** *À gauche : le mécanisme d'activation du site catalytique. Lors de l'entrée d'un nucléotide entrant correct dans le site actif, F549 change de conformation et rompt le pont salin entre D472 et K534. D472 peut alors participer à la catalyse, et K534 change de conformation pour interagir avec le résidu D571 du motif SD2, grâce à des liaisons hydrogènes médiées par des molécules d'eau. À droite : le mouvement de la boucle 3 permettant le positionnement optimal de l'ADN modèle dans le site actif. L'entrée du nucléotide entrant correct dans le site actif stabilise le nucléotide template par l'intermédiaire de R514 et R517, ce qui permet à H530 de stabiliser le brin d'ADN modèle. Cela déclenche l'interaction de la boucle 3 avec ce brin d'ADN grâce à de multiples ponts salins, stabilisant ainsi l'ADN dans le site actif dans une position optimale pour la catalyse. Modèle basé sur les résultats obtenus avec les constructions mutantes de l'ADN polymérase λ.*

Il est intéressant de noter que ce mouvement de la boucle 3 joue un rôle similaire au mouvement de fermeture du domaine catalytique observé chez l'ADN polymérase β : dans ces deux cas, le résultat est la stabilisation de l'ADN instructeur au sein du domaine catalytique. Ces deux mécanismes semblent avoir des origines différentes, mais les deux dépendent du motif SD2 : chez l'ADN polymérase β, c'est la liaison de ce motif avec l'arginine du second motif catalytique qui provoque un déplacement du domaine pouce et la fermeture du domaine catalytique ; chez l'ADN polymérase λ, le motif SD2 commence la cascade d'interactions entre la boucle 3 et l'ADN, après une pré-stabilisation de ce dernier par d'autres résidus, permise par la forme fermée permanente du domaine catalytique.

De façon surprenante, d'autres ADN polymérases de la famille X partagent les résidus à l'origine de ces deux mécanismes de fidélité : les ADN polymérases X des plantes présentent l'équivalent de l'arginine 258 de l'ADN polymérase β sous forme de lysine, ainsi qu'une boucle 3, à l'instar de l'ADN polymérase λ. Leurs motifs *steric gate* (AWTGN) et SD2 (DDT) s'éloignent des séquences connues, mais en gardent les caractéristiques physico-chimiques, et ces enzymes pourraient donc présenter des mécanismes similaires. Certaines ADN polymérases fongiques "λ/μ-like" semblent également partager des équivalents de la boucle 3, mais avec des motifs SD2 plus divergents. Cela met en lumière l'évolution de ces mécanismes dans cette famille d'ADN polymérases, qui devra être étudiée plus en détail. Les ADN polymérases X présentant des mécanismes de fidélité similaires à ceux des ADN polymérases β et λ pourraient constituer un chaînon manquant dans l'évolution de cette famille d'ADN polymérases, entre les ADN polymérases β et λ.

Un autre élément d'intérêt chez ces ADN polymérases X est le linker reliant le domaine BRCT N-terminal au domaine polymérase, qui est particulièrement long et conservé (ce qui suggère une importance fonctionnelle). Son rôle est inconnu, mais il pourrait être lié à l'organisation de la machinerie NHEJ spécialisée de *Paramecium tetraurelia*, pour former ou stabiliser des complexes avec les autres protéines impliquées (Ku 70/80) ou l'ADN.

Enfin, l'interaction de ces ADN polymérases avec leurs partenaires protéiques devra être étudiée. La caractérisation biophysique de ces interactions pourrait être réalisée par ultracentrifugation analytique, et permettrait de mieux comprendre la formation des complexes de réparation des CDB programmées, ainsi que la spécialisation des quatre ADN polymérases X de *Paramecium tetraurelia*.

Chapitre additionnel

Reclassification of family A DNA polymerases reveals novel functional subfamilies and distinctive structural features

Dariusz Czernecki, Antonin Nourisson, Pierre Legrand et Marc Delarue

Nucleic Acids Research

Publié le 18 avril 2023

1 Introduction

1.1 Les ADN polymérases de la famille A

Historiquement, la première famille d'ADN polymérases à avoir été étudiée est la famille A (Lehman *et al.*, 1958). Les ADN polymérases de cette famille sont constituées *a minima* de deux domaines : un domaine ADN polymérase et un domaine 3'-5' exonucléase parfois inactivé (Aliotta *et al.*, 1996). Ce dernier est dédié au contrôle et à la correction du brin d'ADN synthétisé (Ollis *et al.*, 1985), ce qui assure la fidélité de sa réplication. Certains membres de cette famille possèdent également un domaine 5'-3' exonucléase impliqué dans la dégradation des amorces ribonucléiques présentes dans les fragments d'Okazaki (Fukushima *et al.*, 2007).

Le membre le plus connu de cette famille est l'ADN polymérase I d'*Escherichia coli*, en particulier sous sa forme dérivée, le fragment de Klenow, dont le domaine 5'-3' exonucléase a été supprimé (Klenow and Henningsen, 1970). Celui-ci a donné son nom au pli structural de ces polymérases.

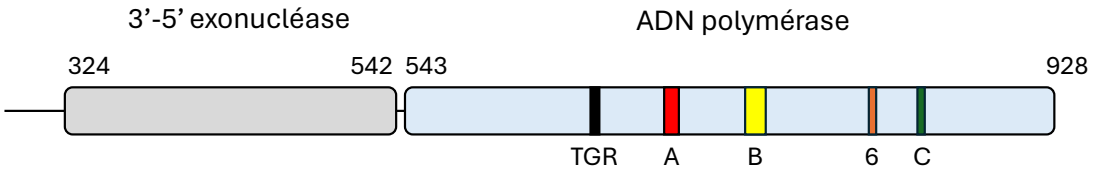


Figure 180 : Schéma global des ADN polymérases de la famille A, basé sur le fragment de Klenow (ADN polymérase I de E.coli). Les motifs impliqués dans l'activité de l'enzyme sont indiqués.

La famille A regroupe des ADN polymérases très variées : certaines interagissent avec des thiorédoxines pour être actives (Hinkle and Richardson, 1975), et d'autres sont fusionnées à d'autres protéines ou domaines (Lindner *et al.*, 2011; Newman *et al.*, 2015; Stewart *et al.*, 2009). Certaines de ces ADN polymérases, grâce à leur thermostabilité (Abramson, 1995), sont fréquemment utilisées dans des approches de biologie moléculaire (Moser *et al.*, 2012; Saiki *et al.*, 1988).

Au sein des cellules, les ADN polymérases de la famille A, connues pour leurs rôles dans la réplication, présentent également une fonction de réparation de l'ADN (Hernández-Tamayo *et al.*, 2019; Makiela-Dzbenska *et al.*, 2009; Okazaki *et al.*, 1971; Uphoff *et al.*, 2013). L'ADN polymérase θ joue un rôle dans la réparation des CDB (Black *et al.*, 2019). Cette ADN polymérase porte un domaine polymérase, un domaine 3'-5' exonucléase inactif, ainsi qu'un

domaine supplémentaire en position N terminale, homologue aux hélicases, et une région centrale prédite comme désordonnée. Parmi ses nombreuses activités, l'ADN polymérase θ peut réparer des cassures doubles brins en présence de grandes microhomologies au sein du système MMEJ, ou ajouter un dA en face d'un site abasique (Black *et al.*, 2019).

1.2 Les ions impliqués dans l'activité des ADN polymérases

Il est admis que l'activité des ADN polymérases dépend d'un mécanisme à deux ions, tel qu'il a été défini par Thomas Steitz en 1991 (Steitz, 1999). D'après ce modèle, deux ions divalents (généralement des ions Mg^{2+}) sont stabilisés par des groupements carboxylates, situés sur des aspartates ou glutamates du site actif. Le premier ion (l'ion A) interagit avec le groupement 3'-OH du brin amorce et permet l'attaque sur le phosphate α du nucléotide triphosphate entrant. Les deux ions stabilisent l'état de transition de la réaction, et l'ion B se fixe aux phosphates β et γ du nucléotide ajouté à l'amorce, et facilite leur sortie du site actif. Il a été proposé qu'un troisième ion pourrait être impliqué dans la catalyse par certaines ADN polymérases. D'après l'article de Yang Gao et Wei Yang (Gao and Yang, 2016), ce troisième ion serait nécessaire pour rompre la liaison entre les phosphates α et β du nucléotide entrant, et pour protoner le pyrophosphate produit par la réaction.

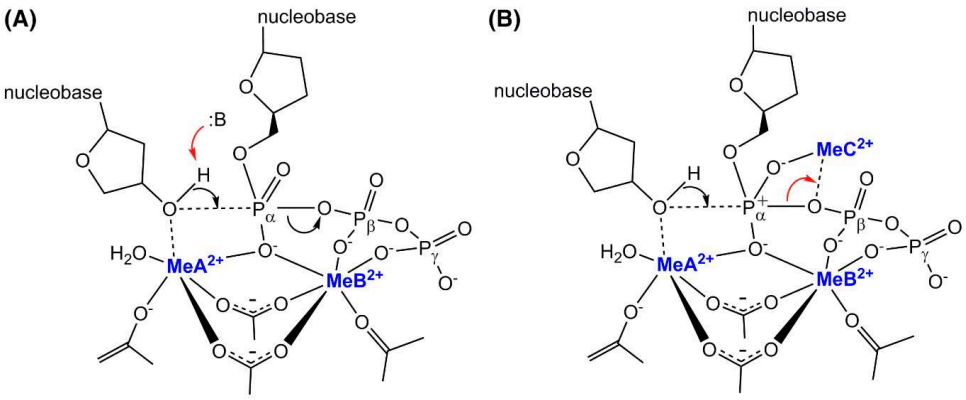


Figure 181 : Mécanisme à deux ions (A) ou trois ions (B) des ADN polymérases. D'après Tsai, M.-D Protein Sci. (2019). (Tsai, 2019)

Cependant, cette hypothèse a été réfutée récemment par Jimin Wang et William Konigsberg, qui affirment plutôt que la présence de ce métal C n'est possible qu'à de hautes concentrations, inhibitrices pour l'activité de l'enzyme. Selon eux, la présence de ce troisième ion pourrait retenir le pyrophosphate normalement libéré du site actif, ce qui permettrait une réaction de pyrophosphorolyse (Wang and Konigsberg, 2022).

Les ions divalents pouvant être utilisés par les ADN polymérases sont variés (Vashishtha *et al.*, 2016), mais la plupart de ces enzymes peuvent être actives en présence de Mg2+, de Mn2+ ou de Co^{2+}. Cependant, les ions utilisés peuvent modifier l'activité de l'ADN polymérase, par exemple en modifiant l'affinité du site actif pour les nucléotides entrants (Vashishtha and Konigsberg, 2016) ou en modifiant la sélectivité de la polymérase pour les nucléotides (Goodman *et al.*, 1983; Miyaki *et al.*, 1977). Selon leur spécialisation, certaines ADN polymérases ont cependant besoin d'ions spécifiques, comme les ADN polymérases spécialisées dans la réparation NHEJ, dans la réparation par excision de base ou par synthèse translésionnelle (Balint and Unk, 2020; Blanca *et al.*, 2003; Frank and Woodgate, 2007; Garcia-Diaz *et al.*, 2007; Gouge *et al.*, 2013; Kirby *et al.*, 2012; Kuznetsova *et al.*, 2022; Martin *et al.*, 2013; Park *et al.*, 2022; Wang *et al.*, 1977) qui sont plus actives dans certains contextes en présence d'ions Mn^{2+}.

1.3 La reclassification des ADN polymérases de la famille A

Malgré l'ancienneté de leur découverte, la classification des ADN polymérases de la famille A est soit incomplète car concentrée sur des sous familles (Chan *et al.*, 2011; Schoenfeld *et al.*, 2013), soit datée (Filée *et al.*, 2002). De plus, les techniques de classifications utilisées jusqu'ici reposaient sur des alignements de séquences, qui perdent en intérêt face à la divergence des grandes quantités de données à analyser (Nuin *et al.*, 2006; Pervez *et al.*, 2014). Une nouvelle méthode de classification appelée CLANS (pour CLuster ANalysis of Sequences) a été utilisée récemment pour étudier la famille des ADN polymérases B et la superfamille AEP (Archaeo-Eukaryotic Primases). Cette classification a permis de découvrir de nouvelles sous familles (Kazlauskas *et al.*, 2020, 2018). Cette méthode (Frickey and Lupas, 2004) est basée sur un algorithme (Fruchterman and Reingold, 1991) qui permet de projeter des séquences dans un espace tridimensionnel puis de les regrouper selon leur similarité. Pour déterminer la distance entre deux séquences, le programme utilise les scores de similarité produits par l'algorithme BLAST (Altschul *et al.*, 1990). C'est également la méthode utilisée dans la première partie de cette thèse.

Récemment, au sein du laboratoire, la méthode CLANS a été utilisée pour analyser une banque de données de séquences d'ADN polymérases A. Les sous familles déjà connues ont été identifiées, et de nouveaux groupes non caractérisés ont été découverts. Les structures de représentants de ces groupes ont été prédites en utilisant Alphafold2 (Jumper *et al.*, 2021), et

l’analyse de ces prédictions a permis de découvrir des particularités des ADN polymérases de ces nouveaux groupes.

1.4 Contribution

Ma contribution à ces travaux a été la production et la purification de deux ADN polymérases issues des nouveaux clusters d’ADN polymérases A, et leur caractérisation afin d’étudier leur activité. Ces expériences ont permis de proposer des spécialisations de ces deux ADN polymérases, basées sur leurs activités 3’-5’ exonucléase et ADN polymérase. Cette activité a été testée en présence de différents ions catalytiques, et c’est ce qui a permis de définir l’activité physiologique de réplication ou de réparation de ces ADN polymérases.

4488–4507 *Nucleic Acids Research, 2023, Vol. 51, No. 9* *Published online 18 April 2023*
https://doi.org/10.1093/nar/gkad242

Reclassification of family A DNA polymerases reveals novel functional subfamilies and distinctive structural features

Dariusz Czernecki[1,2,*], Antonin Nourisson[1,2], Pierre Legrand[1,3] and Marc Delarue[1,*]

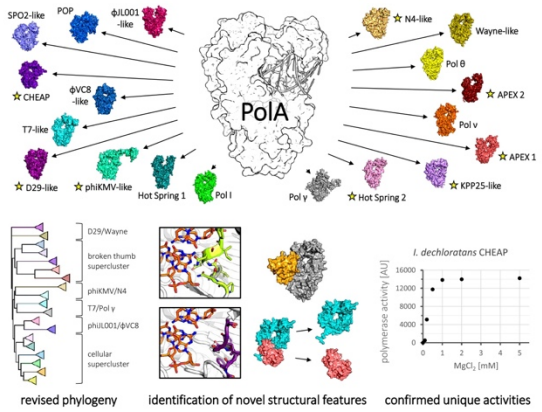
[1]Institut Pasteur, Université Paris Cité, CNRS UMR 3528, Unit of Architecture and Dynamics of Biological Macromolecules, 75015 Paris, France, [2]Sorbonne Université, Collège Doctoral, ED 515, 75005 Paris, France and [3]Synchrotron SOLEIL, L'Orme des Merisiers, 91190 Saint-Aubin, France

Received August 02, 2022; Revised March 07, 2023; Editorial Decision March 20, 2023; Accepted March 24, 2023

ABSTRACT

Family A DNA polymerases (PolAs) form an important and well-studied class of extant polymerases participating in DNA replication and repair. Nonetheless, despite the characterization of multiple subfamilies in independent, dedicated works, their comprehensive classification thus far is missing. We therefore re-examine all presently available PolA sequences, converting their pairwise similarities into positions in Euclidean space, separating them into 19 major clusters. While 11 of them correspond to known subfamilies, eight had not been characterized before. For every group, we compile their general characteristics, examine their phylogenetic relationships and perform conservation analysis in the essential sequence motifs. While most subfamilies are linked to a particular domain of life (including phages), one subfamily appears in Bacteria, Archaea and Eukaryota. We also show that two new bacterial subfamilies contain functional enzymes. We use AlphaFold2 to generate high-confidence prediction models for all clusters lacking an experimentally determined structure. We identify new, conserved features involving structural alterations, ordered insertions and an apparent structural incorporation of a uracil-DNA glycosylase (UDG) domain. Finally, genetic and structural analyses of a subset of T7-like phages indicate a splitting of the 3′–5′ exo and pol domains into two separate genes, observed in PolAs for the first time.

GRAPHICAL ABSTRACT



INTRODUCTION

In the course of evolution, nucleic acids emerged as the universal information carriers of life. The 'central dogma' of molecular biology describes how this information flows from DNA to RNA (and back), and from RNA to proteins (1). Despite having an auto-replicative potential that may have played a role during the very origins of life (2), contemporary nucleic acids are efficiently replicated by protein enzymes in a condensation reaction of nucleotide triphosphates, exploiting Watson–Crick base pairing with the templating strand as the fundamental mechanism for replication fidelity. DNA is the dominant support of genetic information found in all cellular organisms, as well as in an important fraction of the virus world, and its synthesis *in cellulo* relies on a variety of DNA polymerases (3).

There are eight distinct families of DNA polymerases (DNAPs or Pols): A, B, C, D, X, Y, PrimPol (AEP

*To whom correspondence should be addressed. Tel: +44 1223 267597; Email: dczernecki@mrc-lmb.cam.ac.uk
Correspondence may also be addressed to Marc Delarue. Tel: +33 1 45 68 86 05; Email: marc.delarue@pasteur.fr
Present address: Dariusz Czernecki, Medical Research Council Laboratory of Molecular Biology, Francis Crick Avenue, Cambridge, UK.

superfamily) and reverse transcriptases (4,5). While the first seven are DNA dependent, the last one uses RNA as a template; with the exception of PrimPol and several members of family B (6,7), all families need RNA or DNA primers to initiate replication. Some of the eight families are specialized in processive replication, whereas others are involved in re-priming this replication at blocked replication forks or in the repair of DNA damage (4,7). Family A polymerases (PolAs)—the first DNA polymerase family to be discovered and biochemically characterized (3)—essentially consist of a polypeptide chain folded into a polymerase (pol) domain and a proofreading 3′–5′ exonuclease (exo) domain (8); the additional 5′–3′ exonuclease domain is dispensable and does not interfere with the polymerase function and fidelity. While in cellular organisms PolA enzymes perform roles pertaining to whole-genome replication, recombination or repair (9–11), they also efficiently duplicate DNA of eukaryotic organelles (mitochondria and plastids), bacterial plasmids (12) and bacteriophages (e.g. T3/T7, SPO1 and SPO2) (13–15). Due to their relative simplicity, they are routinely used in diverse DNA amplification techniques (16,17), for instance in polymerase chain reaction (PCR)-based diagnostics against the severe acute respiratory syndrome-coronavirus 2 (SARS-CoV-2), which recently caused a major global health crisis (18). The archetypal member of the PolA family is *Escherichia coli* Pol I DNA polymerase and the derived so-called Klenow fragment, which lacks the 5′–3′ exonuclease domain.

Intriguingly, up to now PolAs represent the sole known family of DNA polymerases with representatives that moved away from the strict classical Watson–Crick base paring scheme. These polymerases (called DpoZ) operate in ZTGC-DNA phages, and have evolved a shifted specificity towards 2-aminoadenine (Z) instead of adenine at the templating thymine's Watson–Crick edge (19,20). Other phages, such as YerA41, developed PolA variants that are able to accept hypermodified thymine nucleotides in a templating strand during the replication of the phage DNA, which cannot be processed by commercially available polymerases (21). However, phages with similar modifications on 5-hydroxymethylcytosine, such as glucosylation (phages T-even: T2, T4 and T6) or arabinosylation (phage RB69), possess dedicated PolBs instead of PolA (22,23).

At the structural level, all PolAs share the Klenow fold, first identified in the X-ray structure of the Klenow fragment of *E. coli* DNA polymerase I (8). DNA polymerases B, Y, PrimPol and reverse transcriptases, as well as phage-like single-subunit RNA-dependent RNA polymerases, also possess the Klenow fold and are evolutionarily interconnected (24). On the other hand, families X and C share a different nucleotidyltransferase fold, the Polβ fold, while PolIDs have yet another fold most commonly found in multisubunit RNA polymerases (the so-called double-psi β-barrel fold) (4).

The PolA family has been explored in the past using various phylogenetic analyses. Nevertheless, such classifications are either outdated (25), or focus only on a particular subfamily (26,27) or a particular group of biological entities (28–30). These methods rely heavily on multiple sequence alignments (MSAs), which are challenged by the divergent nature of very large datasets of sequences and vary in accuracy depending on the algorithm used, demanding specialized solutions for different cases (31,32). Lastly, the resulting phylogenetic trees assume *a priori* a fixed (and common) mutation rate—a synchronized molecular clock—for all PolA sequences; yet, a significant departure from this hypothesis is expected there, due to the divergent molecular specializations or horizontal transfers between carrier species having various life cycles, among other factors (33).

A non-hierarchical sequence clustering based on the Fruchterman–Reingold algorithm (34) avoids the above issues altogether. It aims not to provide exact phylogenetic relationships between the particulars, but rather to project and regroup them in Euclidean space based on their pairwise similarities. In short, the algorithm first randomly distributes the sequences in three-dimensional (3D) space, and then updates the position of each sequence by using pseudo-forces derived from the resemblance of each pair, until convergence is achieved. It was implemented in a computer program CLANS (35), which employs the BLAST algorithm (36) to assign the similarity scores to each sequence pair. CLANS has recently been used to investigate family B DNA polymerases (PolBs) and superfamily AEP, complementing the phylogenetic analyses and leading to the discovery of new subfamilies (37,38).

Inspired by these findings, we applied this clustering analysis to PolAs, using an up to date and fully comprehensive library of sequences. We confirmed its accuracy by correctly delineating known PolA subfamilies, unifying them in one global distribution that still captures their reported relationships. We distinguished five previously uncharacterized major groups, and three minor ones showing high similarity to other subfamilies. For each new PolA cluster, we determined its composition, occurrence and phylogenetic connections with other subfamilies. We used AlphaFold2 (39), the most recent and powerful protein structure prediction program (40), to investigate the architecture of the eight new and four still structurally undescribed PolA subfamilies. Consistent, high-confidence structural predictions revealed novel structural features, in the form of ordered insertions, domain assimilation or exo/pol gene splitting. Additionally, we determined that representatives of a known hot spring-associated subfamily consistently appear in archaea as well, demonstrating the presence of these PolAs across all domains of life. Finally, we tested the catalytic activity of the enzymes from two previously unexplored bacterial groups. Both act as a templated DNA polymerase, as expected, yet display distinct polymerase and exonuclease activity levels as well as different divalent metal ion dependencies.

MATERIALS AND METHODS

PolA protein sequence acquisition and clustering

All 60 975 sequences tagged as DNA polymerase A (HMMER-defined PFAM ID: PF00476 (41)) in the UniProt database (42) were downloaded in October 2021 and filtered, removing fragmentary (shorter than 350 amino acids) or incomplete (containing residue X) sequences. Using the BLAST algorithm (version 2.2.26) (36), the remaining sequences were compared against each other to assess their likeness; sequences with identity >70% were

further removed from the dataset. The final selection (8109 sequences) was manually supplemented with several described PolAs of interest, sequences with a known 3D structure as well as all available DpoZ sequences (φVC8-like and Wayne-like) with pairwise identity <90%, due to their current under-representation.

The FASTA file containing 8136 final sequences was processed by the CLANS web-utility from the MPI Bioinformatics Toolkit (43). The clustering simulation was conducted using the Java version of CLANS (35) with default parameters. The sequences were randomly distributed in 3D space, converging to individual clusters after several hundred steps of the simulation. The simulation was let to run for a total of 6000 steps, during which no further modification of the positions appeared. The simulation was run without applying a *P*-value cut-off: additional simulations with cut-offs of 10^{-10} and 10^{-20} resulted, respectively, in an identical cluster distribution but with a reduction in size for most clusters, or further shrinking and fragmentation of the clusters. Clusters were determined with the network-based clustering tool, with a minimum of 40 sequences per cluster and active offset. Clusters φVC8-like and Wayne-like were selected manually, although they can be detected automatically using a lower threshold of at least 20 sequences per cluster. Several, independent simulations converged to almost identical cluster distributions, with no discrepancy regarding the critical details. Simulations not enriched with additional φVC8-like/Wayne-like DpoZ sequences resulted in equivalent distributions, barring the smaller size of clusters #18 and #19.

Similarities with PolAs of interest outside of the dataset were routinely assessed with BLAST searches at the NCBI (44).

Sequence analysis, structure prediction and phylogeny

Protein sequences making up each cluster were extracted with CLANS. For every subfamily, the sequences were aligned with Clustal Omega (45) applying default parameters. The alignments were used to generate sequence logos with WebLogo (46), or were displayed directly with ESPript3 (47).

The existing PolA structures were found through UniProt (42); completeness of the dataset was confirmed with Dali (48) queries. A local version of ColabFold (49) running 18 iterations of the AlphaFold2 algorithm (39) was used to predict PolA 3D models for all sequences selected for phylogenetic analysis (five per cluster) without a crystallographic structure. The highest ranking models of each run were superposed and analysed: they converged towards similar conformations—especially close within the clusters—and obtained high predicted local distance difference test (pLDDT) confidence scores, generally in the [70, 98] range. Known and predicted structures were visualized with Pymol (50).

Five representatives of each cluster were selected to create an MSA that comprises all clusters/subfamilies. Due to misalignments of additional, unrelated domains, all 95 sequences were truncated to their Klenow-like large fragments, i.e. the core PolA fold (3′–5′ exo and pol domains), based on PDB and AlphaFold2 structural models. The MSA was constructed on the truncated sequences in Clustal Omega with default parameters. The program also calculated a Neighbour–Joining tree without distance corrections, that was visualized with iTOL (51) on an unrooted dendrogram. The MSA was additionally used as an input for bootstrap analysis with MEGA X (52) (Maximum Likelihood method, 100 replicates, JTT model, uniform substitution rates).

Purification of *Streptomyces* sp. CT34 APEX and *I. dechloratans* CHEAP

The synthetic genes of *Streptomyces* CT34 (Actinobacterial Polymerases with a potentially Eclipsed eXonuclease or APEX subfamily) and *Ideonella dechloratans* (Cellular Highly Efficient Auxiliary Polymerases or CHEAP subfamily) were optimized for expression in *E. coli* (Supplementary Table S1) and synthesized using ThermoFisher's GeneArt service. The genes were cloned into a modified pRSF1-Duet expression vector with an N-terminal 14-histidine tag using New England Biolabs and Anza (Thermo Fisher Scientific) restriction enzymes. *Escherichia coli* BL21 Star (DE3) cells (Invitrogen) were transformed with the engineered plasmids. Bacteria were cultivated at 37°C in LB medium with kanamycin resistance selection and induced at an optical density (OD) = 0.6–1.0 with 0.5 mM isopropyl-β-D-thiogalactopyranoside (IPTG). After incubation overnight at 20°C, cells were harvested and homogenized in suspension buffer: 50 mM HEPES pH 8.0 (APEX) or 50 mM Tris pH 9.0 (CHEAP), 500 mM NaCl, 10 mM imidazole. After sonication and centrifugation of bacterial debris, corresponding lysate supernatants were supplemented with Benzonase (Sigma-Aldrich) and protease inhibitors (Thermo Fisher Scientific), 1 μl and one tablet per 50 ml, respectively. The proteins of interest were isolated by purification of the lysates on a HisTrap column (suspension buffer as washing buffer, 500 mM imidazole in elution buffer). Collected proteins were diluted to 75 mM NaCl and repurified on a HiTrap Heparin column with an elution at 1 M NaCl. Both purification columns were from Cytiva. Protein purity was assessed on a sodium dodecylsulphate–polyacrylamide gel electrophoresis (SDS–PAGE) 4–15% gel (BioRad) with a molecular weight ladder (Precision Plus Protein, Biorad) as control. The enzymes were concentrated to 3.4–3.9 mg/ml with Amicon Ultra 30k MWCO centrifugal filters (Merck), flash-frozen in liquid nitrogen and stored directly at −20°C, with no glycerol added.

Primer extension, exonuclease and thermostability assays

Polymerase activity tests were performed in 20 mM Tris–HCl pH 8.0, 20 mM NaCl, 5 mM $MgCl_2$ and 1 mM $MnCl_2$ (APEX only), unless specified otherwise. Reaction solutions contained 1 μM of templating oligo (dT_{10} and dN_{10} for APEX, dN_{10} for CHEAP; see Supplementary Table S2), 1 μM of FAM 5′-labelled DNA primer, 1 mM of dATP or a mix of four dNTPs and 1 μM of PolA. Solutions were incubated for 30 min at 37°C (APEX) or for 5 min at 20°C (CHEAP). The concentration of 3′→5′ exo-Klenow

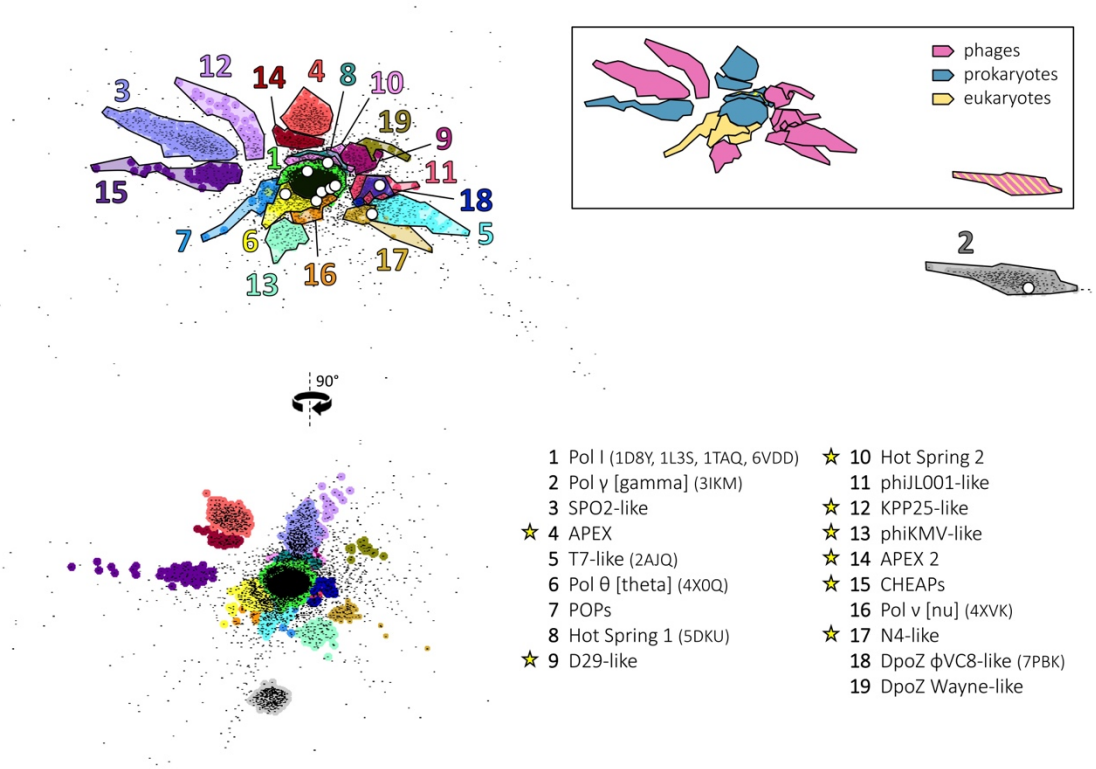


Figure 1. Non-hierarchical clustering of PolA sequences. A comprehensive dataset of all PolA sequences (black points) was obtained from UniProt in October 2021. The 3D distribution of the points was generated with CLANS (35) using the pairwise scores of sequence similarity (evolutionary distance) between individual polymerases. Two planar projections of this distribution are shown on the left (90° rotation). The 19 major clusters of PolA sequences, corresponding to PolA subfamilies, are coloured and numbered in decreasing order of size. Their names (known or proposed) and PDB codes (if applicable) are given in the key on the bottom right: yellow stars denote new subfamilies identified in this work. White dots in the top-left projection mark the position of known PolA X-ray structures. The same cluster projection was recoloured (framed insert on the top right), assigning one colour to each type of carrier species: phages (magenta), prokaryotes (blue) and eukaryotes (yellow). Double affiliation of clusters #2 and #8 is highlighted by coloured stripes.

polymerase used as a control was set at 1 U in 10 μl. Before adding the protein, DNA was hybridized by heating up to 90°C and gradually cooled to room temperature. Reactions were terminated by adding two volumes of a buffer containing 10 mM EDTA, 98% formamide and 1 mg/ml bromophenol blue, and stored at –20°C. Products were preheated at 95°C for 10 min, before being separated using PAGE and visualized by FAM fluorescence on a Typhoon FLA 9000 imager.

Exonuclease assays on the dN_{10} overhang template were conducted similarly for all enzymes (*E. coli* Pol I exo+, CT34 APEX and *I. dechloratans* CHEAP), except for the absence of four dNTPs in the reaction mixture. Solutions were incubated for 30 min at 37°C with no additional $MnCl_2$, and for 2 h at 20°C with 5 mM $MnCl_2$. Thermostability of *I. dechloratans* CHEAP was assessed by preincubating the enzyme for 10 min at room temperature, 70, 80 or 90°C before the primer extension test. All oligonucleotides were from Eurogentec, dNTPs from Fermentas (Thermo Fisher Scientific), chemicals from Sigma-Aldrich and 3'→5' exo-Klenow polymerase from New England Biolabs.

RESULTS

Clustering of available PolA sequences into 19 major subfamilies

In October 2021, we accessed the UniProt database (42) and extracted all non-fragmentary PolA protein sequences that share at most 70% sequence identity within the set. We expanded this dataset manually with 27 representative PolA entries, resulting in 8136 sequences on which we ran the clustering simulation using CLANS (35). After several hundred steps of the simulation, the sequences self-organized in 3D space in a stable manner, forming 19 distinguishable clusters (Figure 1): this distribution remained constant until the simulation's end at 6000 iterations.

The characteristics of the most prominent groups, numbered in decreasing order of size, are listed in Table 1. Seven of the eight largest clusters (#1, #2, #3, #5, #6, #7 and #8) determined in this work match the subfamilies annotated at NCBI's Conserved Domain Database (CDD; superfamily cd06444) (53)—these are the most represented and explored PolA clades. Four other clusters (#11, #16, #18 and #19) represent lesser known subfamilies introduced recently in

Table 1. Occurrence and main characteristics of the PolA subfamilies/clusters

Cluster (colour)	No. of sequences (<70% id)	Occurrence	Subfamily (CDD ID)	Representative species (PDB structure)	Polymerase function	Conservation of motif DxE in the 3′–5′ exonuclease domain	Additional domains or binding partners
1 (lime)	4560	■ Bacteria (all major phyla)	Canonical Pol I (cd08637)	■ *E. coli* (1D8Y) ■ *G. stearothermophilus* (1L3S) ■ *T. aquaticus* (1TAQ) ■ *M. smegmatis* (6VDD)	■ Lagging strand synthesis ■ Single-strand gap repair ■ Removal of RNA primers ■ Replication of plasmids	Partial (~42% sequences with catalytic residues conserved)	■ 5′–3′ exonuclease (94% sequences, N-terminal fusion)
2 (silver)	313	■ Mitochondria (opisthokonts) ■ Several cyanophages (*Caudovirales*)	Pol γ (cd08641)	■ *H. sapiens* (3IKM) ■ Phages A-HIS1, A-HIS2	■ Replication of mitochondrial DNA ■ Replication of phage DNA (putative)	Yes (all except some Glomeromycetes)	■ Interacts with PolγB dimer (*H. sapiens*), monomer (*D. melanogaster*) or functions as a single subunit (*S. cerevisiae*)
3 (lavender blue)	306	■ Phages (*Caudovirales*) ■ Prophages mainly in Firmicutes	SPO2-like (cd08642)	■ Phage SPO2	■ Replication of phage DNA	Yes (all)	■ No known partners
4 (light red)	290	■ Bacteria (Actinobacteria); does not replace canonical Pol I	APEX (established in this study)	■ *S. coelicolor* ■ *Streptomyces* sp. CT34	■ UV sensitivity reduction ■ Confirmed polymerase activity (this study) ■ DNA damage repair (probable)	No; exo catalytic pocket is tightly sealed	■ No known partners
5 (cyan)	205	■ Phages (*Caudovirales*) ■ Prophages mainly in Proteobacteria	T7-like (cd08643)	■ Phage T7 (2AJQ; in a replisome complex: 5IKN) ■ Phage S-SBP1	■ Replication of phage DNA	Yes (all)	■ Interacts with host's thioredoxin through TBD insertion on thumb's tip (~29% sequences with TBD ≥50 amino acids) ■ Interacts with TOPRIM primase-helicase (shown for T7) ■ Separate 3′–5′ exonuclease domain (S-SBP1-like)
6 (yellow)	143	■ Eukaryotes	Pol θ (cd08638)	■ *H. sapiens* (4 × 0Q)	■ DNA repair (MMEJ) ■ Template-dependent and -independent synthesis	No	■ Superfamily 2 helicase (HELQ, N-terminal fusion)
7 (light blue)	76	■ Mitochondria and plastids (non-opisthokonts); replaced by Pol γ in opisthokonts	POPs (cd08640)	■ *A. thaliana*	■ Replication and repair of organellar DNA	Yes (~89% sequences)	■ 5′–3′ exonuclease (~8% sequences, N-terminal fusion)
8 (dark green)	75	■ Diverse bacteria (mainly Aquificae and Cyanobacteria); replaces canonical Pol I in some Aquificae ■ Euryarchaeota (Methanomicrobia) ■ Apicomplexa apicoplasts	Hot Spring 1: Aquificae-like (cd08639)	■ *A. aeolicus* ■ *M. vulcani* ■ *P. falciparum* (5DKU)	■ Unknown role in prokaryotes ■ Replication of apicoplast DNA ■ Thermostable	Yes (~93% sequences)	■ AEP primase-polymerase (~7% of bacterial sequences, N-terminal fusion) ■ Polyprotein in Apicomplexa: fused to TOPRIM primase and helicase (N-terminal)
9 (purple red)	74	■ Actinomycetia phages (*Caudovirales*) ■ Actinomycetia prophages	D29-like (established in this study)	■ Mycobacterium phage D29	■ Replication of phage DNA (putative)	Yes (~99% sequences)	■ No known partners
10 (pale pink)	68	■ Diverse bacteria (mainly Acidobacteria and candidate division WWE3); does not replace canonical Pol I ■ Diverse archaea (metagenomics-derived, putative)	Hot Spring 2 (shares a recent common ancestor with Aquificae-like)	■ *Pyrinomonas methylaliphatogenes*	■ Unknown role ■ Thermostable (putative)	Yes (~96% sequences)	■ No known partners
11 (pink-red)	59	■ Phages (*Caudovirales*)	φJL001-like	■ Phage φJL001	■ Replication of phage DNA (putative)	Yes (~98% sequences)	■ No known partners
12 (lilac)	59	■ Phages (*Caudovirales*) ■ Prophages mainly in Proteobacteria	KPP25-like (established in this study)	■ Phage KPP25	■ Untested ■ Replication of phage DNA (putative)	Yes (all)	■ No known partners
13 (aquamarine)	59	■ Phages of Proteobacteria (*Caudovirales*) ■ Proteobacteria prophages	phiKMV-like (established in this study)	■ Phage phiKMV	■ Replication of phage DNA (putative)	Yes (~95% sequences)	■ ~100 amino acid insertion on thumb's tip, unrelated to TBD

Table 1. Continued

Cluster (colour)	No. of sequences (<70% id)	Occurrence	Subfamily (CDD ID)	Representative species (PDB structure)	Polymerase function	Conservation of motif DxE in the 3′–5′ exonuclease domain	Additional domains or binding partners
14 (brown-red)	53	■ Bacteria (Actinobacteria); does not replace canonical Pol I	APEX 2 (established in this study)	■ *M. pelagius*	■ DNA damage repair (probable)	No; exo catalytic pocket is tightly sealed	■ No known partners
15 (dark violet)	52	■ Bacteria (mainly Proteobacteria); does not replace canonical Pol I ■ Euryarchaeota (Methanomicrobia) ■ Related to CCPols from staphylococcal MGEs	CHEAPs (established in this study)	■ *I. dechloratans*	■ Confirmed polymerase activity (this study) ■ Highly efficient polymerase and exonuclease activities ■ Replication-related (probable)	Yes (~88% sequences)	■ No known partners ■ CCPols lack the majority of 3′–5′ exonuclease domain and form a primase-helicase complex with MD, Cch2
16 (orange)	52	■ Metazoa	Pol *ν*	■ *H. sapiens* (4XVK)	■ DNA cross-linking rescue ■ Germline meiotic homologous recombination	No	■ Interacts with Pol θ-like superfamily 2 helicase (HELQ)
17 (light brown)	45	■ Phages (*Caudovirales*) ■ Proteobacteria prophages	N4-like (established in this study)	■ Phage N4 ■ Phage KPP21	■ Replication of phage DNA (putative)	Yes (all)	■ N-terminal family 4 UDG domain (catalytic residues unconserved)
18 (dark blue)	21 (<90% id)	■ ZTGC-DNA phages (*Caudovirales*)	DpoZ φVC8-like	■ Phage φVC8 (7PBK)	■ Adenine-discriminative ■ Replication of phage ZTGC-DNA	Yes (all)	No known partners
19 (dark gold)	19 (<90% id)	■ ZTGC-DNA phages (*Caudovirales*)	DpoZ Wayne-like	■ Phage Wayne	■ Adenine-discriminative ■ Replication of phage ZTGC-DNA	Yes (all)	No known partners

The clusters and their characteristics are listed in decreasing order of size. Only seven of the 19 clusters (#1, #2, #3, #5, #6, #7 and #8) correspond to PolA subfamilies with an assigned CDD identifier; a further four have been recognized in the literature (#11, #16, #18 and #19). The established or proposed names of the subfamilies are shown in column 4. Functional tests demonstrated that previously undescribed bacterial clusters #4 (APEX 1) and #15 (CHEAPs) comprise functional polymerases (Figures 7 and 8). Experimental 3D structures of subfamily representatives are available for seven clusters (#1, #2, #5, #6, #8, #16 and #18); their PDB code is provided in parentheses in column 5. In general, PolA subfamilies perform at least two different biological roles: DNA replication or repair. They frequently differ in the functionality of the 3′–5′ exonuclease domain and in the presence of additional domains (such as 5′–3′ exonuclease) or partners. Interestingly, all PolAs from phages seem to preserve their 3′–5′ exonuclease activity, regardless of subfamily.

the literature (19,20,30,54), while the eight remaining clusters (#4, #9, #10, #12, #13, #14, #15 and #17) have not been described or recognized as separate subfamilies before. Importantly, reported phylogenetic relationships among the known subfamilies are consistent with the distribution of the clusters (see the following cluster descriptions).

To further characterize the relationships between clusters, we performed a complementary phylogenetic analysis on representative cluster sequences. A Neighbour–Joining tree calculated in Clustal Omega (45) on PolA Klenow-like large fragments reflects the distribution of the clusters, revealing 5–6 superclusters/clades (Figure 2A). This is supported by a separate bootstrap analysis performed with the Maximum Likelihood method (Figure 2B) and corroborates previous phylogenetic studies (25,28). The most abundant supercluster consists of clusters #1, #6, #7, #8, #10 and #16 present in cellular organisms. The second one contains clusters #3, #4, #12, #14 and #15, all displaying disrupted helices in the thumb subdomain (see below). The remaining clusters form pairs, with either strong (#2 and #5; #11 and #18) or weak (#13 and #17; #9 and #19) bootstrap value support. The link between the #11–#18 pair and the first supercluster is also faint. As the connections between the superclusters are even weaker and ambiguous, their exact relationship and the origin of the tree cannot be unequivocally determined.

We also coloured the 3D map generated by CLANS as a function of the type of carrier species: prokaryotes, eukaryotes or viruses (phages) (Figure 1, framed insert). Whereas cellular—bacterial and eukaryotic—clusters tend to be the largest, phage clusters are abundant. This illustrates the general high diversification of the virosphere (55), drawing from frequent genetic transfers of replication-related genes between phages and their hosts (56). Two clusters belong to more than one type of carrier organisms (#2 and #8): for these cases, the horizontal transfer of *polA* genes between cellular hosts and phages has been evidenced in the literature (see cluster descriptions below).

For each cluster, we generated sequence logos of the key functional motifs in the polymerase domain (57,58) (Figure 3); a mapping of the motifs onto the *E. coli* Pol I structure is provided in Supplementary Figure S1. The strict conservation of crucial catalytic residues implies that every subfamily corresponds to functional polymerases. While this has been proved experimentally for 10 known clades, other clusters were lacking direct experimental data; here, we do provide such data for representatives of related bacterial clusters #4 and #15 (see below). The third conserved residue of motif B—that we refer to as position B_3—participates in dideoxynucleotide discrimination (59) and modulates polymerase activity (60): it is the most variable conserved position, used to separate particular PolA clades in previous metagenomic studies (30). In the following, we pay special attention to this position.

Below, we describe each cluster in turn, starting from #1, by far the most abundant cluster (56% of classified sequences), down to #19, the least abundant.

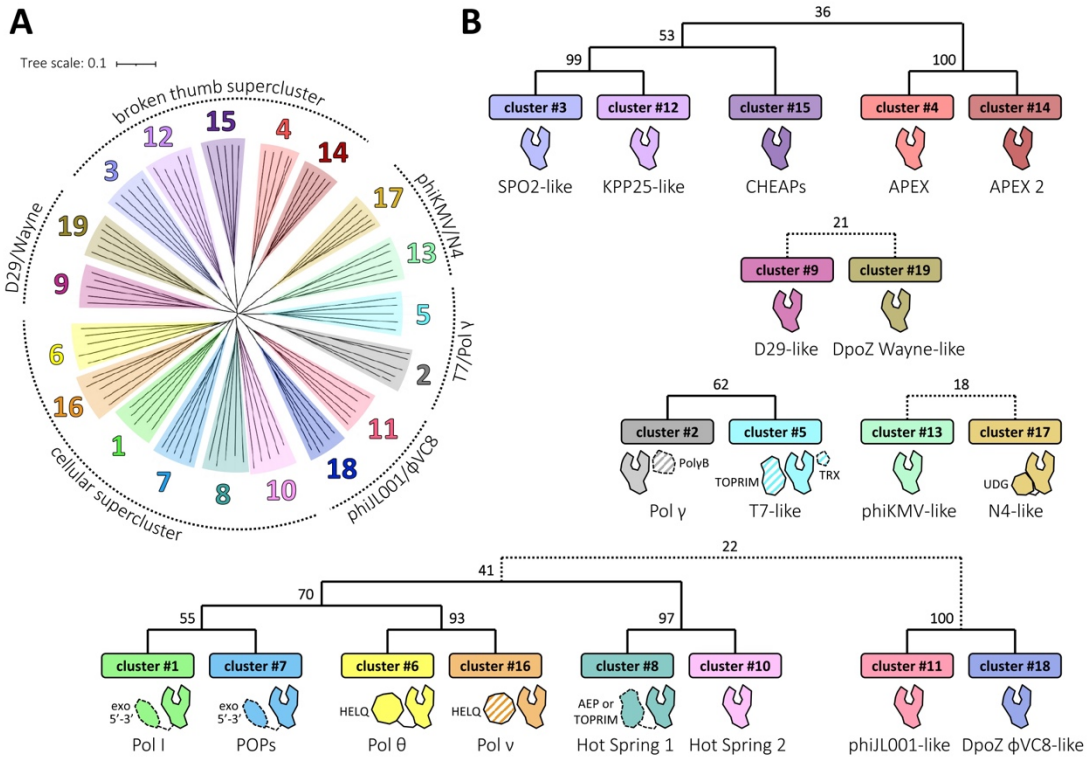


Figure 2. Phylogenetic relationships between the 19 PolA subfamilies. (**A**) Five representatives of each cluster were selected for the multiple sequence alignment of their corresponding Klenow-like large fragments, prepared with Clustal Omega (45) and visualized as an unrooted dendrogram (left). The branches were coloured to match the clusters' colours as in Figure 1; they form five or six distinct groups (superclusters) of closely related subfamilies, presented schematically around the tree. (**B**) The superclusters with supporting bootstrap values between the clusters' branches are presented on the annotated cladograms to the right and below the tree. The bootstrap values were generated with MEGA X (52), using the Maximum Likelihood method and taking 100 replicates. Dotted connections represent more distant relationships. A cartoon domain representation is shown below each cluster. Dashed contours represent domains/proteins present only for some members of a given subfamily; PolA-interacting proteins (separate polypeptide chains) are filled in stripes (see Table 1).

Cluster #1: Pol I. The largest cluster corresponds to the canonical bacterial Pol I (CDD cd08637), encompassing all known major phyla. This subfamily includes several well-described PolAs of species such as *E. coli* or *Thermus aquaticus*: the former was the first isolated and characterized DNA polymerase (3), while the latter is nowadays commonly used for *in vitro* DNA amplification (16). The vast majority of bacterial Pol I enzymes possess an additional N-terminal domain with 5′–3′ exonuclease activity (61). Independently of that fusion, their 3′–5′ exonuclease domain is often found to be deactivated with the mutation of one or several otherwise strictly conserved catalytic residues (62,63). Abundant in the cell (64), *E. coli* Pol I participates in the lagging strand synthesis, single-strand gap repair and in the removal of the RNA primer from Okazaki fragments through its 5′–3′ exonuclease activity; however, such functions can be partly compensated for by other DNA polymerases or nucleases (9,64–68). Cluster #1 PolAs may also be directly involved in the processive replication of plasmids (12).

Cluster #2: Pol γ. The second cluster contains nucleus-encoded DNA polymerases of mitochondria (CDD cd08641) in opisthokonts (animals and fungi), encoded in the cell nucleus. Known as subunit Polγ A, they interact with the accessory subunit Polγ B [homologous to class II aminoacyl-tRNA synthetases (69)] to form the functional Pol γ heterotrimer in humans (70) or the heterodimer in fruit flies (71); alternatively, they operate as a single subunit in yeast (72). Despite clear sequential and structural separation from other PolA subfamilies (73), it was suggested that Pol γ polymerases derive from T7-like PolAs; remarkably, mitochondria share with T7 phage not only their DNA polymerase but also their DNA primase and their RNA polymerase (25,74). This relationship is reproduced in our phylogenetic tree (Figure 2) and supports our clustering results, which place Pol γs as the most remote nebula of sequences, yet precisely behind the T7-like cluster #5 (Figure 1). Intriguingly, several cyanophages of the order *Caudovirales* have been found to contain Pol γ-like polymerases (26). These enzymes show the highest

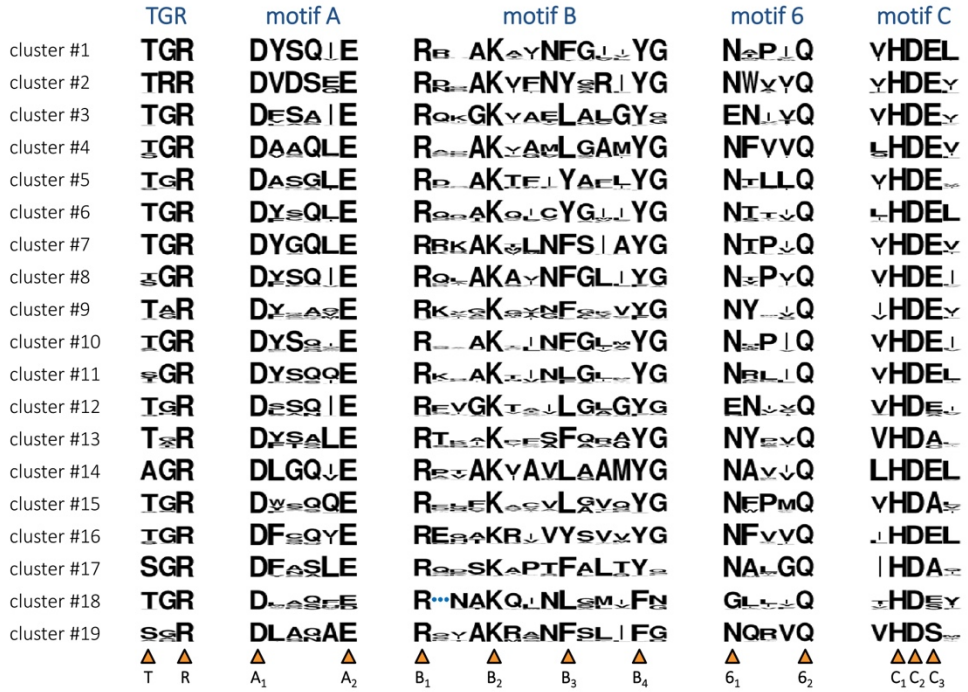


Figure 3. Sequence motifs of the 19 PolA clusters. Each cluster is presented with a sequence logo, where the height of each residue type is proportional to its frequency. Motifs A, B, 6 and C, already recognized in the PolA family (57,58), are complemented with the TGR motif, also conserved in related Klenow-fold polymerases (DNA-dependent DNA and single-subunit RNA) (140,141). Positions of highly conserved residues are marked by orange triangles below. They are given a unique label corresponding to the motif they belong to and their order of appearance. Among them, position B_3 shows the highest variability. Cluster #18-specific insertion in motif B is replaced with a blue ellipsis. See Supplementary Figure S1 for the structural context of the motifs, lining the DNA binding site and the polymerase catalytic site.

similarity with fungal Pol γ, although their functionality remains to be proven.

Cluster #3: SPO2-like. The third cluster represents a collection of bacteriophage PolAs (CDD cd08642), linked to the double-stranded DNA (dsDNA) order *Caudovirales*. Some of the carrier phages are found as prophages in bacterial genomes, mainly of the phylum Firmicutes. Like other PolAs of this subfamily, the polymerase of phage SPO2 (SP02) (15) carries a leucine in position B_3 of motif B as a less common, non-aromatic variation (Figure 3). Cluster #3 is quite divergent in sequence from other explored PolAs and lacks a resolved representative 3D structure; however, polymerase models generated by AlphaFold2 for this and other clusters are convergent and show high confidence levels of prediction (see below).

Cluster #4: APEX 1. The fourth largest cluster determined in this work is new and corresponds to bacterial polymerases of the phylum Actinobacteria. The genetic context of these polymerases does not suggest a prophage origin, yet they co-exist alongside the canonical Pol I on the bacterial chromosome. A deletion of cluster #4 PolA in *Streptomyces* indicated its involvement in DNA repair (75), although this subfamily has not been previously tested for polymerase activity; however, below we confirm the functionality of its representative. These Actinobacterial polymerases carry a distinctive leucine in position B_3 (Figure 3) and are indeed related to members of most other clusters sharing this characteristic, including cluster #3 (Figure 2). The 3′–5′ exonuclease of cluster #4 PolAs is expected to be non-functional, as it lacks a conserved DxE catalytic motif, or a conservative mutation thereof: in the following section, we also describe a distinctive reshaping in the exo catalytic pocket. In order to single out this abundant PolA subfamily and its closely related twin cluster (see below), we give them the name of Actinobacterial Polymerases with a potentially Eclipsed eXonuclease (APEX 1 and 2).

Cluster #5: T7-like. The fifth cluster concerns another, different set of *Caudovirales* PolAs (CDD cd08643), found in phage or prophage sequences predominantly in Proteobacteria. A provisional distribution of subclades constituting this cluster has been previously reported (76). Phage T7 replicative DNA polymerase, a cluster #5 PolA, binds the hosts' thioredoxin for a truly processive polymerase activity (13); grafting the thioredoxin-binding domain (TBD) onto *E. coli* Pol I dramatically increases its processivity upon binding the cofactor (77). Nevertheless, the thioredoxin-binding motif does not consistently appear

in all T7-related phages (78). The structure of the T7 replisome involving a hexameric TOPRIM primase-helicase has been recently determined (79,80). Cluster #5 PolAs contain a tyrosine in position B_3 (Figure 3).

Surprisingly, several T7-like sequences include only the polymerase domain, entirely lacking the proofreading 3′–5′ exonuclease. One such polymerase has been modelled *in silico*, based on a metagenome fragment (29). Further below, we investigate their genomic context, revealing a recent domain splitting event.

Cluster #6: Pol θ. Cluster six encompasses Pol θ (CDD cd08638) present in many eukaryotes, with the exception of fungi. These DNA polymerases are both template dependent and independent. Due to their microhomology-mediated end joining activity, they are recognized as DNA repair enzymes (11,81). Nevertheless, much like cluster #1 PolAs, their physiological role seems partly redundant and pleiotropic, also extending towards replication control during cell division in animals and plants (82,83). PolA domains of Pol θ are fused on the N-termini to a large superfamily 2 (SF2) helicase domain, both having an experimentally determined structure in humans (84,85).

Cluster #7: POPs. The seventh PolA subfamily covers the polymerases of the DNA-containing organelles—plastids and mitochondria—in non-opisthokont eukaryotes (CDD cd08640). Originally detected in plants (86,87), they were dubbed POPs (Plant Organellar DNA Polymerases); due to their universality, they have been proposed to evolutionarily precede Pol γ in organelles (88). Despite also being encoded in the nucleus, phylogenetically these two groups are considerably distinct (87,88). POPs have a functional 3′–5′ exonuclease domain; a fusion with the domain of a 5′–3′ exonuclease was observed only in singular cases (28). Lastly, cluster #7 PolAs seem to share a relatively recent ancestor with several other cellular polymerases (clusters #1, #6, #8 and #16) (28); our phylogenetic tree captures such a relationship (Figure 2).

Cluster #8: Hot Spring 1 (Aquificae-like). Cluster eight (CDD cd08639) consists of products of an adventurous (i.e. appearing in very disparate species) *polA* gene. The cluster's members were previously detected in apicoplasts of eukaryotic Apicomplexa (89) and diverse bacterial phyla (notably Aquificae) (90). In our dataset, we notice that they also appear in a group of archaeal Methanomicrobia from the phylum Euryarchaeota (e.g. *Methanolobus vulcani*, GenBank ID: WP_167879304): their genetic context is not indicative of a prophage sequence. This finding makes it the first PolA subfamily known to span across all three cellular domains of life. Frequent genetic transfers of these PolAs seem to be linked to a gene-sharing network specific to hot springs, populated by thermophilic viruses, their Aquificae hosts and archaea as well (27,91). PolA of *Plasmodium falciparum* shares with Aquificae polymerases not only sequence similarity, but also an unexpected high-temperature activity optimum (89); its X-ray structure has also been determined (92). Thus, a proposed model of their evolution involves horizontal gene transfer between phages, various bacterial phyla and apicoplasts, taking into account the loss of canonical Pol I in Aquificae (27). To underline this unique environmental context and the strong association of cluster #8 with its own twin cluster (#10), we will refer to these two clades as Hot Spring 1 and 2.

Similar polymerases are also found in thermophilic viruses/phages, such as Thermocrinis Great Boiling Spring virus (27,91); another viral metagenome-derived thermostable PolA called '3173 Pol' has been adopted for reverse transcription–PCR applications, as it accepts RNA templates (17). These PolAs are distributed in the immediate proximity of clusters #8 and #10.

In apicoplasts, cluster #8 PolAs are fused on the N-terminus with TOPRIM primase and helicase domains, whose functionality as a polyprotein has been confirmed (89,93); in a subset of phages with a similar polymerase, a polyprotein fusion with a putative helicase domain has also been observed (27). Additionally, we observe that in several bacteria representing various phyla (i.e. Verrucomicrobia, Planctomycetota and Nitrospirae) the cluster #8 PolA is fused to an AEP primase-polymerase. In general, Hot Spring PolAs conserve the 'canonical' motif B_3 phenylalanine (Figure 3), similarly to related Pol I and POPs (clusters #1 and #7).

Cluster #9: D29-like. The ninth largest cluster contains polymerases of *Caudovirales* bacteriophages (and their prophages) preying on Actinomycetia (phylum Actinomycetia). Although phage D29, the prototypical carrier phage of this PolA subfamily, was discovered almost 70 years ago (94), this is the first time that D29-like polymerases are recognized as a separate, highly diverged clade (Figure 2). Nevertheless, the crucial motifs of cluster #9 PolAs stay typical (Figure 3), including the well-conserved DxE catalytic motif in the 3′–5′ exonuclease domain; these polymerases are devoid of any significant insertions.

Cluster #10: Hot Spring 2. The 10th cluster, named Hot Spring 2, closely mimics cluster #8 of Aquificae-like PolAs (Hot Spring 1): the two groups share similar conservation profiles (Figure 3) and a recent common ancestor (Figure 2), although their separation is supported by both clustering and phylogenetic analyses. Organisms carrying cluster #10 polymerases involve bacterial- and metagenomics-derived putative archaeal species: they include predominantly distinct phyla (i.e. Acidobacteria, candidate division WWE, Nanoarchaeota and Thorarchaeota) but exclude Aquificae.

Cluster #11: φJL001-like. PolAs from cluster #11 belong to yet another fraction of *Caudovirales* phages. They are also found in multiple short sequence fragments annotated as bacteria, but such DNA portions directly correspond to the viral ones in length and composition. This subset of PolAs has been observed predominantly in marine virioplankton (29). It is so far the third cluster characterized by a leucine in position B_3, despite an apparent interchangeability with phenylalanine (Figure 3) and lack of close homology with other leucine-bearing clusters, except for #18 (Figure 2). In this group resides the PolA of phage φJL001, whose genome has been described in detail (95). Nonetheless, cluster #11 still lacks an experimentally determined structural representative.

Cluster #12: KPP25-like. Next to cluster #3 one finds one more group of *Caudovirales* PolAs (Figure 1); their genes are often integrated as prophages into proteobacterial genomes as well. Up to now, this subfamily had remained completely undescribed, although a representative was identified in phage KPP25 through routine homology searches (96): we will therefore refer to cluster #12 PolAs as KPP25-like. In agreement with their evident relationship with SPO2-like PolAs (Figure 2), these DNA polymerases display the distinctive position B_3 Leu variant (Figure 3). Their typical length is 600–650 amino acids, with all the activity-related residues conserved in both polymerase and 3′–5′ exonuclease domains.

Cluster #13: phiKMV-like. The 13th cluster comprises another set of PolAs found in phages of Proteobacteria, or in their prophages. This family is represented by the polymerase of phage phiKMV (97), previously described as a T7-like phage: nevertheless, phiKMV belongs to a distinct taxonomic subfamily, while the differences between T7-like and phiKMV-like PolAs are even more pronounced (98) (Figures 1 and 2). For example, cluster #13 polymerases have a phenylalanine in position B_3, instead of a tyrosine specific to the T7 clade (Figure 3). Interestingly, they also possess an extensive (∼100 amino acid) insertion in the thumb subdomain, which could be reliably modelled for multiple representatives (see the following section). Its placement follows the helix H1, in contrast to the TBD of T7 PolA that precedes it (99). There seems to be no phylogenetic relationship between the two, although the peculiar, structured extension of phiKMV-like PolAs may also perform a role related to processivity.

Cluster #14: APEX 2. Similarly to cluster #10, cluster #14 acts as a twin cluster to a larger clade. Despite its strong resemblance to cluster #4 (APEX 1) in sequence and occurrence, the two subfamilies consistently split in a sufficiently large dataset (Figures 1 and 2). We note the exceptional motif conservation of the smaller cluster, named APEX 2, even on usually variable positions, suggesting its relatively recent separation. Cluster #14 represents the only PolA subfamily where an alanine replaces threonine in the TGR motif (Figure 3).

Cluster #15: CHEAPs. Cluster #15 is the last major cloud of bacterial PolA sequences. They are associated essentially with Proteobacteria, but, in a surprising parallel with the unrelated cluster #8 (Figure 2), also with some archaeal Methanomicrobia. Some of these PolAs are annotated as thermostable; however, a thorough search revealed that all such instances were inferred through homology and that no cluster #15 representative has been described before. Cluster #15 PolAs conserve all functional exonuclease and polymerase motifs, with a leucine found in position B_3 (Figure 3), in agreement with their close homology to polymerases SPO2-like, KPP25-like and APEX (Figure 2). Much like APEX (clusters #4 and #14), these enzymes do not replace the canonical bacterial Pol I. An example of a reference carrier organism is *I. dechloratans*, a Betaproteobacteria (GenBank ID: WP_151124575): below, we confirm high templated polymerase and exonuclease activities, as well as the lack of thermostability of its PolA. We therefore name this subfamily Cellular Highly Efficient Auxiliary Polymerases (CHEAPs).

In the proximity of cluster #15, we observed a few truncated PolA sequences: although under-represented, they correspond to a unique class of PolAs (CCPols) with an incomplete exo domain (see below), which are associated with staphylococcal mobile genetic elements (MGEs) (100). CCPol of *Stapylococcus aureus* interacts with a small protein (MP) and a helicase (Cch2); importantly, the CCPol–MP complex displays priming activity (100). A supplementary phylogenetic evaluation confirms that CHEAP is the closest subfamily to CCPols.

Cluster #16: Pol ν. Cluster #16 represents polymerases ν, a young branch of PolAs that arose in animals (54). They display strong sequence similarity with Pol θ as well as with canonical bacterial Pol I (cluster #1), and were reported to match the indel profile of Pol θ (101). These tight evolutionary relationships (Figures 1 and 2) introduce some confusion as to the identity of protozoan Pol ν/Pol θ-like PolAs (102,103). Polymerases ν lack additional domains, although they do interact with a Pol θ-related superfamily 2 helicase (10). Despite being able to rescue DNA cross-linking *in vitro*, their physiological role is associated with meiotic homologous recombination in germline cells (54,10). The crystal structure of human Pol ν has been solved (104).

Cluster #17: N4-like. Cluster #17 is formed by yet another set of PolAs from *Caudovirales* phages of Proteobacteria, including their prophage form. They are comparatively long, usually comprising ∼800–900 amino acids: this results from the fusion on the N-terminus with a family 4 uracil-DNA glycosylase (UDG), an enzyme that typically removes uracil from DNA strands, leaving an abasic site (105). To this cluster belongs the polymerase of phage N4 (106). N4-like PolAs display distant homology to phiKMV-like PolAs (Figure 2), featuring a phenylalanine in position B_3 as well (Figure 3).

There exist other well-known bacteriophages carrying a family A polymerase fused to a family 4 UDG domain: the most notable examples include *Bacillus* phages SPO1 (SP01) (14), SP-10 and SP-15 (107). Their polymerases do not cluster together with N4-like polymerases, although all UDG-PolAs are found in close proximity (Supplementary Figure S2). Importantly, all three phages contain modified uracil nucleotides in place of thymine in their genomes (108–110). The UDG domain of UDG-PolA was speculated to provide selectivity towards 5-hydroxymethyluracil (5hmU) (111), which can then be post-replicatively hypermodifed by glucosylation (112). N4-like phages are not known to modify their DNA, which is consistent with the observation that the catalytic residues in the UDG domain of cluster #17 PolAs have been replaced or deleted (see the following section).

Cluster #18: φVC8-like DpoZ. The penultimate and most recently described PolA clade concerns φVC8-like DpoZ enzymes found in some *Caudovirales* phages that replace their genomic adenine with 2-aminoadenine (Z), resulting in saturated interstrand hydrogen bonding in the phage DNA (19,113). As expected, these

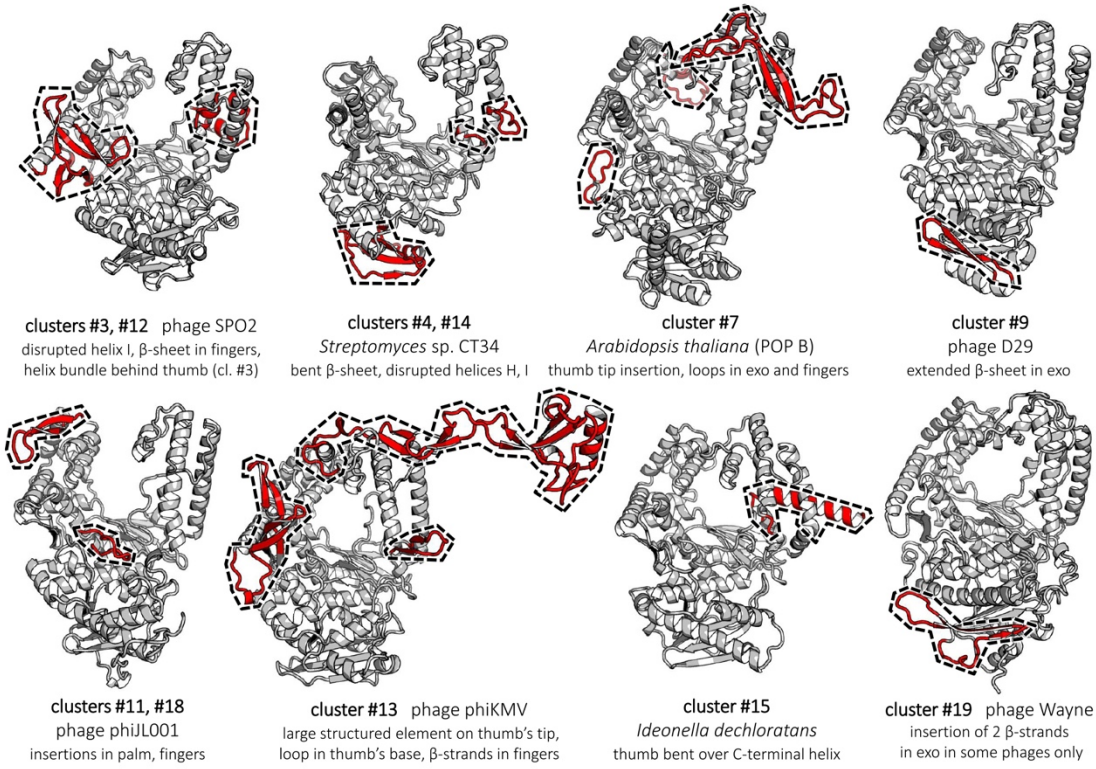


Figure 4. Ribbon representation of PolA structures predicted with AlphaFold2 (39), revealing idiosyncratic structural elements. Keys below the cluster representatives include the associated clusters, source organisms and the description of prominent features conserved within the clusters. These features are highlighted on the models in red and marked by a dashed contour; their amino acid sequence boundaries are specified in Supplementary Table S3. The remaining typical common structural elements are in grey. All structures are viewed from the same angle.

polymerases are Z-specific and substantially—although not completely—discriminate against adenine (19,20). Yet, the net incorporation of Z into the DNA of φVC8 and related phages is also modulated by a conserved dATPase (DatZ): importantly, the ZTGC-DNA cyanophage S-2L has DatZ, but lacks a Z-specific polymerase, demonstrating that DpoZ is in fact dispensable for a complete A-to-Z substitution (113–116). φVC8-like DpoZ show close similarity to φJL001-like PolAs from cluster #11 (Figures 1 and 2). We recently reported the experimental structure of the apo form of φVC8 DpoZ (20), providing a rationale for its specific sequence features (19), some of which were also found to be shared with φJL001 PolA.

Cluster #19: Wayne-like DpoZ. The last cluster contains Wayne-like DpoZ, the second group of PolA enzymes specific to 2-aminoadenine, also found in *Caudovirales* (19). Despite their equivalent functionality, φVC8-like DpoZ are clearly distinct from Wayne-like DpoZ enzymes (19,20) (Figures 1–4). Like the former, the latter show closer homology with an ATGC-DNA-related PolA subfamily (cluster #9), supporting the hypothesis of convergent DpoZ specialization (20), which stands in contradiction to a postulated congruent evolution with PurZ, a key enzyme in Z synthesis (19). Interestingly, both DpoZ clusters share a unique substitution, carrying phenylalanine in position B_4 of motif B (Figure 3): it corresponds to the residue helping to discriminate between the dNTP and NTP substrates (117), in the vicinity of the steric gate with similar functionality (position A_2) (118). Nevertheless, it is unlikely to influence the discrimination of adenine versus 2-aminoadenine (119).

Predicted structural features of PolA subfamilies lacking a crystallographic structure

Structural differences among well-characterized PolA subfamilies are sometimes subtle (Supplementary Figure S3), yet they often prove to be functionally important. Therefore, we aimed to investigate structural features of all 12 clusters that lack an experimental structure (novel or not). We ran AlphaFold2 on five representative sequences in each subfamily. The resulting models, and in particular the observed new features, exhibited high confidence levels of prediction reflected in excellent pLDDT scores (39). These novel elements are characteristic of the individual PolA clusters and appear in all modelled representatives, greatly expanding the known diversity of the PolA fold (Figure 4). Three structured insertions stretch over ∼50 amino acids or

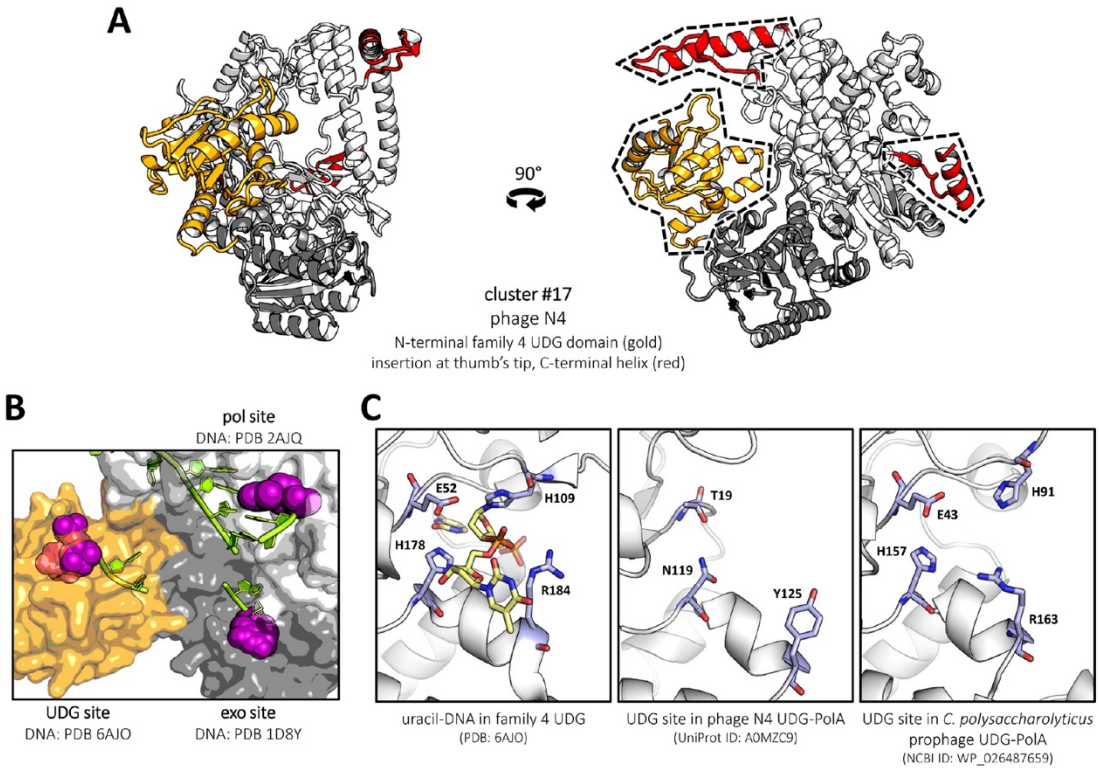


Figure 5. The predicted structure of phage N4 DNA polymerase, representing cluster #17, containing a supplementary domain matching family 4 uracil-DNA glycosylases (UDGs). (**A**) Ribbon representation, front and side view. The particular position of the UDG domain (gold) on the interface between the exo (grey) and pol (white) domains is conserved among the clade and other, unclassified UDG-PolAs. Other features specific to N4-like UDG-PolAs (red) are mapped and labelled as in Figure 4. (**B**) Surface representation, top-sideways angle. The three catalytic sites of N4 UDG-PolA are filled with DNA strands modelled from other structures (DNA strands shown as filled lime sticks, nucleotide substrates represented by purple spheres) and labelled above or below. All three pockets reside equidistantly from each other (30–35 Å). (**C**) Comparison of a functional family 4 UDG protein UdgX (left) with AlphaFold2 models of representative UDG-PolAs. While a deletion and several mutations in the catalytic site clearly inactivate the glycosylase domain of N4-like enzymes (middle), some of the unclassified UDG-PolAs preserve all structural elements and catalytic residues (right). Descriptions including PDB or GenBank IDs are shown below the panels.

more (Supplementary Table S3), joining the group of long PolA-specific insertions also found in T7-like polymerases (TBD) (77) and Pol γ (accessory-interacting determinant, AID) (73).

In cluster #3 PolAs (SPO1-like), we identified an expansion of a β-hairpin in the fingers subdomain by several novel β-strands (Figure 4). This structural element partly overlaps with structural elements in phage T7 PolA (cluster #5), found to be involved in template strand stabilization and in the conformational transition from the elongation to the editing modes (120). Additionally, one helix of the thumb subdomain is disrupted, although an additional stabilization is provided by multiple contacts with a bundle of small helices, whose position corresponds to a β-hairpin in Pol θ and Pol *ν* (Supplementary Figure S3). Cluster #12 polymerases (KPP25-like) show similar features, except for the presence of the helical bundle.

The 'broken thumb' subdomain reappears in related clusters #4, #14 and #15 (APEX 1, APEX 2 and CHEAPs), this time concerning both helices of the stem: in CHEAPs, the thumb bends backwards on a supporting C-terminal helix (Figure 4). We also observe a substantial reshaping of the 3′–5′ exo domain in both APEX subfamilies, resulting in a β-sheet being bent away from the catalytic pocket's side. In APEX 1, a flexible loop blocks the entrance to the inactive 3′–5′ exo site, although APEX 2 models show that this loop can assume a different conformation that unlocks the pocket (Supplementary Figure S4).

Cluster #7 POPs have two loop insertions located below and on the top of fingers, and an additional β-hairpin on the tip of the thumb subdomain. The latter is reminiscent of the TBD of T7 PolA (Figure 4; Supplementary Figure S3), despite their divergent topology.

PolAs D29-like and related Wayne-like DpoZ (clusters #9 and #19) have the typical Klenow fold that is only expanded by two β-strands in the exo domain. Likewise, PolA of phage φJL001 (cluster #11) has two insertions that are shared with φVC8 DpoZ (cluster #18) (20). Indeed, structural predictions of φJL001-like PolAs match closely the experimental structure of the latter (Figure 4; Supplementary

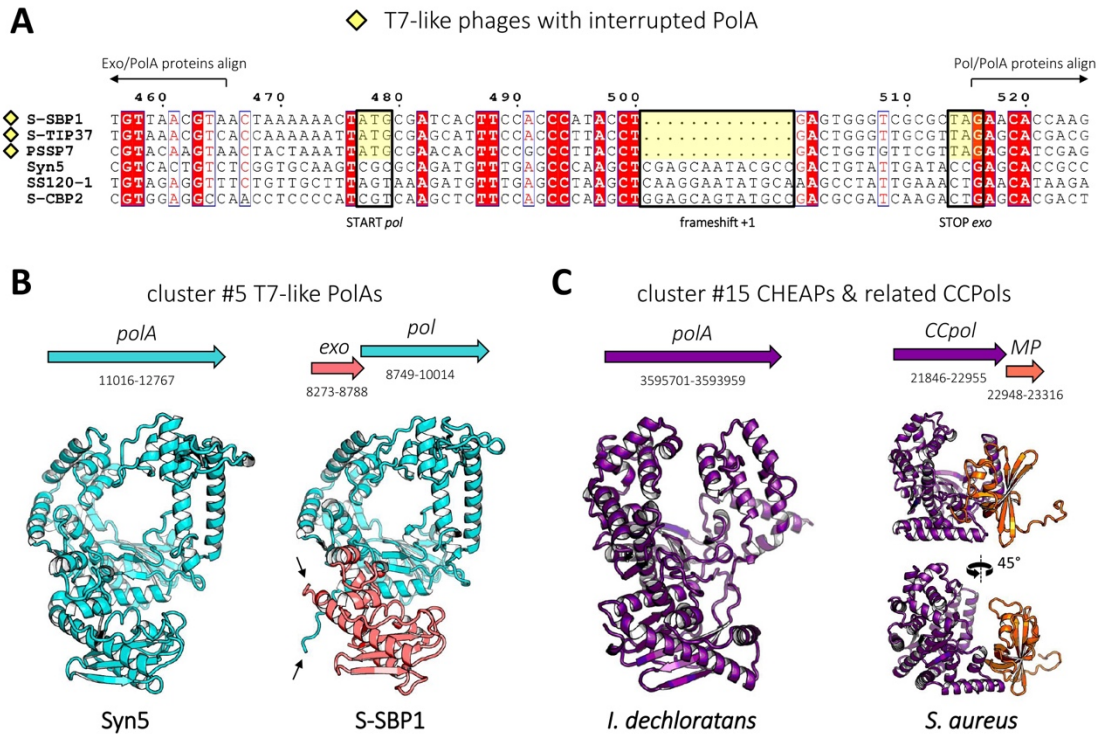


Figure 6. PolA enzymes missing the 3′–5′ exonuclease domain. (**A**) Genomic multialignment of T7-like PolAs (cluster #5) illustrating the genetic separation of the exo and pol domains. The disruption site of S-SBP1-like PolAs (yellow square) is compared with their close, but uninterrupted homologues. Phage names are given on the left. The start codon of *pol*, the stop codon of *exo* and the 1 nt deletion leading to a frameshift and termination of *exo* are conserved only among the interrupted PolAs (yellow boxes). Amino acid sequences of S-SBP1-like PolAs do not align with their uninterrupted homologues in the region of *exo*/*pol* overlap. (**B**) Predicted structure of a complete PolA of phage Syn5 (left, cyan) is essentially identical to the predicted complex of phage S-SBP1 *exo* and *pol* gene products (right, light red and cyan), except for the loose N- and C-termini introduced by the disruption (black arrows). Graphical representation of the corresponding genes along with their nucleotide boundaries is shown above the models, to scale. (**C**) Similar representation for exonuclease-truncated CCPol of *S. aureus* (dark violet, right, standard and 45°-rotated view) and its relative, *I. dechloratans* CHEAP (cluster #15, dark violet, left). CCPol keeps three helices of the exonuclease domain, through which it interacts with the MP protein (orange), according to the AlphaFold2 model.

Figure S3). In addition, despite quite diverging sequences, the shape of the mobile helices E1 and E2 engulfing the exonuclease's catalytic pocket (20) is conserved among the two subfamilies. Similar elements are also observed in phage T7 PolA, although in this case a corresponding phylogenetic connection is missing.

PhiKMV-like polymerases of cluster #13 are more heavily modified. They display a small insertion at the thumb's base, and a larger one involving multiple new β-strands on the interface between the exo domain and fingers subdomain, yet forming a different arrangement from that in SPO2-like/KPP25-like PolAs discussed above. Most importantly, phiKMV-like polymerases possess a long insertion at the thumb's tip that extends far away (60–70 Å) from the enzyme's core fold, through several β-strands. This insertion is structurally and phylogenetically unrelated to the TBD in T7-like PolAs, although it is equally well positioned for a potential interaction with nascent dsDNA (Supplementary Figure S5).

Unlike other N-terminal fusions, the uracil-DNA glycosylase domain of all UDG-PolAs—inside and outside of cluster #17—occupies a well-defined position in the structure, sandwiched between the edges of the 3′–5′ exonuclease and polymerase domains (Figure 5A). Interestingly, the catalytic sites of the three domains face each other and are almost equidistant (Figure 5B). We observe that an unfolded nascent DNA strand could possibly access either the exo or the UDG active site through relatively simple conformational transitions. In some UDG-PolAs, although not in N4-like enzymes, the family 4 UDG domain has retained all catalytic residues participating in the uracil base excision (121) (Figure 5C). In cluster #17 polymerases, however, the UDG domain lacks a short helix and an important β-hairpin carrying a histidine residue crucial for covalent binding of dU along the catalytic path. It is possible that the UDG domain confers a second editing mode in non-N4-like UDG-PolAs with an active UDG domain, which may be linked to the presence of 5hmU in the DNA of

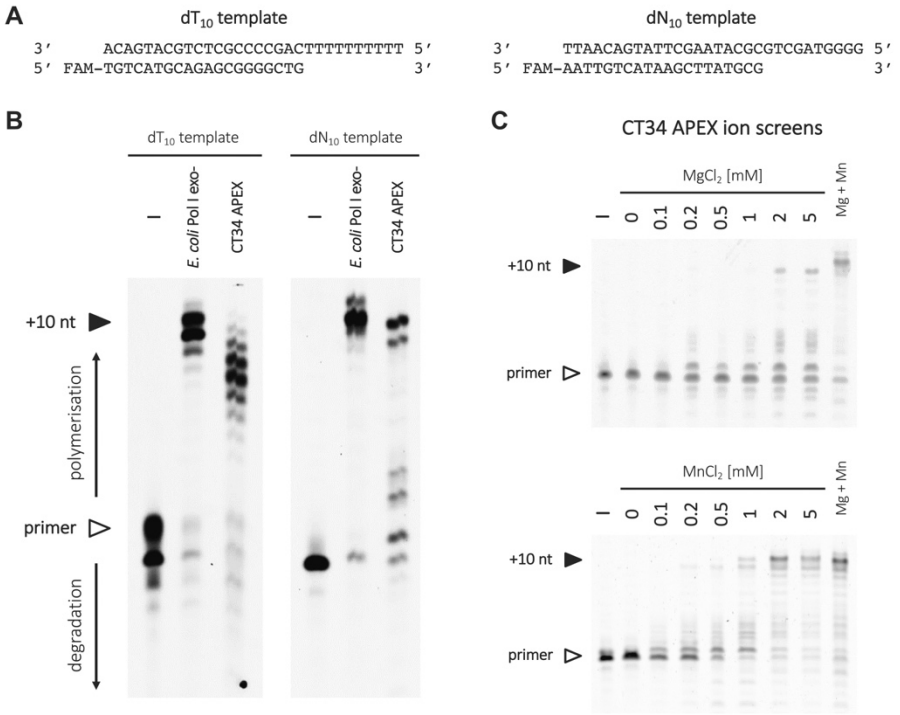


Figure 7. Polymerase activity of *Streptomyces* sp. CT34 APEX (cluster #4). (**A**) Two different substrates were used in the assay, with either dT_{10} or random dN_{10} template overhangs. (**B**) Enzymatic extension of the primer visualized on a polyacrylamide gel. Reaction mixtures were incubated at 37°C for 30 min. Lanes to the left represent a negative control without any polymerase, and a positive control with *E. coli* Pol I (Klenow fragment, 3′–5′ exo-). Bands corresponding to the primer are marked with a white arrow to the left, and fully extended products (+10 nt) with a black arrow. (**C**) Top: a screen for optimal $MgCl_2$ concentration, between 0 and 5 mM, and no additional $MnCl_2$. Bottom: a similar screen for $MnCl_2$, with no $MgCl_2$ added. The lanes to the left (–) correspond to a negative control, without polymerase; the lanes to the right (Mg + Mn) correspond to CT34 APEX in the presence of both 5 mM $MgCl_2$ and 1 mM $MnCl_2$.

phage SPO1 or similar phages with UDG-PolA (108). Such a possibility is supported by the presence of genes related to nucleotide modification and dUTP processing directly upstream of the UDG-PolA gene in a prophage of *Caldanaerobius polysaccharolyticus* (NCBI ID: WP_026487659), where the UDG domain has all the catalytic residues conserved (Figure 5C). Alternatively, in the case of UDG inactivation, the domain could possibly increase polymerase processivity during replication.

PolAs missing the 3′–5′ exonuclease domain: exo/pol domain separation (cluster #5) or formation of a complex with MP (CCPols)

In a number of full-length T7-like phages sequenced recently, such as phage S-SBP1, a truncated polymerase gene is found to correspond only to the pol domain, in agreement with previous results based on metagenomic data (29). Parsing complete genomic sequences, we established that this gene (which we call *pol*) places itself next to a gene of T7-like 3′–5′ exonuclease, which we call *exo*. A sequence comparison of truncated and complete close relatives revealed that the ancestor of S-SBP1-like PolAs arose through a +1 frameshift leading to a new STOP codon 13 bp downstream, which terminates the translation of the exonuclease domain (Figure 6A). The simultaneous appearance of a start codon 21 bp upstream of the frameshift ensures the translation of the remaining domain in the original reading frame. This finding indicates that the domains were separated at the genetic level through gene fission, and dismisses the possibility of a constitutive translational frameshift that has been observed in other phages (122). Moreover, an AlphaFold2 structure prediction of the binary complex between the two separate S-SBP1 *exo* and *pol* gene products results in a perfect superposition with a model of a related, uninterrupted PolA of phage Syn5; the new protein termini of the split polymerase are simply exposed to the solvent (Figure 6B). This implies that the association between the exo and pol domains is most probably preserved.

Although phages with a split PolA are related to phage T7, it is not known whether their polymerases interact with the host's thioredoxin as well. These PolAs display a 32 amino acid deletion in the TBD region, shortening this structural element by half. An AlphaFold2 prediction failed to predict a complex of S-SBP1 PolA with either of the two thioredoxin proteins of the host, *Synechococcus* sp. WH7803 (123) (UniProt IDs: A5GN01, A5GM53).

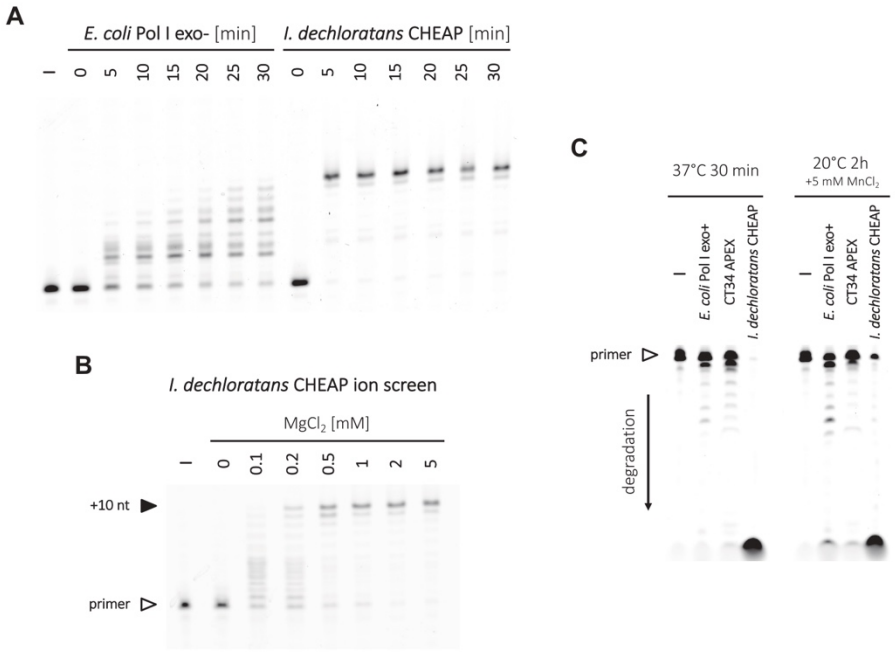


Figure 8. Primer extension activity of *I. dechloratans* CHEAP (cluster #15) and its strong exonuclease activity compared with CT34 APEX and *E. coli* Pol I. (**A**) Primer extension assay visualized on a polyacrylamide gel, with a dN_{10} overhang templating oligonucleotide (Figure 7A). *E. coli* Pol I (Klenow fragment, 3′–5′ exo-) and *I. dechloratans* CHEAP were incubated at 20°C for 0–30 min, as specified above the lanes. The lane to the left represents a negative control without any polymerase. (**B**) A corresponding $MgCl_2$ concentration screen, between 0 and 5 mM, during 5 min incubation of the primed dN_{10} template with *I. dechloratans* CHEAP at 20°C. The negative control on the left had no polymerase added. (**C**) Exonuclease activity of *E. coli* Pol I (Klenow fragment), CT34 APEX and *I. dechloratans* CHEAP, shown for two different conditions specified above the gel images. Reaction mixtures included the dN_{10} overhang template strand and a labelled primer, with no added dNTPs.

In contrast to split T7-like PolAs, in the vicinity of the truncated *CCPol* gene we have not detected the complementary gene of the missing exo fragment (~110 amino acids). Nonetheless, CCPols may retain the appropriate fold stability thanks to the three remaining exonuclease helices and unique interactions with MP and Cch2 (100). Indeed, AlphaFold2 predicts a stable MP–CCPol complex between the remainder of the exo domain and the MP protein, which forms a five-stranded β-barrel: their surface of interaction spans 796.4 Å^2 (Figure 6C).

Catalytic activities of bacterial *Streptomyces* sp. CT34 APEX (cluster #4) and *I. dechloratans* CHEAP (cluster #15)

The conserved catalytic residues present in the pol domain of APEX (cluster #4 and #14 PolAs) indicate that the subfamilies should be functional. To confirm this experimentally, we cloned the gene of one such PolA found in the NCBI RefSeq database, present in the genome of *Streptomyces* sp. CT34 (GenBank WP_043265455). We overexpressed CT34 APEX in *E. coli*, purified its His-tagged version and subjected it to primer extension assays, using polythymine (dT_{10}) or random nucleotide (dN_{10}) overhanging sequence as templates (Figure 7A).

We found the protein to be an active DNA polymerase, albeit not in a very processive way in our experimental conditions (Figure 7B). We screened the optimal Mg^{2+} concentration and found that the activity reaches a plateau at 5 mM; adding 1 mM Mn^{2+} increased it noticeably further (Figure 7C). Conversely, while the Mn^{2+} concentration screen levels off at 1 mM, we observed that adding 5 mM Mg^{2+} moderately improved the activity. We compared this behaviour with the optimal concentrations of divalent ions for different polymerases presented in a recent review (124), and conclude that CT34 APEX most closely resembles DNA polymerases performing DNA repair. This function would be consistent with its low processivity and exonuclease inactivation.

Using the same approach, we tested the activity of a reference enzyme from another newly defined bacterial PolA cluster. After purification of *I. dechloratans* CHEAP, we subjected it to a primer extension assay using the dN_{10} template. The protein is more active than *E. coli* Pol I, reaching its optimum at 1 mM Mg^{2+} with no additional Mn^{2+} needed (Figure 8A, B). Although automatic annotations of close homologues suggested possible thermostability of the enzyme, pre-incubation of *I. dechloratans* CHEAP at 70, 80 or 90°C for 10 min rendered the enzyme inactive (Supplementary Figure S6).

Finally, we examined the exonuclease activities of CT34 APEX and *I. dechloratans* CHEAP (Figure 8C). Marginal degradation of the primer was observed for exonuclease-

inactivated CT34 APEX: this trace activity probably arises from pyrophosphorolysis in the pol domain, which is inherently coupled to DNA and RNA polymerization as their reverse reaction (125,126). Contrastingly, *I. dechloratans* CHEAP shows pronounced exonuclease activity: this capacity for proofreading indicates that it acts as a genuine replicative DNA polymerase.

DISCUSSION

Being an ancient class of DNA replicators (24,127), family A of DNA polymerases displays the expected diversification of its extant progeny. The complete catalogue of major subfamilies, enriched by the newly determined clusters, allows for its holistic, up-to-date description. It lays the foundation for more sophisticated phylogenetic methods that could shed some light on the earliest evolutionary paths from which PolAs emerged. Nonetheless, our clustering does not include smaller subfamilies, such as mitochondrial PolAs of *Trypanosoma*-like euglenozoa and certain phages (128), phage T5-like PolAs (129), the aforementioned MGE-related CCPols (100) or other UDG-PolAs outside of the N4-like family: these already described polymerases do not form a proper cluster at the present time. Yet, the ever-growing number of deposited sequences promises that data available in the near future will be sufficient for more comprehensive analyses.

In our study, we could determine that the replacement of tyrosine or phenylalanine with leucine in position B_3 is present in two separate superclusters: the 'broken thumb' supercluster encompassing APEX and CHEAPs, and the φJL001/φVC8 supercluster (Figure 2). Therefore, the position B_3 as an evolutionary marker (30) should be used with caution and preferably in concert with full sequence data, in order to correctly infer common origins of given clades. To date, only one structure of a PolA carrying the leucine B_3 variant has been experimentally determined—φVC8 DpoZ (PDB ID: 7PBK). The lack of a detectable relationship between the two DpoZ subfamilies—including their divergence in position B_3—is a clear indication for functional convergence inside the PolA family concerning the incorporation of the base Z. New sequencing data could reveal whether the shift of specificity from A towards Z has also arisen in other DNA polymerase families, or other PolA subfamilies.

PolAs are known to structurally require the 3′–5′ exonuclease domain—even in an inactive form—for stability (130,131). The discovery of exo and pol domain separation in a number of T7-like phages (cluster #5) indicates that at least some PolA (S-SBP1-like) enzymes are split *in vivo* into two interacting components. Despite being an oddity among family A members, such a split is reminiscent of constitutive subunits DP1 (proofreading) and DP2 (elongation) of family D DNA polymerases (132), or multisubunit DNA-dependent RNA polymerases that also have their single-subunit counterparts (133). The separation could entail a differential regulation of the two functions at the gene level or might be necessary for a large conformational change and domain rearrangement during the catalytic cycle. In contrast, the unique example of the CCPol–MP complex predicted by AlphaFold2 demonstrates that a structural substitution of the 3′–5′ exonuclease domain is also possible.

The new structural features predicted by AlphaFold2 for 12 structurally unresolved clusters are all located outside of the catalytic sites; yet, they may contribute to the enzymes' processivity, stability, their inherent essential dynamics or the binding of potential partners. Such appendices could indirectly influence the catalytic activity of a polymerase, for instance by modifying the conformational space spanned by helix O in the fingers domain, involved in mismatch detection (134). Globally, reliable 3D models can also inform deep phylogenetic searches, as the conservation of a structure takes precedence over that of a sequence (135).

The third domain with a possible UDG activity found in cluster #17 and other phage-related UDG-PolAs transcends a simple polypeptide chain fusion: in all predicted models, this family 4 UDG homologue maintains its firm position and proximity to both pol and exo sites without apparent clashes with the DNA reactants. In principle, PolA could smoothly integrate the UDG activity after polymerase backtracking and before proofreading (136,137), generating an abasic site before its removal by the exonuclease. It remains to be seen whether the putatively active domain found in some UDG-PolA sequences plays a role in the recognition of uracil, 5-hydroxymethyluracil or thymine, possibly participating in the maintenance of DNA modification in some phages. It is also conceivable that the UDG domain acts merely as a processivity module in UDG-PolAs with an inactive UDG, e.g. in N4-like enzymes.

Finally, we present the evidence that two prominent—so far unexplored—bacterial PolA subfamilies, referred to here as APEX and CHEAPs, consist of functional DNA polymerases: their respective activities may be structurally linked to specific substitutions in helix J, which participates in dsDNA binding and regulates the primer extension–proofreading equilibrium (138). Similar activity tests are still needed for phage-derived enzymes of the other new clusters.

Ultimately, detailed knowledge about the differences among the existing PolA subfamilies may inform the choice of specific polymerase candidates during goal-oriented mutagenesis or directed evolution. In a complementary approach, desirable PolA features—such as processivity factors or accessory domains—could be rationally selected and assembled in chimeric enzymes (139). In this way, engineered PolAs with desired traits would have the potential to meet new laboratory or biotechnological needs.

DATA AVAILABILITY

All data, including the final PolA dataset, all sequences from individual clusters, representative AlphaFold2 models and phylogenetic trees are available in the Supplementary Data.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We thank Sandrine Rosario for mass spectrometry experiments and Dr Sophia Missoury for helpful discussions and critical reading of the manuscript. We thank the Molecular Biophysics and Macromolecular Interactions Platform at Institut Pasteur for help in characterizing the purified proteins by mass spectrometry.

FUNDING

We thank ANR (Grant ANR 20 CE11 002603 Break-Dance) for travelling funds allowing the completion of this project.
Conflict of interest statement. None declared.

REFERENCES

1. Crick,F.H. (1958) On protein synthesis. *Symp. Soc. Exp. Biol.*, **12**, 138–163.
2. Johnston,W.K., Unrau,P.J., Lawrence,M.S., Glasner,M.E. and Bartel,D.P. (2001) RNA-catalyzed RNA polymerization: accurate and general RNA-templated primer extension. *Science*, **292**, 1319–1325.
3. Lehman,I.R., Bessman,M.J., Simms,E.S. and Kornberg,A. (1958) Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from *Escherichia coli*. *J. Biol. Chem.*, **233**, 163–170.
4. Raia,P., Delarue,M. and Sauguet,L. (2019) An updated structural classification of replicative DNA polymerases. *Biochem. Soc. Trans.*, **47**, 239–249.
5. Guilliam,T.A., Keen,B.A., Brissett,N.C. and Doherty,A.J. (2015) Primase-polymerases are a functionally diverse superfamily of replication and repair enzymes. *Nucleic Acids Res.*, **43**, 6651–6664.
6. Blanco,L. and Salas,M. (1984) Characterization and purification of a phage phi 29-encoded DNA polymerase required for the initiation of replication. *Proc. Natl Acad. Sci. USA*, **81**, 5325–5329.
7. Redrejo-Rodríguez,M., Ordóñez,C.D., Berjón-Otero,M., Moreno-González,J., Aparicio-Maldonado,C., Forterre,P., Salas,M. and Krupovic,M. (2017) Primer-independent DNA synthesis by a family B DNA polymerase from self-replicating mobile genetic elements. *Cell Rep.*, **21**, 1574–1587.
8. Ollis,D.L., Brick,P., Hamlin,R., Xuong,N.G. and Steitz,T.A. (1985) Structure of large fragment of *Escherichia coli* DNA polymerase I complexed with dTMP. *Nature*, **313**, 762–766.
9. Okazaki,R., Arisawa,M. and Sugino,A. (1971) Slow joining of newly replicated DNA chains in DNA polymerase I-deficient *Escherichia coli* mutants. *Proc. Natl Acad. Sci. USA*, **68**, 2954–2957.
10. Moldovan,G.-L., Madhavan,M.V., Mirchandani,K.D., McCaffrey,R.M., Vinciguerra,P. and D'Andrea,A.D. (2010) DNA polymerase POLN participates in cross-link repair and homologous recombination. *Mol. Cell. Biol.*, **30**, 1088–1096.
11. Hogg,M., Sauer-Eriksson,A.E. and Johansson,E. (2012) Promiscuous DNA synthesis by human DNA polymerase θ. *Nucleic Acids Res.*, **40**, 2611–2622.
12. Camps,M. and Loeb,L.A. (2004) When Pol I goes into high gear: processive DNA synthesis by Pol I in the cell. *Cell Cycle*, **3**, 114–116.
13. Hinkle,D.C. and Richardson,C.C. (1975) Bacteriophage T7 deoxyribonucleic acid replication in vitro. Purification and properties of the gene 4 protein of bacteriophage T7. *J. Biol. Chem.*, **250**, 5523–5529.
14. De Antoni,G.L., Besso,N.E., Zanassi,G.E., Sarachu,A.N. and Grau,O. (1985) Bacteriophage SPO1 DNA polymerase and the activity of viral gene 31. *Virology*, **143**, 16–22.
15. Rutberg,L., Råden,B. and Flock,J.I. (1981) Cloning and expression of bacteriophage SP02 DNAZ polymerase gene L in *Bacillus subtilis*, using the *Staphylococcus aureus* plasmid pC194. *J. Virol.*, **39**, 407–412.
16. Saiki,R.K., Gelfand,D.H., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. *Science*, **239**, 487–491.
17. Moser,M.J., DiFrancesco,R.A., Gowda,K., Klingele,A.J., Sugar,D.R., Stocki,S., Mead,D.A. and Schoenfeld,T.W. (2012) Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme. *PLoS One*, **7**, e38371.
18. Weissleder,R., Lee,H., Ko,J. and Pittet,M.J. (2020) COVID-19 diagnostics in context. *Sci. Transl. Med.*, **12**, eabc1931.
19. Pezo,V., Jaziri,F., Bourguignon,P.-Y., Louis,D., Jacobs-Sera,D., Rozenski,J., Pochet,S., Herdewijn,P., Hatfull,G.F., Kaminski,P.-A. *et al.* (2021) Noncanonical DNA polymerization by aminoadenine-based siphoviruses. *Science*, **372**, 520–524.
20. Czernecki,D., Hu,H., Romoli,F. and Delarue,M. (2021) Structural dynamics and determinants of 2-aminoadenine specificity in DNA polymerase DpoZ of vibriophage ϕVC8. *Nucleic Acids Res.*, **49**, 11974–11985.
21. Gomez-Raya-Vilanova,M.V., Leskinen,K., Bhattacharjee,A., Virta,P., Rosenqvist,P., Smith,J.L.R., Bayfield,O.W., Homberger,C., Kerrinnes,T., Vogel,J. *et al.* (2022) The DNA polymerase of bacteriophage YerA41 replicates its T-modified DNA in a primer-independent manner. *Nucleic Acids Res.*, **50**, 3985–3997.
22. Karam,J.D. and Konigsberg,W.H. (2000) DNA polymerase of the T4-related bacteriophages. *Prog. Nucleic Acid Res. Mol. Biol.*, **64**, 65–96.
23. Bebenek,A., Carver,G.T., Dressman,H.K., Kadyrov,F.A., Haseman,J.K., Petrov,V., Konigsberg,W.H., Karam,J.D. and Drake,J.W. (2002) Dissecting the fidelity of bacteriophage RB69 DNA polymerase: site-specific modulation of fidelity by polymerase accessory proteins. *Genetics*, **162**, 1003–1018.
24. Mönttinen,H.A.M., Ravantti,J.J. and Poranen,M.M. (2016) Common structural core of three-dozen residues reveals intersuperfamily relationships. *Mol. Biol. Evol.*, **33**, 1697–1710.
25. Filée,J., Forterre,P., Sen-Lin,T. and Laurent,J. (2002) Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. *J. Mol. Evol.*, **54**, 763–773.
26. Chan,Y.-W., Mohr,R., Millard,A.D., Holmes,A.B., Larkum,A.W., Whitworth,A.L., Mann,N.H., Scanlan,D.J., Hess,W.R. and Clokie,M.R.J. (2011) Discovery of cyanophage genomes which contain mitochondrial DNA polymerase. *Mol. Biol. Evol.*, **28**, 2269–2274.
27. Schoenfeld,T.W., Murugapiran,S.K., Dodsworth,J.A., Floyd,S., Lodes,M., Mead,D.A. and Hedlund,B.P. (2013) Lateral gene transfer of family A DNA polymerases between thermophilic viruses, aquificae, and apicomplexa. *Mol. Biol. Evol.*, **30**, 1653–1664.
28. Moriyama,T., Terasawa,K., Fujiwara,M. and Sato,N. (2008) Purification and characterization of organellar DNA polymerases in the red alga *Cyanidioschyzon merolae*. *FEBS J.*, **275**, 2899–2918.
29. Schmidt,H.F., Sakowski,E.G., Williamson,S.J., Polson,S.W. and Wommack,K. (2014) Shotgun metagenomics indicates novel family A DNA polymerases predominate within marine virioplankton. *ISME J.*, **8**, 103–114.
30. Nasko,D.J., Chopyk,J., Sakowski,E.G., Ferrell,B.D., Polson,S.W. and Wommack,K.E. (2018) Family A DNA polymerase phylogeny uncovers diversity and replication gene organization in the virioplankton. *Front. Microbiol.*, **9**, 3053.
31. Nuin,P.A., Wang,Z. and Tillier,E.R. (2006) The accuracy of several multiple sequence alignment programs for proteins. *BMC Bioinformatics*, **7**, 471.
32. Pervez,M.T., Babar,M.E., Nadeem,A., Aslam,M., Awan,A.R., Aslam,N., Hussain,T., Naveed,N., Qadri,S., Waheed,U. *et al.* (2014) Evaluating the accuracy and efficiency of multiple sequence alignment methods. *Evol. Bioinform. Online*, **10**, 205–217 .
33. Drake,J.W. (1999) The distribution of rates of spontaneous mutation over viruses, prokaryotes, and eukaryotes. *Ann. NY Acad. Sci.*, **870**, 100–107.
34. Fruchterman,T.M.J. and Reingold,E.M. (1991) Graph drawing by force-directed placement. *J. Software: Practice Experience*, **21**, 1129–1164.
35. Frickey,T. and Lupas,A. (2004) CLANS: a Java application for visualizing protein families based on pairwise similarity. *Bioinformatics*, **20**, 3702–3704.
36. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
37. Kazlauskas,D., Sezonov,G., Charpin,N., Venclovas,Č., Forterre,P. and Krupovic,M. (2018) Novel families of archaeo-eukaryotic

primases associated with mobile genetic elements of bacteria and archaea. *J. Mol. Biol.*, **430**, 737–750.
38. Kazlauskas,D., Krupovic,M., Guglielmini,J., Forterre,P. and Venclovas,Č. (2020) Diversity and evolution of B-family DNA polymerases. *Nucleic Acids Res.*, **48**, 10142–10156.
39. Jumper,J., Evans,R., Pritzel,A., Green,T., Figurnov,M., Ronneberger,O., Tunyasuvunakool,K., Bates,R., Žídek,A., Potapenko,A. *et al.* (2021) Highly accurate protein structure prediction with AlphaFold. *Nature*, **596**, 583–589.
40. Kryshtafovych,A., Schwede,T., Topf,M., Fidelis,K. and Moult,J. (2021) Critical assessment of methods of protein structure prediction (CASP)—Round XIV. *Proteins Struct. Funct. Bioinf.*, **89**, 1607–1617.
41. Finn,R.D., Bateman,A., Clements,J., Coggill,P., Eberhardt,R.Y., Eddy,S.R., Heger,A., Hetherington,K., Holm,L., Mistry,J. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
42. The UniProt Consortium (2021) UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.*, **49**, D480–D489.
43. Gabler,F., Nam,S.-Z., Till,S., Mirdita,M., Steinegger,M., Söding,J., Lupas,A.N. and Alva,V. (2020) Protein sequence analysis using the MPI bioinformatics toolkit. *Curr. Protoc. Bioinformatics*, **72**, e108.
44. Wheeler,D.L., Church,D.M., Federhen,S., Lash,A.E., Madden,T.L., Pontius,J.U., Schuler,G.D., Schriml,L.M., Sequeira,E., Tatusova,T.A. *et al.* (2003) Database resources of the National Center for Biotechnology. *Nucleic Acids Res.*, **31**, 28–33.
45. Madeira,F., Pearce,M., Tivey,A.R.N., Basutkar,P., Lee,J., Edbali,O., Madhusoodanan,N., Kolesnikov,A. and Lopez,R. (2022) Search and sequence analysis tools services from EMBL-EBI in 2022. *Nucleic Acids Res.*, **50** , W276–W279.
46. Crooks,G.E., Hon,G., Chandonia,J.-M. and Brenner,S.E. (2004) WebLogo: a sequence logo generator. *Genome Res.*, **14**, 1188–1190.
47. Robert,X. and Gouet,P. (2014) Deciphering key features in protein structures with the new ENDscript server. *Nucleic Acids Res.*, **42**, W320–W324.
48. Holm,L. (2019) Benchmarking fold detection by DaliLite v.5. *Bioinformatics*, **35**, 5326–5327.
49. Mirdita,M., Schütze,K., Moriwaki,Y., Heo,L., Ovchinnikov,S. and Steinegger,M. (2022) ColabFold: making protein folding accessible to all. *Nat. Methods*, **19**, 679–682.
50. Schrödinger (2023) The PyMOL Molecular Graphics System, Version 1.8 Schrödinger, LLC. https://pymol.org/2/.
51. Letunic,I. and Bork,P. (2021) Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. *Nucleic Acids Res.*, **49**, W293–W296.
52. Kumar,S., Stecher,G., Li,M., Knyaz,C. and Tamura,K. (2018) MEGA X: molecular evolutionary genetics analysis across computing platforms. *Mol. Biol. Evol.*, **35**, 1547–1549.
53. Lu,S., Wang,J., Chitsaz,F., Derbyshire,M.K., Geer,R.C., Gonzales,N.R., Gwadz,M., Hurwitz,D.I., Marchler,G.H., Song,J.S. *et al.* (2020) CDD/SPARCLE: the conserved domain database in 2020. *Nucleic Acids Res.*, **48**, D265–D268.
54. Takata,K., Reh,S., Yousefzadeh,M.J., Zelazowski,M.J., Bhetawal,S., Trono,D., Lowery,M.G., Sandoval,M., Takata,Y., Lu,Y. *et al.* (2017) Analysis of DNA polymerase *ν* function in meiotic recombination, immunoglobulin class-switching, and DNA damage tolerance. *PLoS Genet.*, **13**, e1006818.
55. Dion,M.B., Oechslin,F. and Moineau,S. (2020) Phage diversity, genomics and phylogeny. *Nat. Rev. Microbiol.*, **18**, 125–138.
56. Moreira,D. (2000) Multiple independent horizontal transfers of informational genes from bacteria to plasmids and phages: implications for the origin of bacterial replication machinery. *Mol. Microbiol.*, **35**, 1–5.
57. Delarue,M., Poch,O., Tordo,N., Moras,D. and Argos,P. (1990) An attempt to unify the structure of polymerases. *Protein Eng.,* **3**, 461–467.
58. Loh,E. and Loeb,L.A. (2005) Mutability of DNA polymerase I: implications for the creation of mutant DNA polymerases. *DNA Repair (Amst.)*, **4**, 1390–1398.
59. Tabor,S. and Richardson,C.C. (1995) A single residue in DNA polymerases of the *Escherichia coli* DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. *Proc. Natl Acad. Sci. USA*, **92**, 6339–6343.
60. Suzuki,M., Yoshida,S., Adman,E.T., Blank,A. and Loeb,L.A. (2000) *Thermus aquaticus* DNA polymerase I mutants with altered fidelity: interacting mutations in the O-helix. *J. Biol. Chem.*, **275**, 32728–32735.
61. Brutlag,D., Atkinson,M.R., Setlow,P. and Kornberg,A. (1969) An active fragment of DNA polymerase produced by proteolytic cleavage. *Biochem. Biophys. Res. Commun.*, **37**, 982–989.
62. Tindall,K.R. and Kunkel,T.A. (1988) Fidelity of DNA synthesis by the *Thermus aquaticus* DNA polymerase. *Biochemistry*, **27**, 6008–6013.
63. Aliotta,J.M., Pelletier,J.J., Ware,J.L., Moran,L.S., Benner,J.S. and Kong,H. (1996) Thermostable Bst DNA polymerase I lacks a 3′ → 5′ proofreading exonuclease activity. *Genet. Anal.*, **12**, 185–195.
64. Uphoff,S., Reyes-Lamothe,R., Leon,F.G.d., Sherratt,D.J. and Kapanidis,A.N. (2013) Single-molecule DNA repair in live bacteria. *Proc. Natl Acad. Sci. USA*, **110**, 8063–8068.
65. Joyce,C.M. and Grindley,N.D. (1984) Method for determining whether a gene of *Escherichia coli* is essential: application to the polA gene. *J. Bacteriol.*, **158**, 636–643.
66. Makiela-Dzbenska,K., Jaszczur,M., Banach-Orlowska,M., Jonczyk,P., Schaaper,R.M. and Fijalkowska,I.J. (2009) Role of *Escherichia coli* DNA polymerase I in chromosomal DNA replication fidelity. *Mol. Microbiol.*, **74**, 1114–1127.
67. Hernández-Tamayo,R., Oviedo-Bocanegra,L.M., Fritz,G. and Graumann,P.L. (2019) Symmetric activity of DNA polymerases at and recruitment of exonuclease ExoR and of PolA to the *Bacillus subtilis* replication forks. *Nucleic Acids Res.*, **47**, 8521–8536.
68. Fukushima,S., Itaya,M., Kato,H., Ogasawara,N. and Yoshikawa,H. (2007) Reassessment of the in vivo functions of DNA polymerase I and RNase H in bacterial cell growth. *J. Bacteriol.*, **189**, 8575–8583.
69. Fan,L., Sanschagrin,P.C., Kaguni,L.S. and Kuhn,L.A. (1999) The accessory subunit of mtDNA polymerase shares structural homology with aminoacyl-tRNA synthetases: implications for a dual role as a primer recognition factor and processivity clamp. *Proc. Natl Acad. Sci. USA*, **96**, 9527–9532.
70. Bolden,A., Noy,G.P. and Weissbach,A. (1977) DNA polymerase of mitochondria is a gamma-polymerase. *J. Biol. Chem.*, **252**, 3351–3356.
71. Iyengar,B., Luo,N., Farr,C.L., Kaguni,L.S. and Campos,A.R. (2002) The accessory subunit of DNA polymerase γ is essential for mitochondrial DNA maintenance and development in *Drosophila melanogaster*. *Proc. Natl Acad. Sci. USA*, **99**, 4483–4488.
72. Viikov,K., Väljamäe,P. and Sedman,J. (2011) Yeast mitochondrial DNA polymerase is a highly processive single-subunit enzyme. *Mitochondrion*, **11**, 119–126.
73. Lee,Y.-S., Kennedy,W.D. and Yin,Y.W. (2009) Structural insight into processive human mitochondrial DNA synthesis and disease-related polymerase mutations. *Cell*, **139**, 312–324.
74. Filée,J. and Forterre,P. (2005) Viral proteins functioning in organelles: a cryptic origin?*Trends Microbiol.*, **13**, 510–513.
75. Huang,T.-W. and Chen,C.W. (2008) DNA polymerase I is not required for replication of linear chromosomes in *Streptomyces*. *J. Bacteriol.*, **190**, 755–758.
76. Labonté,J.M., Reid,K.E. and Suttle,C.A. (2009) Phylogenetic analysis indicates evolutionary diversity and environmental segregation of marine podovirus DNA polymerase gene sequences. *Appl. Environ. Microbiol.*, **75**, 3634–3640.
77. Bedford,E., Tabor,S. and Richardson,C.C. (1997) The thioredoxin binding domain of bacteriophage T7 DNA polymerase confers processivity on *Escherichia coli* DNA polymerase I. *Proc. Natl Acad. Sci. USA*, **94**, 479–484.
78. Liu,B., Gu,S., Liang,N., Xiong,M., Xue,Q., Lu,S., Hu,F. and Zhang,H. (2016) *Pseudomonas aeruginosa* phage PaP1 DNA polymerase is an A-family DNA polymerase demonstrating ssDNA and dsDNA 3′–5′ exonuclease activity. *Virus Genes*, **52**, 538–551.
79. Wallen,J.R., Zhang,H., Weis,C., Cui,W., Foster,B.M., Ho,C.M.W., Hammel,M., Tainer,J.A., Gross,M.L. and Ellenberger,T. (2017) Hybrid methods reveal multiple flexibly linked DNA polymerases within the bacteriophage T7 replisome. *Structure*, **25**, 157–166.
80. Kulczyk,A.W., Moeller,A., Meyer,P., Sliz,P. and Richardson,C.C. (2017) Cryo-EM structure of the replisome reveals multiple interactions coordinating DNA synthesis. *Proc. Natl Acad. Sci.*, **114**, E1848–E1856.

81. Black,S.J., Ozdemir,A.Y., Kashkina,E., Kent,T., Rusanov,T., Ristic,D., Shin,Y., Suma,A., Hoang,T., Chandramouly,G. *et al.* (2019) Molecular basis of microhomology-mediated end-joining by purified full-length Polθ. *Nat. Commun.*, **10**, 4423.
82. Inagaki,S., Suzuki,T., Ohto,M., Urawa,H., Horiuchi,T., Nakamura,K. and Morikami,A. (2006) Arabidopsis TEBICHI, with gelicase and DNA polymerase domains, is required for regulated cell division and differentiation in meristems. *Plant Cell*, **18**, 879–892.
83. Fernandez-Vidal,A., Guitton-Sert,L., Cadoret,J.-C., Drac,M., Schwob,E., Baldacci,G., Cazaux,C. and Hoffmann,J.-S. (2014) A role for DNA polymerase θ in the timing of DNA replication. *Nat. Commun.*, **5**, 4285.
84. Zahn,K.E., Averill,A.M., Aller,P., Wood,R.D. and Doublié,S. (2015) Human DNA polymerase θ grasps the primer terminus to mediate DNA repair. *Nat. Struct. Mol. Biol.*, **22**, 304–311.
85. Newman,J.A., Cooper,C.D.O., Aitkenhead,H. and Gileadi,O. (2015) Structure of the helicase domain of DNA polymerase theta reveals a possible role in the microhomology-mediated end-joining pathway. *Structure*, **23**, 2319–2330.
86. Castroviejo,M., Tarragó-Litvak,L. and Litvak,S. (1975) Partial purification and characterization of two cytoplasmic DNA polymerases from ungerminated wheat. *Nucleic Acids Res.*, **2**, 2077–2090.
87. Christophe,L., Tarrago-Litvak,L., Castroviejo,M. and Litvak,S. (1981) Mitochondrial DNA polymerase from wheat embryos. *Plant Sci. Lett.*, **21**, 181–192.
88. Moriyama,T., Terasawa,K. and Sato,N. (2011) Conservation of POPs, the plant organellar DNA polymerases, in eukaryotes. *Protist*, **162**, 177–187.
89. Seow,F., Sato,S., Janssen,C.S., Riehle,M.O., Mukhopadhyay,A., Phillips,R.S., Wilson,R.J.M.(I.) and Barrett,M.P. (2005) The plastidic DNA replication enzyme complex of *Plasmodium falciparum*. *Mol. Biochem. Parasitol.*, **141**, 145–153.
90. Chang,J.R., Choi,J.J., Kim,H.-K. and Kwon,S.-T. (2001) Purification and properties of *Aquifex aeolicus* DNA polymerase expressed in *Escherichia coli*. *FEMS Microbiol. Lett.*, **201**, 73–77.
91. Palmer,M., Hedlund,B.P., Roux,S., Tsourkas,P.K., Doss,R.K., Stamereilers,C., Mehta,A., Dodsworth,J.A., Lodes,M., Monsma,S. *et al.* (2020) Diversity and distribution of a novel genus of hyperthermophilic aquificae viruses encoding a proof-reading family-A DNA polymerase. *Front. Microbiol.*, **11**, 2809.
92. Milton,M.E., Choe,J.-Y., Honzatko,R.B. and Nelson,S.W. (2016) Crystal structure of the apicoplast DNA polymerase from *Plasmodium falciparum*: the first look at a plastidic A-family DNA polymerase. *J. Mol. Biol.*, **428**, 3920–3934.
93. Lindner,S.E., Llinás,M., Keck,J.L. and Kappe,S.H.I. (2011) The primase domain of PfPrex is a proteolytically matured, essential enzyme of the apicoplast. *Mol. Biochem. Parasitol.*, **180**, 69–75.
94. Froman,S., Will,D.W. and Bogen,E. (1954) Bacteriophage active against virulent *Mycobacterium tuberculosis*—I. Isolation and activity. *Am. J. Public Health Nations Health*, **44**, 1326–1333.
95. Lohr,J.E., Chen,F. and Hill,R.T. (2005) Genomic analysis of bacteriophage ΦJL001: insights into its interaction with a sponge-associated alpha-proteobacterium. *Appl. Environ. Microbiol.*, **71**, 1598–1609.
96. Miyata,R., Yamaguchi,K., Uchiyama,J., Shigehisa,R., Takemura-Uchiyama,I., Kato,S., Ujihara,T., Sakaguchi,Y., Daibata,M. and Matsuzaki,S. (2014) Characterization of a novel *Pseudomonas aeruginosa* bacteriophage, KPP25, of the family Podoviridae. *Virus Res.*, **189**, 43–46.
97. Lavigne,R., Burkal'tseva,M.V., Robben,J., Sykilinda,N.N., Kurochkina,L.P., Grymonprez,B., Jonckx,B., Krylov,V.N., Mesyanzhinov,V.V. and Volckaert,G. (2003) The genome of bacteriophage φKMV, a T7-like virus infecting *Pseudomonas aeruginosa*. *Virology*, **312**, 49–59.
98. Magill,D.J., Kucher,P.A., Krylov,V.N., Pleteneva,E.A., Quinn,J.P. and Kulakov,L.A. (2017) Localised genetic heterogeneity provides a novel mode of evolution in dsDNA phages. *Sci. Rep.*, **7**, 13731.
99. Doublié,S., Tabor,S., Long,A.M., Richardson,C.C. and Ellenberger,T. (1998) Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature*, **391**, 251–258.
100. Bebel,A., Walsh,M.A., Mir-Sanchis,I. and Rice,P.A. (2020) A novel DNA primase–helicase pair encoded by SCCmec elements. *Elife*, **9**, e55478.
101. Takata,K., Arana,M.E., Seki,M., Kunkel,T.A. and Wood,R.D. (2010) Evolutionary conservation of residues in vertebrate DNA polymerase N conferring low fidelity and bypass activity. *Nucleic Acids Res.*, **38**, 3233–3244.
102. Pastor-Palacios,G., Azuara-Liceaga,E. and Brieba,L.G. (2010) A nuclear family A DNA polymerase from *Entamoeba histolytica* bypasses thymine glycol. *PLoS Negl. Trop. Dis.*, **4**, e786.
103. de Lima,L.P., Calderano,S.G., da Silva,M.S., de Araujo,C.B., Vasconcelos,E.J.R., Iwai,L.K., Pereira,C.A., Fragoso,S.P. and Elias,M.C. (2019) Ortholog of the polymerase theta helicase domain modulates DNA replication in *Trypanosoma cruzi*. *Sci. Rep.*, **9**, 2888.
104. Lee,Y.-S., Gao,Y. and Yang,W. (2015) How a homolog of high-fidelity replicases conducts mutagenic DNA synthesis. *Nat. Struct. Mol. Biol.*, **22**, 298–303.
105. Pearl,L.H. (2000) Structure and function in the uracil-DNA glycosylase superfamily. *Mutati. Res.*, **460**, 165–181.
106. Schito,G.C., Rialdi,G. and Pesce,A. (1966) Biophysical properties of N4 coliphage. *Biochim. Biophys. Acta*, **129**, 482–490.
107. Taylor,M.J. and Thorne,C.B. (1963) Transduction of *Bacillus licheniformis* and *Bacillus subtilis* by each of two phages. *J. Bacteriol.*, **86**, 452–461.
108. Okubo,S., Strauss,B. and Stodolsky,M. (1964) The possible role of recombination in the infection of competent *Bacillus subtilis* by bacteriophage deoxyribonucleic acid. *Virology*, **24**, 552–562.
109. Brandon,C., Gallop,P.M., Marmur,J., Hayashi,H. and Nakanishi,K. (1972) Structure of a new pyrimidine from *Bacillus subtilis* phage SP-15 nucleic scid. *Nat. New Biol.*, **239**, 70–71.
110. Witmer,H. and Dosmar,M. (1978) Synthesis of 5-hydroxymethyldeoxyuridine triphosphate in extracts of SP10c phage-infected *Bacillus subtilis* W23. *Curr. Microbiol.*, **1**, 289–292.
111. Stewart,C.R., Casjens,S.R., Cresawn,S.G., Houtz,J.M., Smith,A.L., Ford,M.E., Peebles,C.L., Hatfull,G.F., Hendrix,R.W., Huang,W.M. *et al.* (2009) The genome of *Bacillus subtilis* bacteriophage SPO1. *J. Mol. Biol.*, **388**, 48–70.
112. Witmer,H. (1981) Synthesis of deoxythymidylate and the unusual deoxynucleotide in mature DNA of *Bacillus subtilis* bacteriophage SP10 occurs by postreplicational modification of 5-hydroxymethyldeoxyuridylate. *J. Virol.*, **39**, 536–547.
113. Zhou,Y., Xu,X., Wei,Y., Cheng,Y., Guo,Y., Khudyakov,I., Liu,F., He,P., Song,Z., Li,Z. *et al.* (2021) A widespread pathway for substitution of adenine by diaminopurine in phage genomes. *Science*, **372**, 512–516.
114. Kirnos,M.D., Khudyakov,I.Y., Alexandrushkina,N.I. and Vanyushin,B.F.(1977)2-Aminoadenine is an adenine substituting for a base in S-2L cyanophage DNA. *Nature*, **270**, 369.
115. Czernecki,D., Legrand,P., Tekpinar,M., Rosario,S., Kaminski,P.-A. and Delarue,M. (2021) How cyanophage S-2L rejects adenine and incorporates 2-aminoadenine to saturate hydrogen bonding in its DNA. *Nat. Commun.*, **12**, 2420.
116. Czernecki,D., Bonhomme,F., Kaminski,P.-A. and Delarue,M. (2021) Characterization of a triad of genes in cyanophage S-2L sufficient to replace adenine by 2-aminoadenine in bacterial DNA. *Nat. Commun.*, **12**, 4710.
117. Astatke,M., Ng,K., Grindley,N.D.F. and Joyce,C.M. (1998) A single side chain prevents *Escherichia coli* DNA polymerase I (Klenow fragment) from incorporating ribonucleotides. *Proc. Natl Acad. Sci. USA*, **95**, 3402–3407.
118. Brown,J.A. and Suo,Z. (2011) Unlocking the sugar 'steric gate' of DNA polymerases. *Biochemistry*, **50**, 1135–1142.
119. Suzuki,M., Baskin,D., Hood,L. and Loeb,L.A. (1996) Random mutagenesis of *Thermus aquaticus* DNA polymerase I: concordance of immutable sites in vivo with the crystal structure. *Proc. Natl Acad. Sci. USA*, **93**, 9670–9675.
120. Juarez-Quintero,V., Peralta-Castro,A., Benítez Cardoza,C.G., Ellenberger,T. and Brieba,L.G. (2021) Structure of an open conformation of T7 DNA polymerase reveals novel structural features regulating primer–template stabilization at the polymerization active site. *Biochem. J.*, **478**, 2665–2679.
121. Ahn,W.-C., Aroli,S., Kim,J.-H., Moon,J.H., Lee,G.S., Lee,M.-H., Sang,P.B., Oh,B.-H., Varshney,U. and Woo,E.-J. (2019) Covalent binding of uracil DNA glycosylase UdgX to abasic DNA upon uracil excision. *Nat. Chem. Biol.*, **15**, 607–614.

122. Xu,J., Hendrix,R.W. and Duda,R.L. (2004) Conserved translational frameshift in dsDNA bacteriophage tail assembly genes. *Mol. Cell*, **16**, 11–21.
123. Huang,S., Sun,Y., Zhang,S. and Long,L. (2021) Temporal transcriptomes of a marine cyanopodovirus and its *Synechococcus* host during infection. *MicrobiologyOpen*, **10**, e1150.
124. Wang,J. and Konigsberg,W.H. (2022) Two-metal-ion catalysis: inhibition of DNA polymerase activity by a third divalent metal ion. *Front. Mol. Biosci.*, **9**, 824794.
125. Rozovskaya,T.A., Rechinsky,V.O., Bibilashvili,R.S., Karpiesky,MYa., Tarusova,N.B., Khomutov,R.M. and Dixon,H.B.F. (1984) The mechanism of pyrophosphorolysis of RNA by RNA polymerase. Endowment of RNA polymerase with artificial exonuclease activity. *Biochem. J.*, **224**, 645–650.
126. Shock,D.D., Freudenthal,B.D., Beard,W.A. and Wilson,S.H. (2017) Modulating the DNA polymerase β reaction equilibrium to dissect the reverse reaction. *Nat. Chem. Biol.*, **13**, 1074–1080.
127. Iyer,L.M., Koonin,E.V., Leipe,D.D. and Aravind,L. (2005) Origin and evolution of the archaeo-eukaryotic primase superfamily and related palm-domain proteins: structural insights and new members. *Nucleic Acids Res.*, **33**, 3875–3896.
128. Harada,R. and Inagaki,Y. (2021) Phage origin of mitochondrion-localized family A DNA polymerases in kinetoplastids and diplonemids. *Genome Biol. Evol.*, **13**, evab003.
129. Leavitt,M.C. and Ito,J. (1989) T5 DNA polymerase: structural–functional relationships to other DNA polymerases. *Proc. Natl Acad. Sci. USA*, **86**, 4465–4469.
130. Derbyshire,V., Astatke,M. and Joyce,C.M. (1993) Re-engineering the polymerase domain of Klenow fragment and evaluation of overproduction and purification strategies. *Nucleic Acids Res.*, **21**, 5439–5448.
131. Hadrawi,W.H., Norazman,A., Mohd Shariff,F., Mohamad Ali,M.S. and Raja Abd Rahman,R.N.Z. (2020) Understanding the effect of multiple domain deletion in DNA polymerase I from *Geobacillus* sp. strain SK72. *Catalysts*, **10**, 936.
132. Raia,P., Carroni,M., Henry,E., Pehau-Arnaudet,G., Brûlé,S., Béguin,P., Henneke,G., Lindahl,E., Delarue,M. and Sauguet,L. (2019) Structure of the DP1–DP2 PolD complex bound with DNA and its implications for the evolutionary history of DNA and RNA polymerases. *PLoS Biol.*, **17**, e3000122.
133. Forrest,D., James,K., Yuzenkova,Y. and Zenkin,N. (2017) Single-peptide DNA-dependent RNA polymerase homologous to multi-subunit RNA polymerase. *Nat. Commun.*, **8**, 15774.
134. Wu,E.Y. and Beese,L.S. (2011) The structure of a high fidelity DNA polymerase bound to a mismatched nucleotide reveals an 'ajar' intermediate conformation in the nucleotide selection mechanism. *J. Biol. Chem.*, **286**, 19758–19767.
135. Pál,C., Papp,B. and Lercher,M.J. (2006) An integrated view of protein evolution. *Nat. Rev. Genet.*, **7**, 337–348.
136. Nudler,E. (2012) RNA polymerase backtracking in gene regulation and genome instability. *Cell*, **149**, 1438–1445.
137. Singh,A., Pandey,M., Nandakumar,D., Raney,K.D., Yin,Y.W. and Patel,S.S. (2020) Excessive excision of correct nucleotides during DNA synthesis explained by replication hurdles. *EMBO J.*, **39**, e103367.
138. Singh,K. and Modak,M.J. (2005) Contribution of polar residues of the J-helix in the 3′–5′ exonuclease activity of *Escherichia coli* DNA polymerase I (Klenow fragment): Q677 regulates the removal of terminal mismatch. *Biochemistry*, **44**, 8101–8110.
139. Yamagami,T., Matsukawa,H., Tsunekawa,S., Kawarabayasi,Y., Ishino,S. and Ishino,Y. (2016) A longer finger-subdomain of family A DNA polymerases found by metagenomic analysis strengthens DNA binding and primer extension abilities. *Gene*, **576**, 690–695.
140. Blanco,L., Bernad,A., Blasco,M.A. and Salas,M. (1991) A general structure for DNA-dependent DNA polymerases. *Gene*, **100**, 27–38.
141. Méndez,J., Blanco,L., Lázaro,J.M. and Salas,M. (1994) Primer-terminus stabilization at the psi 29 DNA polymerase active site. Mutational analysis of conserved motif TX2GR. *J. Biol. Chem.*, **269**, 30030–30038.

2 Conclusion

Au cours de ces travaux, les ADN polymérases A de *Streptomyces* sp. CT34 et *Ideonella dechloratans* ont été utilisées comme références pour étudier les nouveaux clusters APEX et CHEAP. Après leur production et leur purification, leurs activités ont été testées *in vitro*.

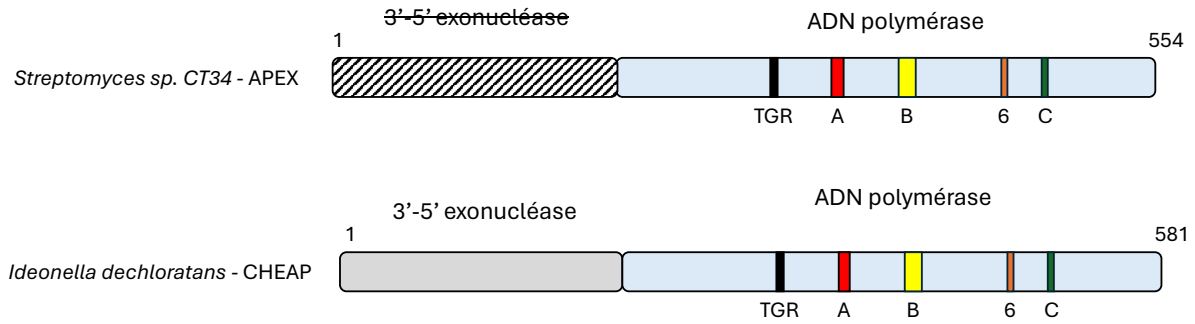


Figure 182 : Schéma global des ADN polymérases de la famille A étudiées dans ces travaux. Les deux enzymes portent un domaine exonucléase, mais pour l'ADN polymérase du cluster APEX, ce domaine est inactivé. Les motifs caractéristiques impliqués dans l'activité ADN polymérase de l'enzyme sont indiqués.

L'ADN polymérase du cluster APEX présente un domaine 3'-5' exonucléase inactivé, et n'a donc pas d'activité de correction. Les tests fonctionnels ont montré que cette enzyme nécessite des ions Mn^{2+} pour être pleinement active, et qu'elle n'est pas processive mais distributive. Ces résultats indiquent que cette ADN polymérase est impliquée *in vivo* dans la réparation de l'ADN : son contexte exact reste cependant à découvrir. Ces conclusions ont été étendues à l'ensemble du cluster correspondant à cette polymérase. L'ADN polymérase du cluster CHEAP a montré une forte processivité (meilleure que l'ADN polymérase I de *E. coli*), une activité 3'-5' exonucléase, et une dépendance aux ions magnésium : ces éléments tendent à la catégoriser parmi les ADN polymérases réplicatives, tout comme le reste des séquences de son cluster.

Cependant, d'autres ADN polymérases A restent non caractérisées à ce jour, et devront être l'objet de futurs travaux, comme celles du cluster 10.

Annexes

1 Clonages des gènes et productions et purifications des protéines

1.1 Vecteur plasmidique utilisé

Le plasmide utilisé dans tous ces travaux est une construction optimisée basée sur le plasmide pRSF-Duet appelée LS05.

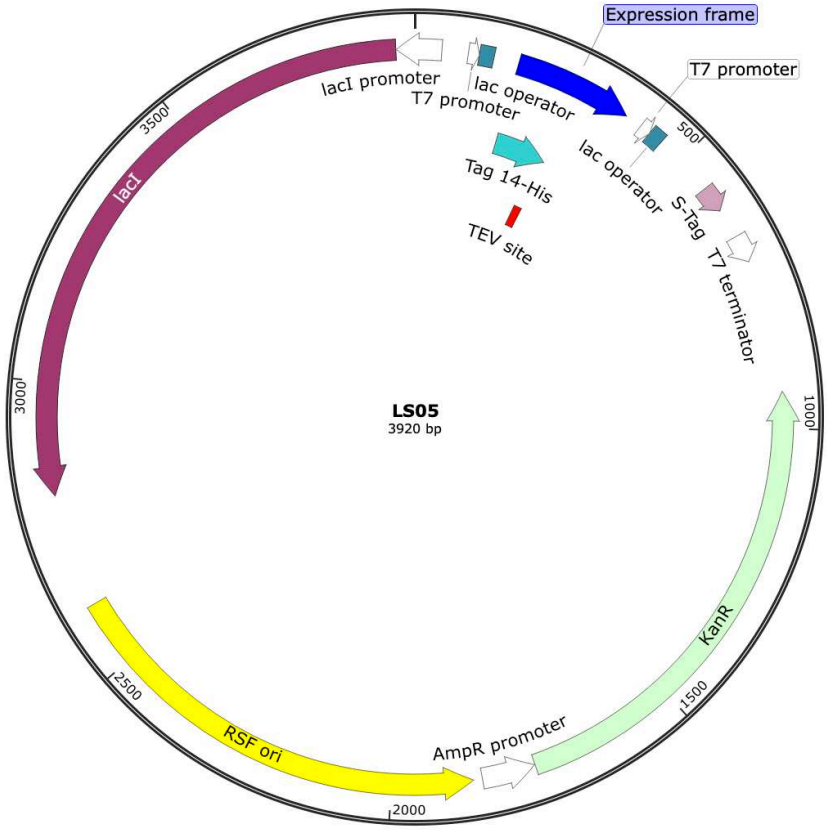


Figure 183 : Carte du vecteur plasmidique utilisé.

Ce plasmide a une origine de réplication RSF, qui lui permet d'être répliqué en un grand nombre de copies dans les bactéries. Il porte aussi un gène de résistance à la kanamycine, qui permet une sélection des bactéries transformées grâce à cet antibiotique. Deux sites multiples de clonage sont présents dans ce plasmide, et précédés de promoteurs de l'ARN polymérase du phage T7. Le premier, et le seul utilisé dans ces travaux, permet l'expression d'un gène fusionné à un tag de 14 histidines, avec un site de clivage par la protéase du virus de la gravure du tabac (ou TEV pour *Tobacco Etch Virus*), qui permet de séparer la protéine de son tag. Enfin, ce plasmide porte aussi le gène lacI, qui permet de produire le répresseur de l'opéron lactose. Celui-ci, en absence d'inducteur (lactose ou isopropyl β-D-1-thiogalactopyranoside (IPTG),

analogue utilisé ici), se fixe sur l'opérateur lac, ce qui bloque la transcription du gène qu'il précède par l'ARN polymérase du phage T7. Cette répression est levée lors de l'ajout d'IPTG dans le milieu de culture.

1.2 Souches bactériennes utilisées

1.2.1 E. coli DH5α

La première souche bactérienne chimiocompétente utilisée pour surproduire les plasmides était la souche DH5α. Celle-ci a plusieurs particularités, résumées dans son génotype (les éléments d'importance dans le cadre de ces travaux d'obtention de plasmides portant des gènes d'intérêt sont indiqués en **gras**):

F– φ80lacZΔM15 Δ(lacZYA-argF)U169 ***recA1 endA1 hsdR17(rK–, mK+)*** *phoA supE44 λ–thi-1 gyrA96 relA1*

- F- : elle ne porte pas de facteur de conjugaison et ne peut donc pas transmettre de matériel génétique à ses voisines ;
- φ80lacZΔM15 : délétion partielle du gène lacZ utile pour l'a-complémentation. Utile pour les méthodes de criblage blanc/bleu ;
- Δ(lacZYA-argF)U169 : amélioration de la résistance au peroxyde d'hydrogène ;
- **recA1** : inactivation de la recombinase RecA, ce qui limite la recombinaison chez les bactéries entre les plasmides et le génome ;
- **endA1** : inactivation de l'endonucléase endA, ce qui limite la dégradation de l'ADN par la bactérie ;
- **hsdR17(rK–, mK+)** : la bactérie n'a plus de système de restriction et de modification de l'ADN, et ne peut donc pas le modifier ou le dégrader. Elle peut cependant méthyler l'ADN ;
- phoA : suppression de l'activité de la phosphatase alcaline ;
- supE44 : suppression du codon stop « ambre » (UAG) ;
- λ–thi-1 : mutation du métabolisme de la thiamine ;
- gyrA96 : mutation d'une ADN gyrase permettant la résistance à l'acide nalidixique ;
- relA1 : permet la synthèse d'ARN en absence de synthèse de protéines.

1.2.2 E. coli Top10

La seconde souche bactérienne chimiocompétente utilisée était la souche Top10. Son génotype est le suivant :

F- ***mcrA Δ(mrr-hsdRMS-mcrBC)*** *Φ80lacZΔM15 Δ lacX74* ***recA1*** *araD139 Δ(araleu) 7697 galU galK rpsL (StrR)* ***endA1 nupG***

- F- : elle ne porte pas de facteur de conjugaison et ne peut donc pas transmettre de matériel génétique à ses voisines ;
- **mcrA Δ(mrr-hsdRMS-mcrBC)** : suppression d'un système de restriction de l'ADN ;
- φ80lacZΔM15 : délétion partielle du gène lacZ utile pour l'a-complémentation. Utile pour les méthodes de criblage blanc/bleu ;
- Δ lacX74 : suppression de l'opéron lactose ;
- **recA1** : inactivation de la recombinase RecA, ce qui limite la recombinaison chez les bactéries entre les plasmides et le génome ;
- araD139 : blocage du métabolisme de l'arabinose ;
- Δ(araleu) 7697 : les bactéries ne peuvent pas synthétiser de leucine, et donc ne peuvent pas pousser sur milieu minimum ;
- galU galK : mutations du métabolisme du galactose ;
- rpsL (StrR) : résistance à la streptomycine ;
- **endA1** : inactivation de l'endonucléase endA, ce qui limite la dégradation de l'ADN par la bactérie ;
- **nupG** : permet l'expression constitutive de gènes permettant l'expression de desoxyribose, ce qui permet à la bactérie d'accepter de plus gros plasmides.

1.2.3 E. coli BL21star(DE3)

La souche bactérienne utilisée pour exprimer les protéines étudiées était la souche BL21star(DE3). Celle-ci a plusieurs particularités, résumées dans son génotype (les éléments d'importance pour cette étape d'expression des protéines sont indiqués en **gras**):

F^-**ompT hsdS$_B$(r_B^- m_B^-)** gal dcm **rne131 (DE3)**

- F- : elle ne porte pas de facteur de conjugaison et ne peut donc pas transmettre de matériel génétique à ses voisines ;
- **ompT** : la bactérie porte une version mutante d'une protéase, ce qui limite la dégradation des protéines produites ;
- **hsdSB (rB-mB-)** : la bactérie n'a plus de système de méthylation ou de restriction de l'ADN, et ne peut donc pas le modifier ;
- gal : le métabolisme du galactose est inefficace chez ces bactéries, qui ne peuvent donc pas se développer en présente de galactose uniquement ;
- dcm : cette mutation empêche la bactérie de méthyler les cytosines ;
- **rne131** : une mutation dans la RNAse E limite la dégradation des ARN, ce qui augmente leur demi-vie et donc la synthèse des protéines ;
- **DE3** : la bactérie porte sous contrôle du promoteur lacUV5 (réprimé en absence de lactose ou analogue par le répresseur Lac) le gène de l'ARN polymérase du phage T7, ce qui permet l'expression des gènes sous le promoteur de cette polymérase (ce qui est le cas avec le vecteur LS05).

1.3 Protocole de transformation de bactéries chimiocompétentes

Pour obtenir des colonies bactériennes portant les plasmides produits, le protocole de transformation bactérienne suivant a été utilisé (en conditions stériles):

- Décongeler un aliquot de 100 µL de bactéries compétentes, dans de la glace, pendant 15 min.
- Y ajouter 3 µL de solution de plasmide, et mélanger délicatement.
- Incuber à 4°C pendant 30 min.
- Placer au bain marie à 42°C pendant 30 secondes.
- Incuber à 4°C pendant 2 min.
- Ajouter 1 mL de milieu SOC.
- Incuber 1h à 37°C.
- Étaler 100 µL de la culture sur une boite de petri contenant du milieu LB gélosé et de la kanamycine (ou autre antibiotique adapté).
- Incuber une nuit à 37°C.

Les colonies obtenus portent le plasmide utilisé.

1.4 Techniques de clonage moléculaire utilisées

1.4.1 STRU-cloning

Dans cette méthode (Bellini *et al.*, 2011), 2 µg de plasmide donneur (résistant à l'ampicilline) sont mélangés avec 1 µg de plasmide receveur LS05 (résistant à la kanamycine) et les enzymes de restriction choisies et leur tampon. L'ensemble est incubé dans les conditions optimales (durée et température) d'activité de ces enzymes. Le mélange est ensuite chauffé à 70°C pendant 10 min pour inactiver les enzymes, puis transféré dans un tube concentrateur (Vivaspin) avec une haute limite de poids moléculaire (100 kDa). La centrifugation permet de ne conserver que les plasmides linéarisés et les gènes d'intérêt, en se débarrassant de la séquence située entre les sites de restriction sur le plasmide receveur. Après centrifugation, le volume n'ayant pas traversé la membrane du concentrateur (contenant l'ADN) est récupéré, et transféré dans une réaction de ligation avec l'ADN ligase T4 en présence d'ATP. Après incubation à température ambiante pendant 30 min, des bactéries *E. coli* Top10 sont transformées avec le produit de ligation. Ces bactéries sont alors cultivées en milieu LB gélosé en présence de kanamycine, une nuit à 37°C. Ainsi, si elles ont été transformées avec le plasmide recircularisé par la ligation avec le gène d'intérêt, elles pourront pousser. Si la ligation n'a pas eu lieu, le plasmide reste linéaire et les bactéries ne peuvent théoriquement pas l'utiliser résister à la kanamycine. Si elles ont été transformées avec le plasmide donneur qui n'aurait pas été coupé par les enzymes de restriction, elles sont résistantes à l'ampicilline et non la kanamycine, et ne peuvent donc pas pousser. Cette sélection garantit donc en théorie que les seules colonies bactériennes sont celles issues de bactéries ayant intégré le plasmide souhaité : LS05 avec le gène d'intérêt inséré.

1.4.2 Gibson assembly

Des amorces de PCR sont créées et utilisées en PCR avec un gradient de températures d'hybridation en présence d'ADN polymérase Q5 Hot Start High Fidelity 1X (New England Biolabs), avec ou sans ajout de *GC enhancer* (un mélange permettant d'optimiser la PCR en cas de taux de GC élevé, composé entre autres d'un mélange de glycérol, de diméthylsulfoxyde et de tétraméthylammonium, trois composants réduisant la formation de structures secondaires dans l'ADN et facilitant la fixation des amorces). Cette PCR a pour but d'amplifier le gène depuis le plasmide donneur d'une part, et de linéariser le plasmide receveur d'autre part. Les amorces sont utilisées à une concentration de 500 nM, et les extrémités des amorces doivent

présenter des débordements correspondant aux séquences aux extrémités du site de clonage du plasmide (pour les amorces permettant d'amplifier le gène d'intérêt), et aux extrémités du gène d'intérêt (pour celles permettant d'amplifier le plasmide receveur). Chaque réaction est incubée 3 min à 98°C, puis suit 30 cycles (10 secondes 98°C, 30 sec à différentes températures testées en gradient entre 55, 30 sec/kb 72°C) et termine l'incubation par 2 min d'élongation à 72°C. La taille des ADN amplifiés par PCR est ensuite contrôlée par électrophorèse sur gel d'agarose 1% comme indiqué précédemment. S'ils ont la bonne taille, les deux ADN linéaires (gène et plasmide) sont mélangés, à hauteur de 60 fmoles de vecteur pour 120 fmoles d'insert, avec un mélange d'enzymes et leur tampon (NEBuilder HiFi DNA Assembly, New England Biolabs), et incubés à 50°C pendant 15 minutes. Ce mélange contient une 5' exonucléase, qui dégrade les extrémités 5' de l'ADN pour révéler des extrémités 3' débordantes. Celles-ci peuvent donc s'hybrider les unes aux autres, puisque les extrémités des ADN amplifiés contiennent des extensions compatibles, et une ADN polymérase étend les extrémités 3' hybridées, jusqu'à rattraper l'exonucléase et la décrocher de l'ADN. Enfin, une ADN ligase rétablit les liaisons phosphodiester : le gène est cloné dans le vecteur.

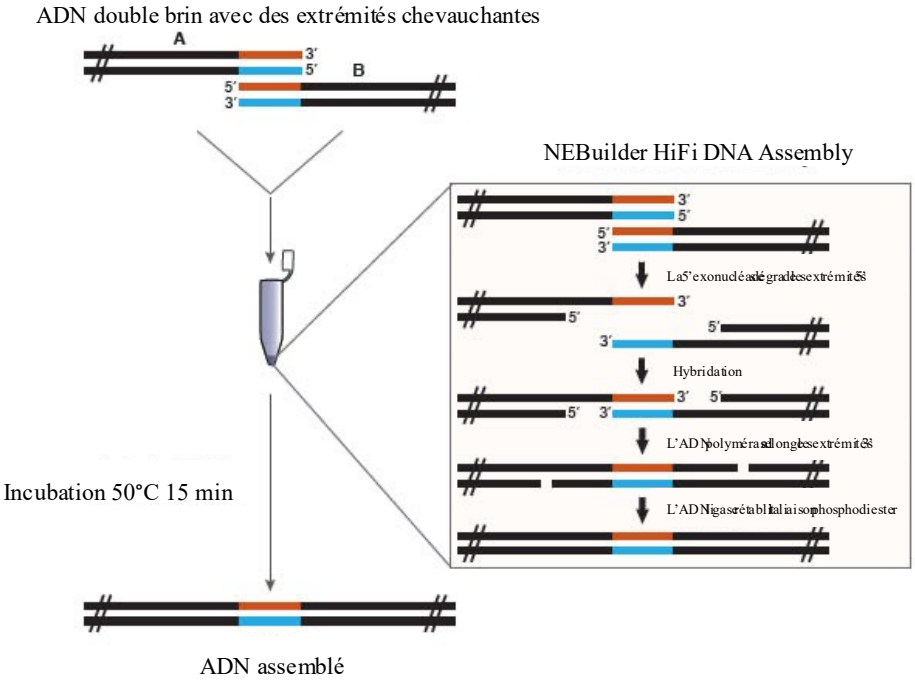


Figure 184 : Principe de la méthode NEBuilder HiFi DNA Assembly. D'après https://www.neb-online.fr/cloning-synthetic-biology/dna-assembly/nebuilder-hifi-dna-assembly/

1.4.3 Règles pour la préparation des amorces de PCR

La température d'hybridation des amorces doit être d'environ 60°C, leur taux de nucléotides G et C doit se situer entre 40 et 60%, et leur extrémité 3' doit être composée d'au moins un nucléotide G ou C, de façon à stabiliser l'hybridation avec l'ADN plasmidique.

1.5 Techniques de mutagenèse dirigée

1.5.1 Les délétions de domaines ou de séquences

Pour supprimer une partie d'une séquence, une PCR est réalisée avec l'ADN polymérase Q5, avec des amorces s'hybridant à chaque extrémité de la séquence à éliminer au sein du plasmide, avec les extrémités 3' orientées vers la séquence à conserver. Cela permet d'amplifier la séquence, sauf la partie située entre les deux amorces. Une fois la PCR faite, la taille des fragments amplifiés est contrôlée par électrophorèse en gel d'agarose 1%. Si le plasmide linéarisé a la taille attendue, il est mélangé à un ensemble d'enzymes appelé KLD (pour Kinase, Ligase, DpnI, (New England Biolabs)) et incubé 5 min à température ambiante. Plusieurs réactions enzymatiques ont lieu à cette étape, pour permettre la circularisation du plasmide : une phosphorylation des extrémités de l'ADN, réalisée par la kinase, puis une ligation, qui permet de lier les extrémités d'ADN phosphorylées, et une dégradation de l'ADN parental par l'enzyme de restriction DpnI, qui reconnait les sites GATC méthylés. Ces méthylations étant réalisées par les bactéries au moment de la surproduction des plasmides, l'ADN néosynthétisé lors de la PCR n'est pas méthylé et n'est donc pas dégradé par DpnI.

1.5.2 Les mutations ponctuelles

Pour introduire des mutations ponctuelles, une méthode proche de la précédente est utilisée. Ici, les deux amorces utilisées sont « dos à dos », et l'une des deux intègre la mutation à réaliser. Le choix des codons à utiliser est fait en se référant aux tables d'usage des codons chez *E. coli*, afin d'éviter d'utiliser un codon rare. Dans certains cas, des optimisations des PCR doivent être réalisées, en faisant varier différents paramètres, le plus souvent la quantité d'ADN initiale.

Tableau 14 : Usage des codons chez E.coli. [2] : lettre entre parenthèse : code 1 lettre pour l'acide aminé ; [3] : fréquence moyenne du codon parmi 100 codons ; [4] : abondance du codon par rapport aux autres codons codant pour cet acide aminé. D'après https://2014.igem.org/Team:Penn_State/CodonOptimization

	Codon	Amino acid[2]	%[3]	Ratio[4]	Codon	Amino acid	%	Ratio	Codon	Amino acid	%	Ratio	Codon	Amino acid	%	Ratio	
U	**UUU**	**Phe (F)**	1.9	0.51	**UCU**	**Ser (S)**	1.1	0.19	**UAU**	**Tyr (Y)**	1.6	0.53	**UGU**	**Cys (C)**	0.4	0.43	**U**
	UUC	**Phe (F)**	1.8	0.49	**UCC**	**Ser (S)**	1.0	0.17	**UAC**	**Tyr (Y)**	1.4	0.47	**UGC**	**Cys (C)**	0.6	0.57	**C**
	UUA	**Leu (L)**	1.0	0.11	**UCA**	**Ser (S)**	0.7	0.12	**UAA**	**STOP**	0.2	0.62	**UGA**	**STOP**	0.1	0.30	**A**
	UUG	**Leu (L)**	1.1	0.11	**UCG**	**Ser (S)**	0.8	0.13	**UAG**	**STOP**	0.03	0.09	**UGG**	**Trp (W)**	1.4	1.00	**G**
C	**CUU**	**Leu (L)**	1.0	0.10	**CCU**	**Pro (P)**	0.7	0.16	**CAU**	**His (H)**	1.2	0.52	**CGU**	**Arg (R)**	2.4	0.42	**U**
	CUC	**Leu (L)**	0.9	0.10	**CCC**	**Pro (P)**	0.4	0.10	**CAC**	**His (H)**	1.1	0.48	**CGC**	**Arg (R)**	2.2	0.37	**C**
	CUA	**Leu (L)**	0.3	0.03	**CCA**	**Pro (P)**	0.8	0.20	**CAA**	**Gln (Q)**	1.3	0.31	**CGA**	**Arg (R)**	0.3	0.05	**A**
	CUG	**Leu (L)**	5.2	0.55	**CCG**	**Pro (P)**	2.4	0.55	**CAG**	**Gln (Q)**	2.9	0.69	**CGG**	**Arg (R)**	0.5	0.08	**G**
A	**AUU**	**Ile (I)**	2.7	0.47	**ACU**	**Thr (T)**	1.2	0.21	**AAU**	**Asn (N)**	1.6	0.39	**AGU**	**Ser (S)**	0.7	0.13	**U**
	AUC	**Ile (I)**	2.7	0.46	**ACC**	**Thr (T)**	2.4	0.43	**AAC**	**Asn (N)**	2.6	0.61	**AGC**	**Ser (S)**	1.5	0.27	**C**
	AUA	**Ile (I)**	0.4	0.07	**ACA**	**Thr (T)**	0.1	0.30	**AAA**	**Lys (K)**	3.8	0.76	**AGA**	**Arg (R)**	0.2	0.04	**A**
	AUG	**Met (M)**	2.6	1.00	**ACG**	**Thr (T)**	1.3	0.23	**AAG**	**Lys (K)**	1.2	0.24	**AGG**	**Arg (R)**	0.2	0.03	**G**
G	**GUU**	**Val (V)**	2.0	0.29	**GCU**	**Ala (A)**	1.8	0.19	**GAU**	**Asp (D)**	3.3	0.59	**GGU**	**Gly (G)**	2.8	0.38	**U**
	GUC	**Val (V)**	1.4	0.20	**GCC**	**Ala (A)**	2.3	0.25	**GAC**	**Asp (D)**	2.3	0.41	**GGC**	**Gly (G)**	3.0	0.40	**C**
	GUA	**Val (V)**	1.2	0.17	**GCA**	**Ala (A)**	2.1	0.22	**GAA**	**Glu (E)**	4.4	0.70	**GGA**	**Gly (G)**	0.7	0.09	**A**
	GUG	**Val (V)**	2.4	0.34	**GCG**	**Ala (A)**	3.2	0.34	**GAG**	**Glu (E)**	1.9	0.30	**GGG**	**Gly (G)**	0.9	0.13	**G**
	U				**C**				**A**				**G**				

Une fois la PCR réalisée, le protocole était le même que pour la délétion : vérification de la taille du plasmide par électrophorèse en gel d'agarose 1%, puis utilisation du mélange KLD.

1.6 Tampons utilisés lors des purifications

Tampon	ADN Polymérases X de *Paramecium tetraurelia* (et versions mutées)	ADN polymérase λ humaine (et versions mutées)	ADN polymérase β humaine
A	50 mM Tris-HCl pH 8, 600 mM NaCl, 10 mM imidazole	50 mM Tris-HCl pH 8, 1 mM EDTA, 1 mM DTT, 5% glycérol, 500 mM NaCl, 20 mM imidazole	50 mM Tris-HCl pH8, 500 mM NaCl, 20 mM imidazole
Lavage	Uniquement pour PolXbΔBRCT : Sodium Phosphate 25 mM pH8, NaCl 1M		
B	50 mM Tris-HCl pH 8, 600 mM NaCl, 500 mM imidazole	50 mM Tris-HCl pH 8, 1 mM EDTA, 1 mM DTT, 5% glycérol, 500 mM NaCl, 500 mM imidazole	50 mM Tris-HCl pH8, 500 mM NaCl, 500 mM imidazole
C	50 mM Tris-HCl pH 8	50 mM Tris-HCl pH 8, 1 mM EDTA, 1 mM DTT, 5% glycérol	50 mM Tris-HCl pH8
D	50 mM Tris-HCl pH 8, 100 mM NaCl	50 mM Tris-HCl pH 8, 1 mM EDTA, 1 mM DTT, 5% glycérol, 100 mM NaCl	50 mM Tris-HCl pH8, 100 mM NaCl,
E	50 mM Tris-HCl pH 8, 1 M NaCl	50 mM Tris-HCl pH 8, 1 mM EDTA, 1 mM DTT, 5% glycérol, 1 M NaCl	50 mM Tris-HCl pH8, 1 M NaCl
F	50 mM Tris-HCl pH, 300 mM NaCl	50 mM Tris-HCl pH 8, 100 mM NaCl	50 mM Tris-HCl pH8, 500 mM NaCl

1.7 Techniques utilisées lors des purifications

1.7.1 Sonication

La sonication permet de lyser les cellules en utilisant des ultrasons dans une solution, ce qui provoque la formation de bulles qui implosent par cavitation. C'est cette implosion des qui endommage les membranes des bactéries, et provoque la lyse bactérienne. Cependant, ce processus génère une chaleur qui peut dénaturer les protéines, c'est pourquoi la sonication doit être faite à 4°C et de façon intermittente.

1.7.2 CellDisruptor (ou French Press)

Cette méthode consiste quant à elle en l'utilisation d'une forte pression pour lyser les bactéries. En bref, la suspension bactérienne est poussée par une forte pression (1,4 kbar) à travers un trou, ce qui force le passage des bactéries et les fait exploser.

1.7.3 Étapes des purifications

1.7.3.1 Chromatographie d'affinité His-Trap

Le tag de 14 histidines situé en N-terminal des protéines produites a une forte affinité pour les ions Ni^{2+}, présents sur les résines Ni-NTA. Cela permet, lorsque la fraction de protéines solubles est injectée sur la résine de chromatographie, de fixer à la colonne de façon spécifique les protéines ayant une affinité pour ces ions. Une fois les protéines fixées sur la résine, on lave la colonne pour éliminer les protéines résiduelles fixées de façon aspécifique. Enfin, une concentration croissante d'imidazole est appliquée sur la résine, sous forme de gradient. L'imidazole, qui a une très forte affinité pour les ions Ni^{2+}, a pour fonction de venir en compétition avec les protéines fixées à la résine, pour les décrocher et les remplacer. Les protéines éluées sont séparées en fractions de 1 mL sur une plaque 96 puits.

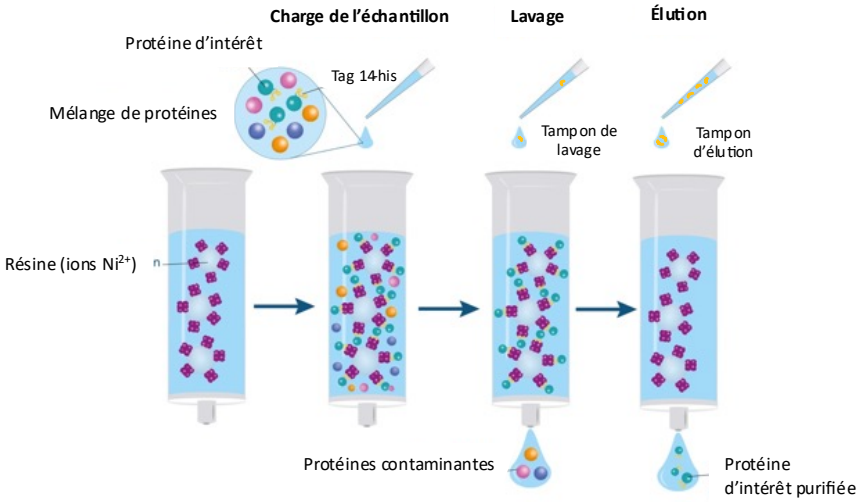


Figure 185 : Principe de la chromatographie d'affinité Histrap. D'après https://www.iba-lifesciences.com/applications/protein-affinity-chromatography/

1.7.3.2 Chromatographie sur résine Héparine

Cette seconde étape de chromatographie est particulièrement adaptée aux protéines se liant à l'ADN. En effet, l'héparine utilisée pour fonctionnaliser la résine mime la structure de

l'ADN, puisqu'elle est recouverte d'anions sulfate, proches du squelette de phosphate de l'ADN.

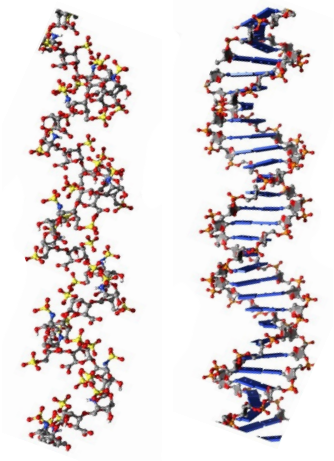


Figure 186 : Comparaison des structures d'une molécule d'héparine (à gauche) et d'ADN (à droite). D'après https://www.york.ac.uk/chemistry/news/deptnews/cracking-the-chiral-code-of-dna-binding/

Cette résine est donc à la fois une résine d'affinité et une résine échangeuse de cations pour les protéines qui se fixent à l'ADN. En effet, elles peuvent se fixer aux groupements sulfate grâce à leurs résidus arginine, lysine et histidine, chargés positivement. Cette étape permet également d'éliminer les contaminations aux acides nucléiques, puisque les ADN polymérases ont généralement une plus grande affinité pour l'héparine que l'ADN (Brennessel *et al.*, 1978).

1.7.3.3 Clivage de l'étiquette de 14 histidines

Le plasmide LS05 utilisé pour exprimer toutes les protéines étudiées ici permet l'expression des protéines avec une étiquette de 14 histidines suivie d'un site de clivage par la protéase du virus de la gravure du tabac (TEV pour Tobacco Etch Virus), qui permet de ne laisser que quelques résidus (5 au maximum : GTGDS) fusionnés à la protéine d'intérêt. La protéase TEV a un site de reconnaissance constitués des résidus suivants : ENLYFQG. Elle a pour rôle de reconnaitre cette séquence et de couper la liaison peptidique en N-terminal de la glycine C-terminale. Cette protéase est produite et purifiée au sein du laboratoire, couplée à un tag de 6 histidines non clivable.

1.7.3.4 Chromatographie d'exclusion stérique

Cette étape de purification vise à faire passer l'échantillon dans un tamis moléculaire en 3 dimensions, formé de billes de gel de taille homogène présentant des pores. Ainsi, lorsque

l'échantillon est injecté sur ce type de colonne, les plus petites molécules, qui peuvent entrer dans les billes de gel, vont y être retardées, tandis que les plus grosse protéines vont traverser la colonne sans passer dans les pores, et donc sans être ralenties. On obtient ainsi une séparation selon le rayon hydrodynamique des protéines : les plus grosses protéines sortent de la colonne en premières, et les plus petites en dernières.

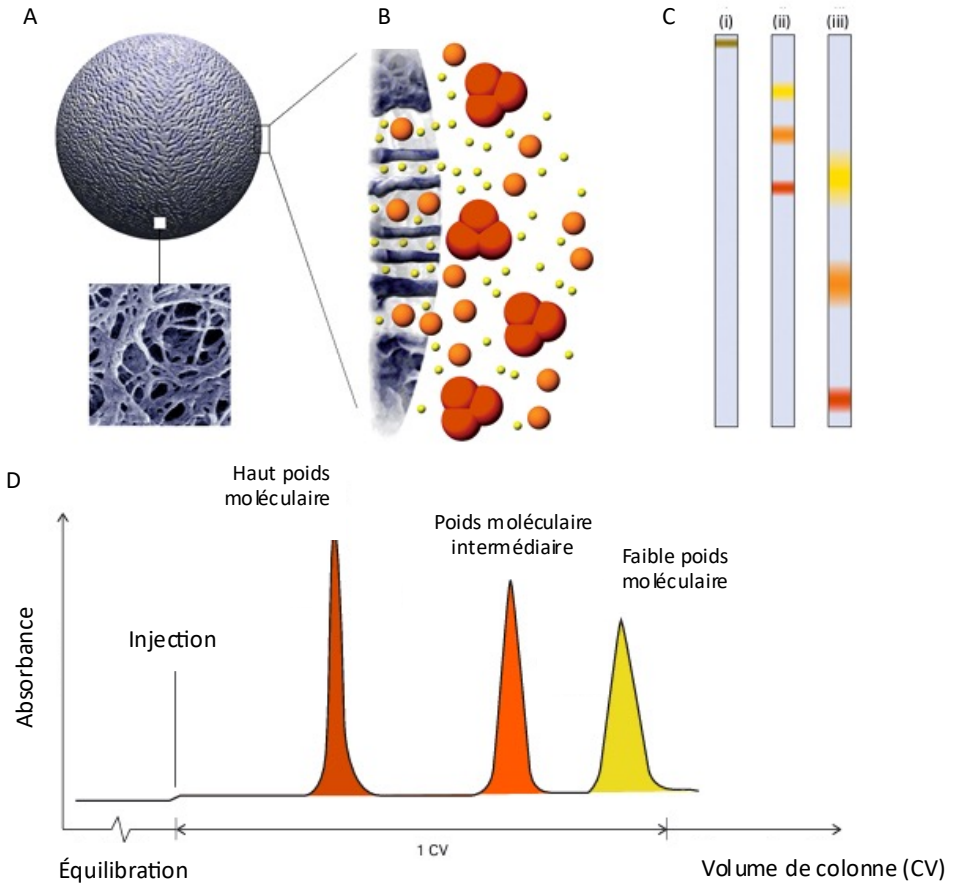


Figure 187 : Principe de la chromatographie par exclusion de taille. A : La résine est formée de billes de gel poreuses. B : Les protéines peuvent entrer ou non dans les pores des billes, selon leur rayon hydrodynamique (résumé par leur poids moléculaire). C : Les protéines de plus grand rayon hydrodynamique ne rentrent pas dans les billes, n'y sont pas retardées, et sont éluées en premières. Plus une molécule est petite, plus elle va passer dans des billes et y être retardée, donc elle sera éluée plus tard. D : En suivant l'absorbance au cours de l'élution, on peut donc suivre la sortie de chaque protéine de la résine. D'après Size Exclusion Chromatography : Principles and Methods (Cytiva, accessible depuis https://www.cytivalifesciences.com/en/us/support/handbooks).

1.7.4 Contrôle qualité des protéines purifiées

1.7.4.1 SDS-PAGE (Laemmli, 1970)

Les électrophorèses en gel de polyacrylamide en conditions dénaturantes ou SDS-PAGE permettent d'obtenir des informations importantes sur la pureté des protéines. Cette méthode consiste en la dénaturation des protéines de l'échantillon par ajout de sels de sulfate (SDS ou

LDS), d'agent réducteur (Dithiothréitol ou DTT) et par chauffage de l'échantillon à 95°C pendant 5 minutes. Cette dénaturation linéarise les protéines et les couvre de charges négatives grâce aux ions sulfate. L'échantillon est alors déposé sur un gel réticulé de polyacrylamide, et un champ électrique est appliqué pour faire migrer les protéines vers une anode grâce à la charge du SDS dans le gel. La vitesse de migration des protéines ne dépend alors que de leur poids moléculaire : plus une protéine est petite, plus elle migre vite, car elle traverse facilement le réseau d'acrylamide formé dans le gel. En utilisant un échantillon marqueur de poids moléculaire (qui contient plusieurs molécules de poids moléculaire connu), on peut déterminer approximativement la masse moléculaire des protéines présentes sur le gel, après coloration de ce dernier.

Dans ces travaux, deux systèmes SDS-PAGE ont été utilisés : Mini Protean TGX (BioRad) et mPAGE Bis-Tris (Merck). Ce second système a été utilisé avec un tampon MOPS ou MES : le MOPS propose une meilleure séparation des protéines de haut poids moléculaire et le MES une meilleure résolution pour les petits poids moléculaires. La coloration des gels a été réalisée rapidement avec un colorant appelé InstantBlue (AbCam) et les gels ont été visualisés grâce à un Chemidoc (BioRad).

1.7.4.2 La mesure du spectre d'absorbance entre 220 et 350 nm

L'autre élément fondamental du contrôle qualité des échantillons protéiques est la mesure du spectre d'absorbance de l'échantillon entre 220 et 350 nm. Dans ces travaux, ces mesures ont toujours été réalisées au NanoDrop. En cours de purification, cela permet d'obtenir plusieurs informations :

- Sur la concentration de la protéine, grâce à son absorbance à 280 nm. Celle-ci est liée à la présence de résidus tryptophane et tyrosine. Dès lors que la séquence de la protéine est connue, il est possible de déterminer son coefficient d'extinction molaire (ε) à 280 nm (par le calcul, ou *via* des services en ligne comme Expasy ProtParam); c'est cette valeur qui permet de déterminer la concentration molaire de la protéine en solution, grâce à la loi de Beer-Lambert : $A = \varepsilon . l . c$. Dans le cas des mesures au NanoDrop, la longueur du trajet optique l est négligeable, donc la concentration molaire de la protéine peut être calculée : $c\ (mol/L) = \frac{A}{\varepsilon}$. On peut ensuite obtenir la concentration massique de la protéine en multipliant c par la masse molaire de la protéine en daltons (Da) ;

- Sur la présence d'acides nucléiques contaminants (avec un pic d'absorbance à 260 nm : si le ratio $\frac{Abs260nm}{Abs280nm}$ est supérieur à 0,6, on considère en théorie qu'il y a une contamination aux acides nucléiques ; dans les faits, certains résidus pouvant absorber à 260 nm (la phénylalanine surtout), la valeur seuil est plus haute, autour de 0,8-0,9) ;
- Sur l'état d'agrégation des protéines en solution : si le pic d'absorbance à 280 nm ne revient pas à 0 UA à 340 nm (si une pente est présente au lieu d'un retour à la ligne de base) cela peut indiquer une agrégation des protéines en solution (Pignataro *et al.*, 2020). Un indice d'agrégation peut être calculé : $\frac{100\times Abs340nm}{Abs280nm-Abs340nm}$. Si cet indice est supérieur à 2, alors on considère qu'il y a agrégation.

Cependant, cette méthode a ses limites, en particulier après l'étape de chromatographie HisTrap, car l'échantillon contient de l'imidazole qui absorbe à 280 nm. Mais lors des étapes suivantes, lorsque l'échantillon n'en contient plus, cette méthode de quantification est parfaitement adaptée.

1.7.4.3 Spectrométrie de masse MALDI-TOF

Le MALDI-TOF (*Matrix Assisted Laser Desorption Ionization – Time Of Flight*) consiste à ioniser l'échantillon protéique à l'aide d'un laser (MALDI), puis à mesurer le temps qu'il met à exciter un détecteur après avoir traversé un champ électromagnétique. Ce temps (temps de vol, TOF) est dépendant du rapport masse/charge (m/z) de la molécule. Dans le cas du MALDI, lors de l'ionisation, la grande majorité des protéines ne reçoivent qu'une charge supplémentaire (z=1), donc la mesure m/z est égale à la masse moléculaire de la protéine. Cela permet d'obtenir une mesure très précise de la masse moléculaire d'une protéine purifiée (ou de plusieurs dans le cas d'un mélange hétérogène).

1.7.4.4 Diffusion dynamique de la lumière (DLS)

La diffusion dynamique de la lumière se base sur le mouvement brownien des molécules en solution : en résumé, les plus grosses molécules bougent moins vite que les petites. L'idée ici est d'appliquer à l'échantillon une lumière, et de mesurer sa diffusion Rayleigh à un angle de 90°. Un traitement mathématique (fonction d'autocorrélation) permet d'obtenir des informations sur la variation du signal diffusé : si la fonction d'autocorrélation décroit

rapidement, c'est que le mouvement des espèces en solution est rapide. Un coefficient de diffusion peut alors être calculé, et l'équation de Stokes-Einstein permet alors de déterminer le rayon hydrodynamique des espèces présentes dans la solution. Cela permet donc de savoir si l'échantillon est monodisperse, c'est-à-dire qu'il contient des populations de molécules avec un rayon hydrodynamique uniforme, ou s'il est polydisperse, c'est-à-dire qu'il contient des espèces de différentes tailles (agrégats, contaminants, etc).

2 Oligonucléotides utilisés dans les expériences de caractérisation enzymatique

- Extension d'amorce

Séquences des oligonucléotides :

5′-GGGGTAGCTGCGCATAAGCTTATGACAATT-3′

5′-FAM-AATTGTCATAAGCTTATGCG-3′

Structure adoptée :

```
  3′ TTAACAGTATTCGAATACGCGTCGATGGGG 5′
FAM 5′ AATTGTCATAAGCTTATGCG 3′
```

- Microhomology End Joining

Séquences des oligonucléotides :

5′-FAM-AATCACCAGTACGCGCCGTTGCGTCATCTGACGTC-3′

5′-p-CGGCCATGACGCCAAGACCAGG-3′

5′-GACGCAACGGCGCGTACTGGTGATT-3′

5′-CCTGGTCTTGGCGTCATGGCGCGACGTCAGAT-3′

Structure adoptée :

```
  3′ TTAGTGGTCATGCGCGGCAACGCAG TAGACTGCAGCGCGGTACTGCGGTTCTGGTCC 5′
FAM 5′ AATCACCAGTACGCGCCGTTGCGTC-ATCTGACGTC CGGCCATGACGCCAAGACCAGG 3′
                         P
```

- NHEJ

Séquences des oligonucléotides :

5′-FAM-GCCGTTGCGTCAT-3′

5′-p-CCGCCATGACGC-3′

5′-GCGTCATGGCGGCAT-3′

5′-ACGCAACGGC-3′

Structure adoptée :

```
  3′ CGGCAACGCA TACGGCGGTACTGCG 5′
FAM 5′ GCCGTTGCGTCAT CCGCCATGACGC 3′
           P
```

- NHEJ-cis

Séquences des oligonucléotides :

5′-FAM-AATCACCAGTACGCCGTTGCGT-3′

5′-p-TATCGCCATGACGCGGTTCTGGTCC-3′

5′-TACACGCAACGGCGTACTGGTGATT-3′

5′-GGACCAGAACCGCGTCATGGCG-3′

Structure adoptée :

```
  3′ TTAGTGGTCATGCGGCAACGCACAT GCGGTACTGCGCCAAGACCAGG 5′
FAM 5′ AATCACCAGTACGCCGTTGCGT TATCGCCATGACGCGGTTCTGGTCC 3′
               P
```

- Gap-filling

Séquences des oligonucléotides :

5′-FAM-AATCACCAGTACGCCGTTGCGT-3′

5′-p-TATCGCCATGACGCGGTTCTGGTCC-3′

5′-GGACCAGAACCGCGTCATGGCGATACACGCAACGGCGTACTGGTGATT-3′

Structure adoptée :

```
  3′ TTAGTGGTCATGCGGCAACGCACATAGCGGTACTGCGCCAAGACCAGG 5′
FAM 5′ AATCACCAGTACGCCGTTGCGT TATCGCCATGACGCGGTTCTGGTCC 3′
               P
```

- Terminal transférase

Séquences des oligonucléotides :

5′-AAGCTTATGACAATT-3′

5′-FAM-AATTGTCATAAGCTTATGCG-3′

Structure adoptée :

```
  3′ TTAACAGTATTCGAA 5′
FAM 5′ AATTGTCATAAGCTTATGCG 3′
```

3 Caractérisation cinétique des ADN polymérases X

Lors de la caractérisation enzymatique des ADN polymérases, on parle de k_{obs} et non de k_{cat}, car la réaction enzymatique catalysée par les ADN polymérases est de second ordre : A+B→AB (avec A l'ADN initial et B le dGTP), ce qui donne $k_{cat} = k[A][B]$. L'expérience réalisée est dite de pseudo premier ordre, car la concentration en ADN est beaucoup plus élevée que celle de dGTP et ne varie pas pendant l'expérience, donc la vitesse de formation du produit ne dépend virtuellement que de la concentration en dGTP. On obtient donc une approximation du k_{cat}, appelée k_{obs}, avec $k_{obs} = k[A]$ (Pollard and De La Cruz, 2013).

4 Bases de cristallographie

L'objectif de cette partie est de présenter des éléments basiques de cristallographie, utiles à la compréhension des résultats présentés. Cette partie n'est pas une présentation exhaustive de la technique de cristallographie aux rayons X.

4.1 Cristallogenèse

Pour étudier la structure de protéines et de complexes par diffraction des rayons X, il faut obtenir des cristaux de ces molécules. Le cristal est une forme organisée et compacte des macromolécules, qui permet d'amplifier le signal lors d'expériences de diffraction des rayons X, pour obtenir des informations détaillées sur la structure de ces macromolécules.

La cristallogenèse consiste à permettre à une molécule soluble dans un liquide de passer dans un état solide, où l'ensemble des molécules présentes structurées de façon identique s'organisent dans l'espace tridimensionnel de façon régulière et répétée par des contacts cristallins. Cette étape est empirique : aujourd'hui, on ne peut pas prédire dans quelles conditions une molécule va cristalliser. Il faut donc tester de nombreuses conditions (pH, force ionique, solvants, autres polymères présents en solution, etc.), qui permettront de diminuer la solubilité de la molécule d'intérêt pour former un cristal. Ces conditions permettent un changement de phase, et on représente les états possibles sous forme d'un diagramme de phase, au sein duquel la protéine doit atteindre un état métastable, sans précipiter. Pour cela, il est possible de faire varier les espèces présentes dans la solution de cristallisation, ainsi que leurs concentrations, et celle de la protéine. Une fois que les conditions sont atteintes, un germe de cristal va pousser tant qu'il reste de la protéine disponible en solution et que celle-ci reste dans la zone métastable.

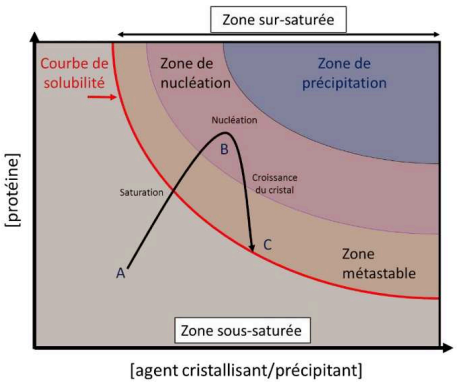


Figure 188 : Diagramme de phases à deux dimensions d'une macromolécule. A : Macromolécule soluble. B : Nucléation du cristal. C : Cristal final stable. D'après Giegé et al. Biocrystallography: past, present, future. HFSP J. 2010 Jun;4(3-4):109-21.

L'obtention de cristaux nécessite une très grande pureté des protéines, ainsi qu'une grande homogénéité, c'est pourquoi les protéines utilisées pour la cristallographie ont toutes été purifiées jusqu'à une étape de chromatographie d'exclusion stérique.

4.2 Anatomie du cristal

Les cristaux formés sont un agencement organisé et répété des molécules dans l'espace tridimensionnel. L'unité de base du cristal est la maille élémentaire, qui forme un système cristallin par répétition et translation. Cette maille peut être caractérisée pour chaque cristal à 6 paramètres : les longueurs a, b et c et les angles α, β et γ. Dans la nature, on peut obtenir 7 systèmes cristallins (cubique, hexagonal, trigonal, tétragonal, orthorhombique, monoclinique ou triclinique). Au sein d'une maille, on peut retrouver une molécule ou plusieurs, répétées par des éléments de symétrie organisés autour de nœuds. La position de ces nœuds détermine le type de maille élémentaire (primitive, base centrée, corps centrée, face centrée). L'ensemble de ces éléments, ainsi que d'autres opérations symétriques, déterminent 65 groupes d'espace possibles, qui permettent de connaitre l'organisation des molécules au sein du cristal. La détermination du groupe d'espace d'un cristal est la première étape du traitement des données de cristallographie.

4.3 La problème de la phase

Le premier objectif de l'étape de résolution de la structure est d'obtenir une carte de densité électronique correspondant à la protéine étudiée, et de l'utiliser pour placer les acides aminés formant la protéine d'une façon s'accordant le mieux avec cette densité obtenue expérimentalement. Cette densité électronique dépend en chaque point de l'espace du volume de la maille et est obtenue par transformée de Fourier des facteurs de structure, définis par leur amplitude et leur phase. L'obtention des clichés de diffraction permet d'obtenir l'intensité des taches obtenues (qui correspond au carré de l'amplitude du facteur de structure). Cependant, pour obtenir la densité électronique d'une protéine, il est indispensable de connaitre la phase de tous les facteurs de structure, or celle-ci est perdue lors de l'obtention des données. Il existe deux méthodes principales pour obtenir la phase : le phasage expérimental (utilisation d'atomes lourds dans les cristaux et utilisation de leurs caractéristiques particulières pour obtenir la phase), et le remplacement moléculaire (utilisation du modèle structural d'une protéine proche et placement de celui-ci dans la maille considérée).

4.4 Bases du remplacement moléculaire

Dans un premier temps, il est important de connaître le nombre de molécules présentes dans chaque maille (unité asymétrique). Pour cela, on utilise le coefficient de Matthews (V_m, compris entre 1,68 et 3,53 Å^3) : $V_m = \frac{V}{n \times M}$. Ici, V est le volume de la maille cristalline, M le poids moléculaire de la molécule, et n le nombre de molécules par unité asymétrique.

Il est ensuite possible d'utiliser le modèle proche choisi et le nombre de molécules par unité asymétrique pour comparer des données cristallographiques théoriques basées sur ce modèle aux données expérimentales (les données comparées sont appelées cartes de Patterson, basées sur les intensités des facteurs de structure). Il est nécessaire d'explorer toutes les rotations et translations possibles pour placer au mieux le modèle dans la maille de la protéine étudiée.

4.5 Fonctions de Buster utilisées dans les affinements

Affinement « Rigid-Body » (commande -RB) : on considère que chaque chaîne est un objet rigide qui se déplace en bloc : c'est l'approche utilisée juste après l'étape de remplacement moléculaire.

Information des symétries non cristallines (commande -autoNCS) : lorsque plusieurs copies de la molécule se trouvent dans une unité asymétrique, on peut penser que les modifications d'une molécule peuvent aussi bien s'appliquer aux autres. Cela permet de moyenner les variations, et donc d'augmenter le rapport signal/bruit. Cette commande est paramétrée de façon optimale par le logiciel en mode automatique et est très utile lorsque la diffraction du cristal est anisotrope (si la diffraction n'est pas homogène dans les trois directions de l'espace). Il peut cependant être utile d'empêcher le logiciel de supprimer les éléments aberrants lui-même (avec la commande -autoNCS_noprune).

Utilisation de tenseurs TLS (commande -TLS) (*Translation Rotation and Screw rotation*) : Ce sont des opérateurs mathématiques permettant de décrire le désordre et les vibrations moléculaires par plusieurs éléments indépendants. Cette fonction permet de décrire les vibrations moléculaires, et d'affiner le modèle (à moyenne résolution) qui correspondra alors au mieux aux données expérimentales.

Utilisation d'une cible d'affinement (commande -target) : lorsqu'un modèle de bonne qualité à plus haute résolution existe déjà, il peut être utilisé comme cible, de sorte que l'affinement tende vers une structure similaire au modèle. Cependant, cette commande est à utiliser en analysant le modèle lors de son affinement, car si les données expérimentales ne correspondent pas parfaitement au modèle cible, cela peut ralentir l'affinement. Dans mon cas, avec la structure du mutant « Polβ-like », le placement d'une boucle d'après les données expérimentales était différent du modèle utilisé initialement (7M43) comme cible pour l'affinement, ce qui ne permettait pas un bon affinement. Le simple fait de changer de modèle cible pour un modèle avec un placement différent de cette boucle (1XSL) a permis d'améliorer drastiquement l'affinement ;

Placement automatique de molécules d'eau (commande -WAT). Dans certains modèles, la résolution permet d'obtenir des densités électroniques correspondant à des molécules d'eau. Celles-ci peuvent être ajoutées ou supprimées automatiquement par le logiciel au besoin.

Il existe de nombreuses autres commandes de Buster, et dans les cas difficiles, il peut être utile d'en tester un certain nombre avant de trouver les meilleures.

4.6 R_{work} et R_{free}

$$R_{work} = \frac{\sum_{hkl}\big||Fobs| - k|Fcalc|\big|}{\sum_{hkl}|Fobs|}$$

$$R_{free} = \frac{\sum_{hkl\in T}\big||Fobs| - k|Fcalc|\big|}{\sum_{hkl\in T}|Fobs|}$$

Avec |Fobs| l'amplitude d'un facteur de structure observé ; |Fcalc| l'amplitude d'un facteur de structure calculé, k le facteur de mise à l'échelle des données, et T une sous partie des données n'ayant pas été utilisée par l'affinement.

Références bibliographiques

Abello, A., 2019. Spécialisation de Ku80c dans le couplage entre coupure et réparation de l'ADN lors des réarrangements programmés du génome chez Paramecium tetraurelia (phdthesis). Université Paris Saclay (COmUE).
Abello, A., Régnier, V., Arnaiz, O., Bars, R.L., Bétermier, M., Bischerour, J., 2020. Functional diversification of Paramecium Ku80 paralogs safeguards genome integrity during precise programmed DNA elimination. PLOS Genet. 16, e1008723. https://doi.org/10.1371/journal.pgen.1008723
Abramson, R.D., 1995. 4 - Thermostable DNA Polymerases, in: Innis, M.A., Gelfand, D.H., Sninsky, J.J. (Eds.), PCR Strategies. Academic Press, San Diego, pp. 39–57. https://doi.org/10.1016/B978-012372182-2/50006-X
Ahel, I., Rass, U., El-Khamisy, S.F., Katyal, S., Clements, P.M., McKinnon, P.J., Caldecott, K.W., West, S.C., 2006. The neurodegenerative disease protein aprataxin resolves abortive DNA ligation intermediates. Nature 443, 713–716. https://doi.org/10.1038/nature05164
Ahnesorg, P., Smith, P., Jackson, S.P., 2006. XLF Interacts with the XRCC4-DNA Ligase IV Complex to Promote DNA Nonhomologous End-Joining. Cell 124, 301–313. https://doi.org/10.1016/j.cell.2005.12.031
Albright, L.M., Slatko, B.E., 1994. Denaturing Polyacrylamide Gel Electrophoresis. Curr. Protoc. Hum. Genet. 00, A.3F.1-A.3F.4. https://doi.org/10.1002/0471142905.hga03fs00
Alhmoud, J.F., Woolley, J.F., Al Moustafa, A.-E., Malki, M.I., 2020. DNA Damage/Repair Management in Cancers. Cancers 12, 1050. https://doi.org/10.3390/cancers12041050
Aliotta, J.M., Pelletier, J.J., Ware, J.L., Moran, L.S., Benner, J.S., Kong, H., 1996. Thermostable *Bst* DNA polymerase I lacks a 3′ → 5′ proofreading exonuclease activity. Genet. Anal. Biomol. Eng. 12, 185–195. https://doi.org/10.1016/S1050-3862(96)80005-2
Allen, S.E., Hug, I., Pabian, S., Rzeszutek, I., Hoehener, C., Nowacki, M., 2017. Circular Concatemers of Ultra-Short DNA Segments Produce Regulatory RNAs. Cell 168, 990-999.e7. https://doi.org/10.1016/j.cell.2017.02.020
Allfrey, V.G., Faulkner, R., Mirsky, A.E., 1964. ACETYLATION AND METHYLATION OF HISTONES AND THEIR POSSIBLE ROLE IN THE REGULATION OF RNA SYNTHESIS. Proc. Natl. Acad. Sci. U. S. A. 51, 786–794. https://doi.org/10.1073/pnas.51.5.786
Allinson, S.L., Dianova, I.I., Dianov, G.L., 2001. DNA polymerase β is the major dRP lyase involved in repair of oxidative base lesions in DNA by mammalian cell extracts. EMBO J. 20, 6919–6926. https://doi.org/10.1093/emboj/20.23.6919
Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J., 1990. Basic local alignment search tool. J. Mol. Biol. 215, 403–410. https://doi.org/10.1016/S0022-2836(05)80360-2
Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., Lipman, D.J., 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25, 3389–3402. https://doi.org/10.1093/nar/25.17.3389
Anand, R., Ranjha, L., Cannavo, E., Cejka, P., 2016. Phosphorylated CtIP Functions as a Co-factor of the MRE11-RAD50-NBS1 Endonuclease in DNA End Resection. Mol. Cell 64, 940–950. https://doi.org/10.1016/j.molcel.2016.10.017
Aoufouchi, S., Flatter, E., Dahan, A., Faili, A., Bertocci, B., Storck, S., Delbos, F., Cocea, L., Gupta, N., Weill, J.-C., Reynaud, C.-A., 2000. Two novel human and mouse DNA polymerases of the polX family. Nucleic Acids Res. 28, 3684–3693. https://doi.org/10.1093/nar/28.18.3684
Aravind, L., Koonin, E.V., 2001. Prokaryotic homologs of the eukaryotic DNA-end-binding protein Ku, novel domains in the Ku protein and prediction of a prokaryotic double-strand break repair system. Genome Res. 11, 1365–1374. https://doi.org/10.1101/gr.181001
Aravind, L., Koonin, E.V., 1999. DNA polymerase β-like nucleotidyltransferase superfamily: Identification of three new families, classification and evolutionary history. Nucleic Acids Res. 27, 1609–1618. https://doi.org/10.1093/nar/27.7.1609
Aravind, L., Koonin, E.V., 1998. Phosphoesterase domains associated with DNA polymerases of diverse origins. Nucleic Acids Res. 26, 3746–3752. https://doi.org/10.1093/nar/26.16.3746
Arnaiz, O., Mathy, N., Baudry, C., Malinsky, S., Aury, J.-M., Wilkes, C.D., Garnier, O., Labadie, K., Lauderdale, B.E., Mouël, A.L., Marmignon, A., Nowacki, M., Poulain, J., Prajer, M., Wincker, P., Meyer, E., Duharcourt, S., Duret, L., Bétermier, M., Sperling, L., 2012. The Paramecium Germline Genome Provides a Niche for Intragenic Parasitic DNA: Evolutionary Dynamics of Internal Eliminated Sequences. PLOS Genet. 8, e1002984. https://doi.org/10.1371/journal.pgen.1002984
Arnaiz, O., Van Dijk, E., Bétermier, M., Lhuillier-Akakpo, M., de Vanssay, A., Duharcourt, S., Sallet, E., Gouzy, J., Sperling, L., 2017. Improved methods and resources for paramecium genomics: transcription units, gene annotation and gene expression. BMC Genomics 18, 483. https://doi.org/10.1186/s12864-017-3887-z
Arnould, C., Rocher, V., Finoux, A.-L., Clouaire, T., Li, K., Zhou, F., Caron, P., Mangeot, P.E., Ricci, E.P., Mourad, R., Haber, J.E., Noordermeer, D., Legube, G., 2021. Loop extrusion as a mechanism for formation of DNA damage repair foci. Nature 590, 660–665. https://doi.org/10.1038/s41586-021-03193-z
Arnould, C., Rocher, V., Saur, F., Bader, A.S., Muzzopappa, F., Collins, S., Lesage, E., Le Bozec, B., Puget, N.,

Clouaire, T., Mangeat, T., Mourad, R., Ahituv, N., Noordermeer, D., Erdel, F., Bushell, M., Marnef, A., Legube, G., 2023. Chromatin compartmentalization regulates the response to DNA damage. Nature 623, 183–192. https://doi.org/10.1038/s41586-023-06635-y

Aury, J.-M., Jaillon, O., Duret, L., Noel, B., Jubin, C., Porcel, B.M., Ségurens, B., Daubin, V., Anthouard, V., Aiach, N., Arnaiz, O., Billaut, A., Beisson, J., Blanc, I., Bouhouche, K., Câmara, F., Duharcourt, S., Guigo, R., Gogendeau, D., Katinka, M., Keller, A.-M., Kissmehl, R., Klotz, C., Koll, F., Le Mouël, A., Lepère, G., Malinsky, S., Nowacki, M., Nowak, J.K., Plattner, H., Poulain, J., Ruiz, F., Serrano, V., Zagulski, M., Dessen, P., Bétermier, M., Weissenbach, J., Scarpelli, C., Schächter, V., Sperling, L., Meyer, E., Cohen, J., Wincker, P., 2006. Global trends of whole-genome duplications revealed by the ciliate Paramecium tetraurelia. Nature 444, 171–178. https://doi.org/10.1038/nature05230

Aziz, R.K., Breitbart, M., Edwards, R.A., 2010. Transposases are the most abundant, most ubiquitous genes in nature. Nucleic Acids Res. 38, 4207–4217. https://doi.org/10.1093/nar/gkq140

Balint, E., Unk, I., 2020. Selective Metal Ion Utilization Contributes to the Transformation of the Activity of Yeast Polymerase η from DNA Polymerization toward RNA Polymerization. Int. J. Mol. Sci. 21, 8248. https://doi.org/10.3390/ijms21218248

Baños, B., Lázaro, J.M., Villar, L., Salas, M., de Vega, M., 2008. Editing of misaligned 3′-termini by an intrinsic 3′–5′ exonuclease activity residing in the PHP domain of a family X DNA polymerase. Nucleic Acids Res. 36, 5736–5749. https://doi.org/10.1093/nar/gkn526

Baños, B., Villar, L., Salas, M., de Vega, M., 2010. Intrinsic apurinic/apyrimidinic (AP) endonuclease activity enables Bacillus subtilis DNA polymerase X to recognize, incise, and further repair abasic sites. Proc. Natl. Acad. Sci. 107, 19219–19224. https://doi.org/10.1073/pnas.1013603107

Baudat, F., Imai, Y., de Massy, B., 2013. Meiotic recombination in mammals: localization and regulation. Nat. Rev. Genet. 14, 794–806. https://doi.org/10.1038/nrg3573

Baudry, C., Malinsky, S., Restituito, M., Kapusta, A., Rosa, S., Meyer, E., Bétermier, M., 2009. PiggyMac, a domesticated piggyBac transposase involved in programmed genome rearrangements in the ciliate Paramecium tetraurelia. Genes Dev. 23, 2478–2483. https://doi.org/10.1101/gad.547309

Beard, W.A., 2020. DNA polymerase β: Closing the gap between structure and function. DNA Repair, Tribute to Samuel H. Wilson: Shining Light on Base Excision DNA Repair 93, 102910. https://doi.org/10.1016/j.dnarep.2020.102910

Beard, W.A., Shock, D.D., Batra, V.K., Prasad, R., Wilson, S.H., 2014. Substrate-induced DNA Polymerase β Activation. J. Biol. Chem. 289, 31411–31422. https://doi.org/10.1074/jbc.M114.607432

Beard, W.A., Wilson, S.H., 2014. Structure and Mechanism of DNA Polymerase β. Biochemistry 53, 2768–2780. https://doi.org/10.1021/bi500139h

Bębenek, A., Ziuzia-Graczyk, I., 2018. Fidelity of DNA replication—a matter of proofreading. Curr. Genet. 64, 985–996. https://doi.org/10.1007/s00294-018-0820-1

Bebenek, K., Garcia-Diaz, M., Blanco, L., Kunkel, T.A., 2003. The Frameshift Infidelity of Human DNA Polymerase λ: IMPLICATIONS FOR FUNCTION*. J. Biol. Chem. 278, 34685–34690. https://doi.org/10.1074/jbc.M305705200

Bebenek, K., Garcia-Diaz, M., Foley, M.C., Pedersen, L.C., Schlick, T., Kunkel, T.A., 2008. Substrate-induced DNA strand misalignment during catalytic cycling by DNA polymerase λ. EMBO Rep. 9, 459–464. https://doi.org/10.1038/embor.2008.33

Bebenek, K., Garcia-Diaz, M., Patishall, S.R., Kunkel, T.A., 2005. Biochemical Properties of Saccharomyces cerevisiae DNA Polymerase IV*. J. Biol. Chem. 280, 20051–20058. https://doi.org/10.1074/jbc.M501981200

Bebenek, K., Kunkel, T.A., 2004. Functions of DNA Polymerases, in: Advances in Protein Chemistry, DNA Repair and Replication. Academic Press, pp. 137–165. https://doi.org/10.1016/S0065-3233(04)69005-X

Bebenek, K., Pedersen, L.C., Kunkel, T.A., 2014. Structure–Function Studies of DNA Polymerase λ. Biochemistry 53, 2781–2792. https://doi.org/10.1021/bi4017236

Bellini, D., Fordham-Skelton, A.P., Papiz, M.Z., 2011. STRU-Cloning: A Fast, Inexpensive and Efficient Cloning Procedure Applicable to Both Small Scale and Structural Genomics Size Cloning. Mol. Biotechnol. 48, 30–37. https://doi.org/10.1007/s12033-010-9345-7

Belousova, E.A., Lavrik, O.I., 2015. DNA polymerases β and λ and their roles in cell. DNA Repair, DNA polymerases 29, 112–126. https://doi.org/10.1016/j.dnarep.2015.02.001

Bentchikou, E., Servant, P., Coste, G., Sommer, S., 2007. Additive effects of SbcCD and PolX deficiencies in the in vivo repair of DNA double-strand breaks in Deinococcus radiodurans. J. Bacteriol. 189, 4784–4790. https://doi.org/10.1128/JB.00452-07

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., Bourne, P.E., 2000. The Protein Data Bank. Nucleic Acids Res. 28, 235–242. https://doi.org/10.1093/nar/28.1.235

Bertrand, C., Thibessard, A., Bruand, C., Lecointe, F., Leblond, P., 2019. Bacterial NHEJ: a never ending story. Mol. Microbiol. 111, 1139–1151. https://doi.org/10.1111/mmi.14218

Betancurt-Anzola, L., Martínez-Carranza, M., Delarue, M., Zatopek, K.M., Gardner, A.F., Sauguet, L., 2023.

Molecular basis for proofreading by the unique exonuclease domain of Family-D DNA polymerases. Nat. Commun. 14, 8306. https://doi.org/10.1038/s41467-023-44125-x
Bétermier, M., 2004. Large-scale genome remodelling by the developmentally programmed elimination of germ line sequences in the ciliate Paramecium. Res. Microbiol., Genome plasticity and the evolution of microbial genomes 155, 399–408. https://doi.org/10.1016/j.resmic.2004.01.017
Bétermier, M., Bertrand, P., Lopez, B.S., 2014. Is Non-Homologous End-Joining Really an Inherently Error-Prone Process? PLoS Genet. 10, e1004086. https://doi.org/10.1371/journal.pgen.1004086
Bétermier, M., Duharcourt, S., 2015. Programmed Rearrangement in Ciliates: Paramecium, in: Mobile DNA III. John Wiley & Sons, Ltd, pp. 369–388. https://doi.org/10.1128/9781555819217.ch17
Betermier, M., Duharcourt, S., 2014. Programmed Rearrangement in Ciliates: Paramecium. Microbiol. Spectr. 2, 2.6.26. https://doi.org/10.1128/microbiolspec.MDNA3-0035-2014
Bétermier, M., Duharcourt, S., Seitz, H., Meyer, E., 2000. Timing of developmentally programmed excision and circularization of Paramecium internal eliminated sequences. Mol. Cell. Biol. 20, 1553–1561. https://doi.org/10.1128/MCB.20.5.1553-1561.2000
Bétermier, M., Klobutcher, L.A., Orias, E., 2023. Programmed chromosome fragmentation in ciliated protozoa: multiple means to chromosome ends. Microbiol. Mol. Biol. Rev. 87, e00184-22. https://doi.org/10.1128/mmbr.00184-22
Beukers, R., Eker, A.P.M., Lohman, P.H.M., 2008. 50 years thymine dimer. DNA Repair 7, 530–543. https://doi.org/10.1016/j.dnarep.2007.11.010
Bhattarai, H., Gupta, R., Glickman, M.S., 2014. DNA ligase C1 mediates the LigD-independent nonhomologous end-joining pathway of Mycobacterium smegmatis. J. Bacteriol. 196, 3366–3376. https://doi.org/10.1128/JB.01832-14
Bienstock, R.J., Beard, W.A., Wilson, S.H., 2014. Phylogenetic analysis and evolutionary origins of DNA polymerase X-family members. DNA Repair 22, 77–88. https://doi.org/10.1016/j.dnarep.2014.07.003
Bischerour, J., Bhullar, S., Denby Wilkes, C., Régnier, V., Mathy, N., Dubois, E., Singh, A., Swart, E., Arnaiz, O., Sperling, L., Nowacki, M., Bétermier, M., 2018. Six domesticated PiggyBac transposases together carry out programmed DNA elimination in Paramecium. eLife 7, e37927. https://doi.org/10.7554/eLife.37927
Black, S.J., Ozdemir, A.Y., Kashkina, E., Kent, T., Rusanov, T., Ristic, D., Shin, Y., Suma, A., Hoang, T., Chandramouly, G., Siddique, L.A., Borisonnik, N., Sullivan-Reed, K., Mallon, J.S., Skorski, T., Carnevale, V., Murakami, K.S., Wyman, C., Pomerantz, R.T., 2019. Molecular basis of microhomology-mediated end-joining by purified full-length Polθ. Nat. Commun. 10, 4423. https://doi.org/10.1038/s41467-019-12272-9
Blanca, G., Shevelev, I., Ramadan, K., Villani, G., Spadari, S., Hübscher, U., Maga, G., 2003. Human DNA polymerase λ diverged in evolution from DNA polymerase β toward specific Mn++ dependence: A kinetic and thermodynamic study. Biochemistry 42, 7467–7476. https://doi.org/10.1021/bi034198m
Blanca, G., Villani, G., Shevelev, I., Ramadan, K., Spadari, S., Hübscher, U., Maga, G., 2004. Human DNA Polymerases λ and β Show Different Efficiencies of Translesion DNA Synthesis past Abasic Sites and Alternative Mechanisms for Frameshift Generation. Biochemistry 43, 11605–11615. https://doi.org/10.1021/bi049050x
Blasius, M., Shevelev, I., Jolivet, E., Sommer, S., Hübscher, U., 2006. DNA polymerase X from Deinococcus radiodurans possesses a structure-modulated 3′→5′ exonuclease activity involved in radioresistance. Mol. Microbiol. 60, 165–176. https://doi.org/10.1111/j.1365-2958.2006.05077.x
Blier, P.R., Griffith, A.J., Craft, J., Hardin, J.A., 1993. Binding of Ku protein to DNA. Measurement of affinity for ends and demonstration of binding to nicks. J. Biol. Chem. 268, 7594–7601. https://doi.org/10.1016/S0021-9258(18)53216-6
Boiteux, S., Coste, F., Castaing, B., 2017. Repair of 8-oxo-7,8-dihydroguanine in prokaryotic and eukaryotic cells: Properties and biological roles of the Fpg and OGG1 DNA N-glycosylases. Free Radic. Biol. Med., Oxidative DNA Damage & Repair 107, 179–201. https://doi.org/10.1016/j.freeradbiomed.2016.11.042
Bollum, F.J., 1960. Calf Thymus Polymerase. J. Biol. Chem. 235, 2399–2403. https://doi.org/10.1016/S0021-9258(18)64634-4
Boulé, J.-B., Rougeon, F., Papanicolaou, C., 2001. Terminal Deoxynucleotidyl Transferase Indiscriminately Incorporates Ribonucleotides and Deoxyribonucleotides*. J. Biol. Chem. 276, 31388–31393. https://doi.org/10.1074/jbc.M105272200
Bourque, G., Burns, K.H., Gehring, M., Gorbunova, V., Seluanov, A., Hammell, M., Imbeault, M., Izsvák, Z., Levin, H.L., Macfarlan, T.S., Mager, D.L., Feschotte, C., 2018. Ten things you should know about transposable elements. Genome Biol. 19, 199. https://doi.org/10.1186/s13059-018-1577-z
Bowater, R.P., Gates, A.J., 2015. Nucleotides: Structure and Properties, in: Encyclopedia of Life Sciences. John Wiley & Sons, Ltd, pp. 1–9. https://doi.org/10.1002/9780470015902.a0001333.pub3
Brandt, V.L., Roth, D.B., 2008. G.O.D.'s Holy Grail: discovery of the RAG proteins. J. Immunol. Baltim. Md 1950 180, 3–4. https://doi.org/10.4049/jimmunol.180.1.3
Brennessel, B.A., Buhrer, D.P., Gottlieb, A.A., 1978. Use of insoluble heparin for isolation of DNA polymerase

enzymes from murine myeloma. Anal. Biochem. 87, 411–417. https://doi.org/10.1016/0003-2697(78)90690-5
Brooks, P.J., 2017. The cyclopurine deoxynucleosides: DNA repair, biological effects, mechanistic insights, and unanswered questions. Free Radic. Biol. Med., Oxidative DNA Damage & Repair 107, 90–100. https://doi.org/10.1016/j.freeradbiomed.2016.12.028
Brown, J.A., Fiala, K.A., Fowler, J.D., Sherrer, S.M., Newmister, S.A., Duym, W.W., Suo, Z., 2010. A Novel Mechanism of Sugar Selection Utilized by a Human X-Family DNA Polymerase. J. Mol. Biol. 395, 282–290. https://doi.org/10.1016/j.jmb.2009.11.003
Buchmann, K., 2014. Evolution of Innate Immunity: Clues from Invertebrates via Fish to Mammals. Front. Immunol. 5.
Canela, A., Maman, Y., Jung, S., Wong, N., Callen, E., Day, A., Kieffer-Kwon, K.-R., Pekowska, A., Zhang, H., Rao, S.S.P., Huang, S.-C., Mckinnon, P.J., Aplan, P.D., Pommier, Y., Aiden, E.L., Casellas, R., Nussenzweig, A., 2017. Genome Organization Drives Chromosome Fragility. Cell 170, 507-521.e18. https://doi.org/10.1016/j.cell.2017.06.034
Cannan, W.J., Pederson, D.S., 2016. Mechanisms and Consequences of Double-strand DNA Break Formation in Chromatin. J. Cell. Physiol. 231, 3–14. https://doi.org/10.1002/jcp.25048
Capp, J.-P., Boudsocq, F., Bertrand, P., Laroche-Clary, A., Pourquier, P., Lopez, B.S., Cazaux, C., Hoffmann, J.-S., Canitrot, Y., 2006. The DNA polymerase λ is required for the repair of non-compatible DNA double strand breaks by NHEJ in mammalian cells. Nucleic Acids Res. 34, 2998–3007. https://doi.org/10.1093/nar/gkl380
Castaño, A., Roy, U., Schärer, O.D., 2017. Chapter Sixteen - Preparation of Stable Nitrogen Mustard DNA Interstrand Cross-Link Analogs for Biochemical and Cell Biological Studies, in: Eichman, B.F. (Ed.), Methods in Enzymology, DNA Repair Enzymes: Cell, Molecular, and Chemical Biology. Academic Press, pp. 415–431. https://doi.org/10.1016/bs.mie.2017.03.007
Chagovetz, A.M., Sweasy, J.B., Preston, B.D., 1997. Increased Activity and Fidelity of DNA Polymerase β on Single-nucleotide Gapped DNA *. J. Biol. Chem. 272, 27501–27504. https://doi.org/10.1074/jbc.272.44.27501
Chalker, D.L., Meyer, E., Mochizuki, K., 2013. Epigenetics of Ciliates. Cold Spring Harb. Perspect. Biol. 5, a017764. https://doi.org/10.1101/cshperspect.a017764
Chan, Y.-W., Mohr, R., Millard, A.D., Holmes, A.B., Larkum, A.W., Whitworth, A.L., Mann, N.H., Scanlan, D.J., Hess, W.R., Clokie, M.R.J., 2011. Discovery of cyanophage genomes which contain mitochondrial DNA polymerase. Mol. Biol. Evol. 28, 2269–2274. https://doi.org/10.1093/molbev/msr041
Charmant, O., Gruchota, J., Arnaiz, O., Zangarelli, C., Bétermier, M., Nowak, K., Legros, V., Chevreux, G., Nowak, J., Duharcourt, S., 2023. The nuclear PIWI-interacting protein Gtsf1 controls the selective degradation of small RNAs in Paramecium. https://doi.org/10.1101/2023.09.19.558372
Chen, Q., Luo, W., Veach, R.A., Hickman, A.B., Wilson, M.H., Dyda, F., 2020. Structural basis of seamless excision and specific targeting by piggyBac transposase. Nat. Commun. 11, 3446. https://doi.org/10.1038/s41467-020-17128-1
Chen, S., Lee, L., Naila, T., Fishbain, S., Wang, A., Tomkinson, A.E., Lees-Miller, S.P., He, Y., 2021. Structural basis of long-range to short-range synaptic transition in NHEJ. Nature 593, 294–298. https://doi.org/10.1038/s41586-021-03458-7
Chen, S., Vogt, A., Lee, L., Naila, T., McKeown, R., Tomkinson, A.E., Lees-Miller, S.P., He, Y., 2023. Cryo-EM visualization of DNA-PKcs structural intermediates in NHEJ. Sci. Adv. 9, eadg2838. https://doi.org/10.1126/sciadv.adg2838
Chen, S.-H., Yu, X., 2019. Human DNA ligase IV is able to use NAD+ as an alternative adenylation donor for DNA ends ligation. Nucleic Acids Res. 47, 1321–1334. https://doi.org/10.1093/nar/gky1202
Chen, X., Ballin, J.D., Della-Maria, J., Tsai, M.-S., White, E.J., Tomkinson, A.E., Wilson, G.M., 2009. Distinct kinetics of human DNA ligases I, IIIα, IIIβ, and IV reveal direct DNA sensing ability and differential physiological functions in DNA repair. DNA Repair 8, 961–968. https://doi.org/10.1016/j.dnarep.2009.06.002
Chen, X., Su, S., Chen, Y., Gao, Y., Li, Y., Shao, Z., Zhang, Y., Shao, Q., Liu, H., Li, J., Ma, J., Gan, J., 2020. Structural studies reveal a ring-shaped architecture of deep-sea vent phage NrS-1 polymerase. Nucleic Acids Res. 48, 3343–3355. https://doi.org/10.1093/nar/gkaa071
Chen, X., Tomkinson, A.E., 2011. Yeast Nej1 Is a Key Participant in the Initial End Binding and Final Ligation Steps of Nonhomologous End Joining. J. Biol. Chem. 286, 4931–4940. https://doi.org/10.1074/jbc.M110.195024
Chi, X., Li, Y., Qiu, X., 2020. V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: mechanism and regulation. Immunology 160, 233–247. https://doi.org/10.1111/imm.13176
Chiruvella, K.K., Renard, B.M., Birkeland, S.R., Sunder, S., Liang, Z., Wilson, T.E., 2014. Yeast DNA ligase IV mutations reveal a nonhomologous end joining function of BRCT1 distinct from XRCC4/Lif1 binding. DNA Repair 24, 37–45. https://doi.org/10.1016/j.dnarep.2014.10.003
Cho, J.-E., Jinks-Robertson, S., 2018. Topoisomerase I and Genome Stability: The Good and the Bad. Methods Mol. Biol. Clifton NJ 1703, 21–45. https://doi.org/10.1007/978-1-4939-7459-7_2
Chu, V.T., Weber, T., Wefers, B., Wurst, W., Sander, S., Rajewsky, K., Kühn, R., 2015. Increasing the efficiency

of homology-directed repair for CRISPR-Cas9-induced precise gene editing in mammalian cells. Nat. Biotechnol. 33, 543–548. https://doi.org/10.1038/nbt.3198
Clouaire, T., Legube, G., 2019. A Snapshot on the Cis Chromatin Response to DNA Double-Strand Breaks. Trends Genet. 35, 330–345. https://doi.org/10.1016/j.tig.2019.02.003
Clouaire, T., Rocher, V., Lashgari, A., Arnould, C., Aguirrebengoa, M., Biernacka, A., Skrzypczak, M., Aymard, F., Fongang, B., Dojer, N., Iacovoni, J.S., Rowicka, M., Ginalski, K., Côté, J., Legube, G., 2018. Comprehensive Mapping of Histone Modifications at DNA Double-Strand Breaks Deciphers Repair Pathway Chromatin Signatures. Mol. Cell 72, 250-262.e6. https://doi.org/10.1016/j.molcel.2018.08.020
Cooper, G.M., 2000. Chromosomes and Chromatin, in: The Cell: A Molecular Approach. 2nd Edition. Sinauer Associates.
Cortes Ledesma, F., El Khamisy, S.F., Zuma, M.C., Osborn, K., Caldecott, K.W., 2009. A human 5'-tyrosyl DNA phosphodiesterase that repairs topoisomerase-mediated DNA damage. Nature 461, 674–678. https://doi.org/10.1038/nature08444
Costantini, S., Woodbine, L., Andreoli, L., Jeggo, P.A., Vindigni, A., 2007. Interaction of the Ku heterodimer with the DNA ligase IV/Xrcc4 complex and its regulation by DNA-PK. DNA Repair 6, 712–722. https://doi.org/10.1016/j.dnarep.2006.12.007
Crick, F., 1974. The double helix: a personal view. Nature 248, 766–769. https://doi.org/10.1038/248766a0
Crick, F., 1970. Central Dogma of Molecular Biology. Nature 227, 561–563. https://doi.org/10.1038/227561a0
Crooks, G.E., Hon, G., Chandonia, J.-M., Brenner, S.E., 2004. WebLogo: a sequence logo generator. Genome Res. 14, 1188–1190. https://doi.org/10.1101/gr.849004
Crotty, S., 2015. A brief history of T cell help to B cells. Nat. Rev. Immunol. 15, 185–189. https://doi.org/10.1038/nri3803
Czernecki, D., Nourisson, A., Legrand, P., Delarue, M., 2023. Reclassification of family A DNA polymerases reveals novel functional subfamilies and distinctive structural features. Nucleic Acids Res. gkad242. https://doi.org/10.1093/nar/gkad242
d'Adda di Fagagna, F., Weller, G.R., Doherty, A.J., Jackson, S.P., 2003. The Gam protein of bacteriophage Mu is an orthologue of eukaryotic Ku. EMBO Rep. 4, 47–52. https://doi.org/10.1038/sj.embor.embor709
Dai, X., Zhang, S., Zaleta-Rivera, K., 2020. RNA: interactions drive functionalities. Mol. Biol. Rep. 47, 1413–1434. https://doi.org/10.1007/s11033-019-05230-7
Daly, M.J., Gaidamakova, E.K., Matrosova, V.Y., Vasilenko, A., Zhai, M., Venkateswaran, A., Hess, M., Omelchenko, M.V., Kostandarithes, H.M., Makarova, K.S., Wackett, L.P., Fredrickson, J.K., Ghosal, D., 2004. Accumulation of Mn(II) in Deinococcus radiodurans facilitates gamma-radiation resistance. Science 306, 1025–1028. https://doi.org/10.1126/science.1103185
de Gruijl, F.R., 1999. Skin cancer and solar UV radiation. Eur. J. Cancer Oxf. Engl. 1990 35, 2003–2009. https://doi.org/10.1016/s0959-8049(99)00283-x
Deans, A.J., West, S.C., 2011. DNA interstrand crosslink repair and cancer. Nat. Rev. Cancer 11, 467–480. https://doi.org/10.1038/nrc3088
Delarue, M., 1995. Aminoacyl-tRNA synthetases. Curr. Opin. Struct. Biol. 5, 48–55. https://doi.org/10.1016/0959-440X(95)80008-O
DeRose, E.F., Clarkson, M.W., Gilmore, S.A., Galban, C.J., Tripathy, A., Havener, J.M., Mueller, G.A., Ramsden, D.A., London, R.E., Lee, A.L., 2007. Solution Structure of Polymerase μ's BRCT Domain Reveals an Element Essential for Its Role in Nonhomologous End Joining. Biochemistry 46, 12100–12110. https://doi.org/10.1021/bi7007728
Deshpande, R.A., Wilson, T.E., 2007. Modes of interaction among yeast Nej1, Lif1 and Dnl4 proteins and comparison to human XLF, XRCC4 and Lig4. DNA Repair 6, 1507–1516. https://doi.org/10.1016/j.dnarep.2007.04.014
Di Tommaso, P., Moretti, S., Xenarios, I., Orobitg, M., Montanyola, A., Chang, J.-M., Taly, J.-F., Notredame, C., 2011. T-Coffee: a web server for the multiple sequence alignment of protein and RNA sequences using structural information and homology extension. Nucleic Acids Res. 39, W13–W17. https://doi.org/10.1093/nar/gkr245
Dobell, C., Dobell, C., Leeuwenhoek, A. van, 1932. Antony van Leeuwenhoek and his "Little animals"; being some account of the father of protozoology and bacteriology and his multifarious discoveries in these disciplines. Harcourt, Brace and company, New York. https://doi.org/10.5962/bhl.title.13354
Doherty, A.J., Jackson, S.P., Weller, G.R., 2001. Identification of bacterial homologues of the Ku DNA repair proteins. FEBS Lett. 500, 186–188. https://doi.org/10.1016/S0014-5793(01)02589-3
Domínguez, O., Ruiz, J.F., Laín de Lera, T., García-Díaz, M., González, M.A., Kirchhoff, T., Martínez-A, C., Bernad, A., Blanco, L., 2000. DNA polymerase mu (Pol μ), homologous to TdT, could act as a DNA mutator in eukaryotic cells. EMBO J. 19, 1731–1742. https://doi.org/10.1093/emboj/19.7.1731
Drake, J.W., 1999. The Distribution of Rates of Spontaneous Mutation over Viruses, Prokaryotes, and Eukaryotes. Ann. N. Y. Acad. Sci. 870, 100–107. https://doi.org/10.1111/j.1749-6632.1999.tb08870.x
Dray, E., Etchin, J., Wiese, C., Saro, D., Williams, G.J., Hammel, M., Yu, X., Galkin, V.E., Liu, D., Tsai, M.-S.,

Sy, S.M.-H., Schild, D., Egelman, E., Chen, J., Sung, P., 2010. Enhancement of RAD51 recombinase activity by the tumor suppressor PALB2. Nat. Struct. Mol. Biol. 17, 1255–1259. https://doi.org/10.1038/nsmb.1916
Drews, F., Boenigk, J., Simon, M., 2022. Paramecium epigenetics in development and proliferation. J. Eukaryot. Microbiol. 69, e12914. https://doi.org/10.1111/jeu.12914
Dubois, E., Bischerour, J., Marmignon, A., Mathy, N., Régnier, V., Bétermier, M., 2012. Transposon Invasion of the Paramecium Germline Genome Countered by a Domesticated PiggyBac Transposase and the NHEJ Pathway. Int. J. Evol. Biol. 2012, e436196. https://doi.org/10.1155/2012/436196
Elick, T.A., Bauser, C.A., Fraser, M.J., 1996. Excision of the piggyBac transposable element in vitro is a precise event that is enhanced by the expression of its encoded transposase. Genetica 98, 33–41. https://doi.org/10.1007/BF00120216
Emerson, C.H., Bertuch, A.A., 2016. Consider the workhorse: Nonhomologous end joining in budding yeast. Biochem. Cell Biol. Biochim. Biol. Cell. 94, 396–406. https://doi.org/10.1139/bcb-2016-0001
Evans, P., 2006. Scaling and assessment of data quality. Acta Crystallogr. D Biol. Crystallogr. 62, 72–82. https://doi.org/10.1107/S0907444905036693
Feng, W., Smith, C.M., Simpson, D.A., Gupta, G.P., 2022. Targeting Non-homologous and Alternative End Joining Repair to Enhance Cancer Radiosensitivity. Semin. Radiat. Oncol., Progress Towards Genomically-directed Radiosensitization 32, 29–41. https://doi.org/10.1016/j.semradonc.2021.09.007
Ferraro, P., Franzolin, E., Pontarin, G., Reichard, P., Bianchi, V., 2010. Quantitation of cellular deoxynucleoside triphosphates. Nucleic Acids Res. 38, e85. https://doi.org/10.1093/nar/gkp1141
Fiala, K.A., Abdel-Gawad, W., Suo, Z., 2004. Pre-steady-state kinetic studies of the fidelity and mechanism of polymerization catalyzed by truncated human DNA polymerase lambda. Biochemistry 43, 6751–6762. https://doi.org/10.1021/bi049975c
Fiala, K.A., Duym, W.W., Zhang, J., Suo, Z., 2006. Up-regulation of the Fidelity of Human DNA Polymerase λ by Its Non-enzymatic Proline-rich Domain *. J. Biol. Chem. 281, 19038–19044. https://doi.org/10.1074/jbc.M601178200
Filée, J., Forterre, P., Sen-Lin, T., Laurent, J., 2002. Evolution of DNA polymerase families: evidences for multiple gene exchange between cellular and viral proteins. J. Mol. Evol. 54, 763–773. https://doi.org/10.1007/s00239-001-0078-x
Foley, M.C., Padow, V.A., Schlick, T., 2010. DNA Pol λ's Extraordinary Ability To Stabilize Misaligned DNA. J. Am. Chem. Soc. 132, 13403–13416. https://doi.org/10.1021/ja1049687
Fox, G.E., 2010. Origin and Evolution of the Ribosome. Cold Spring Harb. Perspect. Biol. 2, a003483. https://doi.org/10.1101/cshperspect.a003483
Frank, E.G., Woodgate, R., 2007. Increased Catalytic Activity and Altered Fidelity of Human DNA Polymerase ι in the Presence of Manganese *. J. Biol. Chem. 282, 24689–24696. https://doi.org/10.1074/jbc.M702159200
Freudenthal, B.D., Beard, W.A., Shock, D.D., Wilson, S.H., 2013. Observing a DNA Polymerase Choose Right from Wrong. Cell 154, 157–168. https://doi.org/10.1016/j.cell.2013.05.048
Frickey, T., Lupas, A., 2004. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics 20, 3702–3704. https://doi.org/10.1093/bioinformatics/bth444
Friedberg, E.C., 2003. DNA damage and repair. Nature 421, 436–440. https://doi.org/10.1038/nature01408
Frigerio, C., Di Nisio, E., Galli, M., Colombo, C.V., Negri, R., Clerici, M., 2023. The Chromatin Landscape around DNA Double-Strand Breaks in Yeast and Its Influence on DNA Repair Pathway Choice. Int. J. Mol. Sci. 24, 3248. https://doi.org/10.3390/ijms24043248
Frit, P., Ropars, V., Modesti, M., Charbonnier, J.B., Calsou, P., 2019. Plugged into the Ku-DNA hub: The NHEJ network. Prog. Biophys. Mol. Biol. 147, 62–76. https://doi.org/10.1016/j.pbiomolbio.2019.03.001
Fruchterman, T.M.J., Reingold, E.M., 1991. Graph drawing by force-directed placement. Softw. Pract. Exp. 21, 1129–1164. https://doi.org/10.1002/spe.4380211102
Fukushima, S., Itaya, M., Kato, H., Ogasawara, N., Yoshikawa, H., 2007. Reassessment of the In Vivo Functions of DNA Polymerase I and RNase H in Bacterial Cell Growth. J. Bacteriol. 189, 8575–8583. https://doi.org/10.1128/JB.00653-07
Furrer, D.I., Swart, E.C., Kraft, M.F., Sandoval, P.Y., Nowacki, M., 2017. Two Sets of Piwi Proteins Are Involved in Distinct sRNA Pathways Leading to Elimination of Germline-Specific DNA. Cell Rep. 20, 505–520. https://doi.org/10.1016/j.celrep.2017.06.050
Gabler, F., Nam, S.-Z., Till, S., Mirdita, M., Steinegger, M., Söding, J., Lupas, A.N., Alva, V., 2020. Protein Sequence Analysis Using the MPI Bioinformatics Toolkit. Curr. Protoc. Bioinforma. 72, e108. https://doi.org/10.1002/cpbi.108
Gahlon, H.L., Sturla, S.J., 2019. Determining Steady-State Kinetics of DNA Polymerase Nucleotide Incorporation, in: Shank, N. (Ed.), Non-Natural Nucleic Acids: Methods and Protocols, Methods in Molecular Biology. Springer, New York, NY, pp. 299–311. https://doi.org/10.1007/978-1-4939-9216-4_19
Ganai, R.A., Johansson, E., 2016. DNA Replication-A Matter of Fidelity. Mol. Cell 62, 745–755. https://doi.org/10.1016/j.molcel.2016.05.003

Gao, Y., Yang, W., 2016. Capture of A Third Mg2+ is Essential for Catalyzing DNA Synthesis. Science 352, 1334–1337. https://doi.org/10.1126/science.aad9633
Garcia-Diaz, M., Bebenek, K., Gao, G., Pedersen, L.C., London, R.E., Kunkel, T.A., 2005a. Structure–function studies of DNA polymerase lambda. DNA Repair, The Dale W. Mosbaugh Commemorative DNA Repair Issue 4, 1358–1367. https://doi.org/10.1016/j.dnarep.2005.09.001
Garcia-Diaz, M., Bebenek, K., Krahn, J.M., Blanco, L., Kunkel, T.A., Pedersen, L.C., 2004. A Structural Solution for the DNA Polymerase λ-Dependent Repair of DNA Gaps with Minimal Homology. Mol. Cell 13, 561–572. https://doi.org/10.1016/S1097-2765(04)00061-9
Garcia-Diaz, M., Bebenek, K., Krahn, J.M., Kunkel, T.A., Pedersen, L.C., 2005b. A closed conformation for the Pol λ catalytic cycle. Nat. Struct. Mol. Biol. 12, 97–98. https://doi.org/10.1038/nsmb876
Garcia-Diaz, M., Bebenek, K., Krahn, J.M., Pedersen, L.C., Kunkel, T.A., 2007. Role of the catalytic metal during polymerization by DNA polymerase lambda. DNA Repair 6, 1333–1340. https://doi.org/10.1016/j.dnarep.2007.03.005
García-Díaz, M., Bebenek, K., Kunkel, T.A., Blanco, L., 2001. Identification of an intrinsic 5'-deoxyribose-5-phosphate lyase activity in human DNA polymerase lambda: a possible role in base excision repair. J. Biol. Chem. 276, 34659–34663. https://doi.org/10.1074/jbc.M106336200
García-Díaz, M., Bebenek, K., Sabariegos, R., Domínguez, O., Rodríguez, J., Kirchhoff, T., García-Palomero, E., Picher, A.J., Juárez, R., Ruiz, J.F., Kunkel, T.A., Blanco, L., 2002. DNA Polymerase λ, a Novel DNA Repair Enzyme in Human Cells∗. J. Biol. Chem. 277, 13184–13191. https://doi.org/10.1074/jbc.M111601200
Gehring, K., Leroy, J.-L., Guéron, M., 1993. A tetrameric DNA structure with protonated cytosine-cytosine base pairs. Nature 363, 561–565. https://doi.org/10.1038/363561a0
Ghosh, D., Raghavan, S.C., 2021. 20 years of DNA Polymerase μ, the polymerase that still surprises. FEBS J. 288, 7230–7242. https://doi.org/10.1111/febs.15852
Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., Smith, H.O., 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods 6, 343–345. https://doi.org/10.1038/nmeth.1318
Gilley, D., Blackburn, E.H., 1994. Lack of telomere shortening during senescence in Paramecium. Proc. Natl. Acad. Sci. 91, 1955–1958. https://doi.org/10.1073/pnas.91.5.1955
Godbey, W.T., 2022. Chapter 3 - Proteins, in: Godbey, W.T. (Ed.), Biotechnology and Its Applications (Second Edition). Academic Press, pp. 47–72. https://doi.org/10.1016/B978-0-12-817726-6.00003-4
Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., Ferrin, T.E., 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. Protein Sci. Publ. Protein Soc. 27, 14–25. https://doi.org/10.1002/pro.3235
Gong, C., Bongiorno, P., Martins, A., Stephanou, N.C., Zhu, H., Shuman, S., Glickman, M.S., 2005. Mechanism of nonhomologous end-joining in mycobacteria: a low-fidelity repair system driven by Ku, ligase D and ligase C. Nat. Struct. Mol. Biol. 12, 304–312. https://doi.org/10.1038/nsmb915
Gong, C., Martins, A., Bongiorno, P., Glickman, M., Shuman, S., 2004. Biochemical and genetic analysis of the four DNA ligases of mycobacteria. J. Biol. Chem. 279, 20594–20606. https://doi.org/10.1074/jbc.M401841200
Goodenbour, J.M., Pan, T., 2006. Diversity of tRNA genes in eukaryotes. Nucleic Acids Res. 34, 6137–6146. https://doi.org/10.1093/nar/gkl725
Goodman, M.F., Keener, S., Guidotti, S., Branscomb, E.W., 1983. On the enzymatic basis for mutagenesis by manganese. J. Biol. Chem. 258, 3469–3475. https://doi.org/10.1016/S0021-9258(18)32685-1
Gouge, J., Rosario, S., Romain, F., Beguin, P., Delarue, M., 2013. Structures of Intermediates along the Catalytic Cycle of Terminal Deoxynucleotidyltransferase: Dynamical Aspects of the Two-Metal Ion Mechanism. J. Mol. Biol. 425, 4334–4352. https://doi.org/10.1016/j.jmb.2013.07.009
Gouge, J., Rosario, S., Romain, F., Poitevin, F., Béguin, P., Delarue, M., 2015. Structural basis for a novel mechanism of DNA bridging and alignment in eukaryotic DSB DNA repair. EMBO J. 34, 1126–1142. https://doi.org/10.15252/embj.201489643
Graham, T.G.W., Walter, J.C., Loparo, J.J., 2016. Two-Stage Synapsis of DNA Ends during Non-Homologous End Joining. Mol. Cell 61, 850–858. https://doi.org/10.1016/j.molcel.2016.02.010
Gratias, A., Bétermier, M., 2003. Processing of Double-Strand Breaks Is Involved in the Precise Excision of Paramecium Internal Eliminated Sequences. Mol. Cell. Biol. 23, 7152–7162. https://doi.org/10.1128/MCB.23.20.7152-7162.2003
Gratias, A., Lepère, G., Garnier, O., Rosa, S., Duharcourt, S., Malinsky, S., Meyer, E., Bétermier, M., 2008. Developmentally programmed DNA splicing in Paramecium reveals short-distance crosstalk between DNA cleavage sites. Nucleic Acids Res. 36, 3244–3251. https://doi.org/10.1093/nar/gkn154
Grawunder, U., Zimmer, D., Fugmann, S., Schwarz, K., Lieber, M.R., 1998. DNA ligase IV is essential for V(D)J recombination and DNA double-strand break repair in human precursor lymphocytes. Mol. Cell 2, 477–484. https://doi.org/10.1016/s1097-2765(00)80147-1
Grawunder, Ulf, Zimmer, D., Kulesza, P., Lieber, M.R., 1998. Requirement for an Interaction of XRCC4 with

DNA Ligase IV for Wild-type V(D)J Recombination and DNA Double-strand Break Repairin Vivo *. J. Biol. Chem. 273, 24708–24714. https://doi.org/10.1074/jbc.273.38.24708
Greco, G.E., Matsumoto, Y., Brooks, R.C., Lu, Z., Lieber, M.R., Tomkinson, A.E., 2016. SCR7 is neither a selective nor a potent inhibitor of human DNA ligase IV. DNA Repair 43, 18–23. https://doi.org/10.1016/j.dnarep.2016.04.004
Grob, P., Zhang, T.T., Hannah, R., Yang, H., Hefferin, M.L., Tomkinson, A.E., Nogales, E., 2012. Electron microscopy visualization of DNA–protein complexes formed by Ku and DNA ligase IV. DNA Repair 11, 74–81. https://doi.org/10.1016/j.dnarep.2011.10.023
Gu, J., Lu, H., Tippin, B., Shimazaki, N., Goodman, M.F., Lieber, M.R., 2007a. XRCC4:DNA ligase IV can ligate incompatible DNA ends and can ligate across gaps. EMBO J. 26, 1010–1023. https://doi.org/10.1038/sj.emboj.7601559
Gu, J., Lu, H., Tsai, A.G., Schwarz, K., Lieber, M.R., 2007b. Single-stranded DNA ligation and XLF-stimulated incompatible DNA end ligation by the XRCC4-DNA ligase IV complex: influence of terminal DNA sequence. Nucleic Acids Res. 35, 5755–5762. https://doi.org/10.1093/nar/gkm579
Guérineau, M., Bessa, L., Moriau, S., Lescop, E., Bontems, F., Mathy, N., Guittet, E., Bischerour, J., Bétermier, M., Morellet, N., 2021. The unusual structure of the PiggyMac cysteine-rich domain reveals zinc finger diversity in PiggyBac-related transposases. Mob. DNA 12, 12. https://doi.org/10.1186/s13100-021-00240-4
Guilliam, T.A., Keen, B.A., Brissett, N.C., Doherty, A.J., 2015. Primase-polymerases are a functionally diverse superfamily of replication and repair enzymes. Nucleic Acids Res. 43, 6651–6664. https://doi.org/10.1093/nar/gkv625
Guirouilh-Barbat, J., Huck, S., Bertrand, P., Pirzio, L., Desmaze, C., Sabatier, L., Lopez, B.S., 2004. Impact of the KU80 Pathway on NHEJ-Induced Genome Rearrangements in Mammalian Cells. Mol. Cell 14, 611–623. https://doi.org/10.1016/j.molcel.2004.05.008
Hakem, R., 2008. DNA-damage repair; the good, the bad, and the ugly. EMBO J. 27, 589–605. https://doi.org/10.1038/emboj.2008.15
Hanawalt, P.C., 2020. Tribute to Sam Wilson: Shining a light on base excision DNA repair. DNA Repair, Tribute to Samuel H. Wilson: Shining Light on Base Excision DNA Repair 93, 102933. https://doi.org/10.1016/j.dnarep.2020.102933
Hennequin, C., Giocanti, N., Averbeck, D., Favaudon, V., 1999. La protéine kinase dépendante de l'ADN (DNA-PK), une enzyme clé de la religation des cassures double-brin de l'ADN. Cancer/Radiothérapie 3, 289–295. https://doi.org/10.1016/S1278-3218(99)80070-5
Hernández-Tamayo, R., Oviedo-Bocanegra, L.M., Fritz, G., Graumann, P.L., 2019. Symmetric activity of DNA polymerases at and recruitment of exonuclease ExoR and of PolA to the Bacillus subtilis replication forks. Nucleic Acids Res. 47, 8521–8536. https://doi.org/10.1093/nar/gkz554
Hinkle, D.C., Richardson, C.C., 1975. Bacteriophage T7 deoxyribonucleic acid replication in vitro. Purification and properties of the gene 4 protein of bacteriophage T7. J. Biol. Chem. 250, 5523–5529.
Hoch, S., Schwaber, J., 1996. VH and VL gene elements that encode human antibodies to DNA. Clin. Immunol. Immunopathol. 80, 88–95. https://doi.org/10.1006/clin.1996.0098
Hoeijmakers, J.H., 2001. Genome maintenance mechanisms for preventing cancer. Nature 411, 366–374. https://doi.org/10.1038/35077232
Hoff, G., Bertrand, C., Zhang, L., Piotrowski, E., Chipot, L., Bontemps, C., Confalonieri, F., McGovern, S., Lecointe, F., Thibessard, A., Leblond, P., 2016. Multiple and Variable NHEJ-Like Genes Are Involved in Resistance to DNA Damage in Streptomyces ambofaciens. Front. Microbiol. 7, 1901. https://doi.org/10.3389/fmicb.2016.01901
Hossain, M.A., Lin, Y., Yan, S., 2018. Single-Strand Break End Resection in Genome Integrity: Mechanism and Regulation by APE2. Int. J. Mol. Sci. 19, 2389. https://doi.org/10.3390/ijms19082389
Hsieh, P., Zhang, Y., 2017. The Devil is in the details for DNA mismatch repair. Proc. Natl. Acad. Sci. 114, 3552–3554. https://doi.org/10.1073/pnas.1702747114
Hustedt, N., Durocher, D., 2017. The control of DNA repair by the cell cycle. Nat. Cell Biol. 19, 1–9. https://doi.org/10.1038/ncb3452
Iacovoni, J.S., Caron, P., Lassadi, I., Nicolas, E., Massip, L., Trouche, D., Legube, G., 2010. High-resolution profiling of γH2AX around DNA double strand breaks in the mammalian genome. EMBO J. 29, 1446–1457. https://doi.org/10.1038/emboj.2010.38
Ille, A.M., Lamont, H., Mathews, M.B., 2022. The Central Dogma revisited: Insights from protein synthesis, CRISPR, and beyond. WIREs RNA 13, e1718. https://doi.org/10.1002/wrna.1718
Interthal, H., Chen, H.J., Champoux, J.J., 2005. Human Tdp1 Cleaves a Broad Spectrum of Substrates Including Phosphoamide Linkages. J. Biol. Chem. 280, 36518–36528. https://doi.org/10.1074/jbc.M508898200
Ishida, N., Nakagawa, T., Iemura, S.-I., Yasui, A., Shima, H., Katoh, Y., Nagasawa, Y., Natsume, T., Igarashi, K., Nakayama, K., 2017. Ubiquitylation of Ku80 by RNF126 Promotes Completion of Nonhomologous End Joining-Mediated DNA Repair. Mol. Cell. Biol. 37, e00347-16. https://doi.org/10.1128/MCB.00347-16

Jackson, S.E., 2013. Hsp90: Structure and Function, in: Jackson, S. (Ed.), Molecular Chaperones, Topics in Current Chemistry. Springer, Berlin, Heidelberg, pp. 155–240. https://doi.org/10.1007/128_2012_356
Jamsen, J.A., Shock, D.D., Wilson, S.H., 2022. Watching right and wrong nucleotide insertion captures hidden polymerase fidelity checkpoints. Nat. Commun. 13, 3193. https://doi.org/10.1038/s41467-022-30141-w
Jensen, R.B., Carreira, A., Kowalczykowski, S.C., 2010. Purified human BRCA2 stimulates RAD51-mediated recombination. Nature 467, 678–683. https://doi.org/10.1038/nature09399
Johnson, K.A., 2008. Role of Induced Fit in Enzyme Specificity: A Molecular Forward/Reverse Switch. J. Biol. Chem. 283, 26297–26301. https://doi.org/10.1074/jbc.R800034200
Johnson, K.A., Goody, R.S., 2011. The Original Michaelis Constant: Translation of the 1913 Michaelis-Menten Paper. Biochemistry 50, 8264–8269. https://doi.org/10.1021/bi201284u
Juárez, R., Ruiz, J.F., McElhinny, S.A.N., Ramsden, D., Blanco, L., 2006. A specific loop in human DNA polymerase mu allows switching between creative and DNA-instructed synthesis. Nucleic Acids Res. 34, 4572–4582. https://doi.org/10.1093/nar/gkl457
Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S.A.A., Ballard, A.J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., Back, T., Petersen, S., Reiman, D., Clancy, E., Zielinski, M., Steinegger, M., Pacholska, M., Berghammer, T., Bodenstein, S., Silver, D., Vinyals, O., Senior, A.W., Kavukcuoglu, K., Kohli, P., Hassabis, D., 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596, 583–589. https://doi.org/10.1038/s41586-021-03819-2
Kabsch, W., 2010. XDS. Acta Crystallogr. D Biol. Crystallogr. 66, 125–132. https://doi.org/10.1107/S0907444909047337
Kapusta, A., Matsuda, A., Marmignon, A., Ku, M., Silve, A., Meyer, E., Forney, J.D., Malinsky, S., Bétermier, M., 2011. Highly Precise and Developmentally Programmed Genome Assembly in Paramecium Requires Ligase IV–Dependent End Joining. PLOS Genet. 7, e1002049. https://doi.org/10.1371/journal.pgen.1002049
Karanam, K., Kafri, R., Loewer, A., Lahav, G., 2012. Quantitative Live Cell Imaging Reveals a Gradual Shift between DNA Repair Mechanisms and a Maximal Use of HR in Mid S Phase. Mol. Cell 47, 320–329. https://doi.org/10.1016/j.molcel.2012.05.052
Kato, K., Gonalves, J.M., Houts, G.E., Bollum, F.J., 1967. Deoxynucleotide-polymerizing Enzymes of Calf Thymus Gland: II. PROPERTIES OF THE TERMINAL DEOXYNUCLEOTIDYLTRANSFERASE. J. Biol. Chem. 242, 2780–2789. https://doi.org/10.1016/S0021-9258(18)99635-3
Katoh, K., Rozewicki, J., Yamada, K.D., 2019. MAFFT online service: multiple sequence alignment, interactive sequence choice and visualization. Brief. Bioinform. 20, 1160–1166. https://doi.org/10.1093/bib/bbx108
Kazlauskas, D., Krupovic, M., Guglielmini, J., Forterre, P., Venclovas, Č., 2020. Diversity and evolution of B-family DNA polymerases. Nucleic Acids Res. 48, 10142–10156. https://doi.org/10.1093/nar/gkaa760
Kazlauskas, D., Sezonov, G., Charpin, N., Venclovas, Č., Forterre, P., Krupovic, M., 2018. Novel Families of Archaeo-Eukaryotic Primases Associated with Mobile Genetic Elements of Bacteria and Archaea. J. Mol. Biol. 430, 737–750. https://doi.org/10.1016/j.jmb.2017.11.014
Keijzers, G., Bohr, V.A., Rasmussen, L.J., 2015. Human exonuclease 1 (EXO1) activity characterization and its function on flap structures. Biosci. Rep. 35, e00206. https://doi.org/10.1042/BSR20150058
Khoury, G.A., Baliban, R.C., Floudas, C.A., 2011. Proteome-wide post-translational modification statistics: frequency analysis and curation of the swiss-prot database. Sci. Rep. 1, srep00090. https://doi.org/10.1038/srep00090
Kirby, T.W., DeRose, E.F., Cavanaugh, N.A., Beard, W.A., Shock, D.D., Mueller, G.A., Wilson, S.H., London, R.E., 2012. Metal-induced DNA translocation leads to DNA polymerase conformational activation. Nucleic Acids Res. 40, 2974–2983. https://doi.org/10.1093/nar/gkr1218
Klenow, H., Henningsen, I., 1970. Selective elimination of the exonuclease activity of the deoxyribonucleic acid polymerase from Escherichia coli B by limited proteolysis. Proc. Natl. Acad. Sci. U. S. A. 65, 168–175. https://doi.org/10.1073/pnas.65.1.168
Klobutcher, L.A., Herrick, G., 1997. Developmental Genome Reorganization in Ciliated Protozoa: The Transposon Link, in: Cohn, W.E., Moldave, K. (Eds.), Progress in Nucleic Acid Research and Molecular Biology. Academic Press, pp. 1–62. https://doi.org/10.1016/S0079-6603(08)61001-6
Krokan, H.E., Bjørås, M., 2013. Base Excision Repair. Cold Spring Harb. Perspect. Biol. 5, a012583. https://doi.org/10.1101/cshperspect.a012583
Kuraku, S., Zmasek, C.M., Nishimura, O., Katoh, K., 2013. aLeaves facilitates on-demand exploration of metazoan gene family trees on MAFFT sequence alignment server with enhanced interactivity. Nucleic Acids Res. 41, W22–W28. https://doi.org/10.1093/nar/gkt389
Kuznetsova, A.A., Tyugashev, T.E., Alekseeva, I.V., Timofeyeva, N.A., Fedorova, O.S., Kuznetsov, N.A., 2022. Insight into the mechanism of DNA synthesis by human terminal deoxynucleotidyltransferase. Life Sci. Alliance 5. https://doi.org/10.26508/lsa.202201428
Laemmli, U.K., 1970. Cleavage of Structural Proteins during the Assembly of the Head of Bacteriophage T4.

Nature 227, 680–685. https://doi.org/10.1038/227680a0
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., Funke, R., Gage, D., Harris, K., Heaford, A., Howland, J., Kann, L., Lehoczky, J., LeVine, R., McEwan, P., McKernan, K., Meldrim, J., Mesirov, J.P., Miranda, C., Morris, W., Naylor, J., Raymond, C., Rosetti, M., Santos, R., Sheridan, A., Sougnez, C., Stange-Thomann, Y., Stojanovic, N., Subramanian, A., Wyman, D., Rogers, J., Sulston, J., Ainscough, R., Beck, S., Bentley, D., Burton, J., Clee, C., Carter, N., Coulson, A., Deadman, R., Deloukas, P., Dunham, A., Dunham, I., Durbin, R., French, L., Grafham, D., Gregory, S., Hubbard, T., Humphray, S., Hunt, A., Jones, M., Lloyd, C., McMurray, A., Matthews, L., Mercer, S., Milne, S., Mullikin, J.C., Mungall, A., Plumb, R., Ross, M., Shownkeen, R., Sims, S., Waterston, R.H., Wilson, R.K., Hillier, L.W., McPherson, J.D., Marra, M.A., Mardis, E.R., Fulton, L.A., Chinwalla, A.T., Pepin, K.H., Gish, W.R., Chissoe, S.L., Wendl, M.C., Delehaunty, K.D., Miner, T.L., Delehaunty, A., Kramer, J.B., Cook, L.L., Fulton, R.S., Johnson, D.L., Minx, P.J., Clifton, S.W., Hawkins, T., Branscomb, E., Predki, P., Richardson, P., Wenning, S., Slezak, T., Doggett, N., Cheng, J.F., Olsen, A., Lucas, S., Elkin, C., Uberbacher, E., Frazier, M., Gibbs, R.A., Muzny, D.M., Scherer, S.E., Bouck, J.B., Sodergren, E.J., Worley, K.C., Rives, C.M., Gorrell, J.H., Metzker, M.L., Naylor, S.L., Kucherlapati, R.S., Nelson, D.L., Weinstock, G.M., Sakaki, Y., Fujiyama, A., Hattori, M., Yada, T., Toyoda, A., Itoh, T., Kawagoe, C., Watanabe, H., Totoki, Y., Taylor, T., Weissenbach, J., Heilig, R., Saurin, W., Artiguenave, F., Brottier, P., Bruls, T., Pelletier, E., Robert, C., Wincker, P., Smith, D.R., Doucette-Stamm, L., Rubenfield, M., Weinstock, K., Lee, H.M., Dubois, J., Rosenthal, A., Platzer, M., Nyakatura, G., Taudien, S., Rump, A., Yang, H., Yu, J., Wang, J., Huang, G., Gu, J., Hood, L., Rowen, L., Madan, A., Qin, S., Davis, R.W., Federspiel, N.A., Abola, A.P., Proctor, M.J., Myers, R.M., Schmutz, J., Dickson, M., Grimwood, J., Cox, D.R., Olson, M.V., Kaul, R., Raymond, C., Shimizu, N., Kawasaki, K., Minoshima, S., Evans, G.A., Athanasiou, M., Schultz, R., Roe, B.A., Chen, F., Pan, H., Ramser, J., Lehrach, H., Reinhardt, R., McCombie, W.R., de la Bastide, M., Dedhia, N., Blöcker, H., Hornischer, K., Nordsiek, G., Agarwala, R., Aravind, L., Bailey, J.A., Bateman, A., Batzoglou, S., Birney, E., Bork, P., Brown, D.G., Burge, C.B., Cerutti, L., Chen, H.C., Church, D., Clamp, M., Copley, R.R., Doerks, T., Eddy, S.R., Eichler, E.E., Furey, T.S., Galagan, J., Gilbert, J.G., Harmon, C., Hayashizaki, Y., Haussler, D., Hermjakob, H., Hokamp, K., Jang, W., Johnson, L.S., Jones, T.A., Kasif, S., Kaspryzk, A., Kennedy, S., Kent, W.J., Kitts, P., Koonin, E.V., Korf, I., Kulp, D., Lancet, D., Lowe, T.M., McLysaght, A., Mikkelsen, T., Moran, J.V., Mulder, N., Pollara, V.J., Ponting, C.P., Schuler, G., Schultz, J., Slater, G., Smit, A.F., Stupka, E., Szustakowki, J., Thierry-Mieg, D., Thierry-Mieg, J., Wagner, L., Wallis, J., Wheeler, R., Williams, A., Wolf, Y.I., Wolfe, K.H., Yang, S.P., Yeh, R.F., Collins, F., Guyer, M.S., Peterson, J., Felsenfeld, A., Wetterstrand, K.A., Patrinos, A., Morgan, M.J., de Jong, P., Catanese, J.J., Osoegawa, K., Shizuya, H., Choi, S., Chen, Y.J., Szustakowki, J., International Human Genome Sequencing Consortium, 2001. Initial sequencing and analysis of the human genome. Nature 409, 860–921. https://doi.org/10.1038/35057062
Lecointe, F., Shevelev, I.V., Bailone, A., Sommer, S., Hübscher, U., 2004. Involvement of an X family DNA polymerase in double-stranded break repair in the radioresistant organism Deinococcus radiodurans. Mol. Microbiol. 53, 1721–1730. https://doi.org/10.1111/j.1365-2958.2004.04233.x
Lee, J.W., Blanco, L., Zhou, T., Garcia-Diaz, M., Bebenek, K., Kunkel, T.A., Wang, Z., Povirk, L.F., 2004. Implication of DNA polymerase lambda in alignment-based gap filling for nonhomologous DNA end joining in human nuclear extracts. J. Biol. Chem. 279, 805–811. https://doi.org/10.1074/jbc.M307913200
Lee, K.-J., Saha, J., Sun, J., Fattah, K.R., Wang, S.-C., Jakob, B., Chi, L., Wang, S.-Y., Taucher-Scholz, G., Davis, A.J., Chen, D.J., 2016. Phosphorylation of Ku dictates DNA double-strand break (DSB) repair pathway choice in S phase. Nucleic Acids Res. 44, 1732–1745. https://doi.org/10.1093/nar/gkv1499
Lehman, I.R., Bessman, M.J., Simms, E.S., Kornberg, A., 1958. Enzymatic synthesis of deoxyribonucleic acid. I. Preparation of substrates and partial purification of an enzyme from Escherichia coli. J. Biol. Chem. 233, 163–170.
Letunic, I., Bork, P., 2021. Interactive Tree Of Life (iTOL) v5: an online tool for phylogenetic tree display and annotation. Nucleic Acids Res. 49, W293–W296. https://doi.org/10.1093/nar/gkab301
Leung, C.C.Y., Glover, J.N.M., 2011. BRCT domains. Cell Cycle 10, 2461–2470. https://doi.org/10.4161/cc.10.15.16312
Li, X., Stith, C.M., Burgers, P.M., Heyer, W.-D., 2009. PCNA Is Required for Initiation of Recombination-Associated DNA Synthesis by DNA Polymerase δ. Mol. Cell 36, 704–713. https://doi.org/10.1016/j.molcel.2009.09.036
Li, Y., Chirgadze, D.Y., Bolanos-Garcia, V.M., Sibanda, B.L., Davies, O.R., Ahnesorg, P., Jackson, S.P., Blundell, T.L., 2008. Crystal structure of human XLF/Cernunnos reveals unexpected differences from XRCC4 with implications for NHEJ. EMBO J. 27, 290–300. https://doi.org/10.1038/sj.emboj.7601942
Li, Z., Wen, J., Lin, Y., Wang, S., Xue, P., Zhang, Z., Zhou, Y., Wang, X., Sui, L., Bi, L.-J., Zhang, X.-E., 2011. A Sir2-Like Protein Participates in Mycobacterial NHEJ. PLoS ONE 6, e20045. https://doi.org/10.1371/journal.pone.0020045
Liang, F., Romanienko, P.J., Weaver, D.T., Jeggo, P.A., Jasin, M., 1996. Chromosomal double-strand break repair

in Ku80-deficient cells. Proc. Natl. Acad. Sci. 93, 8929–8933. https://doi.org/10.1073/pnas.93.17.8929
Lindner, S.E., Llinás, M., Keck, J.L., Kappe, S.H.I., 2011. The primase domain of PfPrex is a proteolytically matured, essential enzyme of the apicoplast. Mol. Biochem. Parasitol. 180, 69–75. https://doi.org/10.1016/j.molbiopara.2011.08.002
Loc'h, J., Delarue, M., 2018. Terminal deoxynucleotidyltransferase: the story of an untemplated DNA polymerase capable of DNA bridging and templated synthesis across strands. Curr. Opin. Struct. Biol., Protein–nucleic acid interactions • Catalysis and regulation 53, 22–31. https://doi.org/10.1016/j.sbi.2018.03.019
Loc'h, J., Gerodimos, C.A., Rosario, S., Tekpinar, M., Lieber, M.R., Delarue, M., 2019. Structural evidence for an in trans base selection mechanism involving Loop1 in polymerase μ at an NHEJ double-strand break junction. J. Biol. Chem. 294, 10579–10595. https://doi.org/10.1074/jbc.RA119.008739
Loc'h, J., Rosario, S., Delarue, M., 2016. Structural Basis for a New Templated Activity by Terminal Deoxynucleotidyl Transferase: Implications for V(D)J Recombination. Structure 24, 1452–1463. https://doi.org/10.1016/j.str.2016.06.014
Lopez, M.J., Mohiuddin, S.S., 2023. Biochemistry, Essential Amino Acids, in: StatPearls. StatPearls Publishing, Treasure Island (FL).
Lopez-Martinez, D., Liang, C.-C., Cohn, M.A., 2016. Cellular response to DNA interstrand crosslinks: the Fanconi anemia pathway. Cell. Mol. Life Sci. CMLS 73, 3097–3114. https://doi.org/10.1007/s00018-016-2218-x
Ma, Y., Pannicke, U., Schwarz, K., Lieber, M.R., 2002. Hairpin Opening and Overhang Processing by an Artemis/DNA-Dependent Protein Kinase Complex in Nonhomologous End Joining and V(D)J Recombination. Cell 108, 781–794. https://doi.org/10.1016/S0092-8674(02)00671-2
Madru, C., Henneke, G., Raia, P., Hugonneau-Beaufet, I., Pehau-Arnaudet, G., England, P., Lindahl, E., Delarue, M., Carroni, M., Sauguet, L., 2020. Structural basis for the increased processivity of D-family DNA polymerases in complex with PCNA. Nat. Commun. 11, 1591. https://doi.org/10.1038/s41467-020-15392-9
Makiela-Dzbenska, K., Jaszczur, M., Banach-Orlowska, M., Jonczyk, P., Schaaper, R.M., Fijalkowska, I.J., 2009. Role of Escherichia coli DNA polymerase I in chromosomal DNA replication fidelity. Mol. Microbiol. 74, 1114–1127. https://doi.org/10.1111/j.1365-2958.2009.06921.x
Mari, P.-O., Florea, B.I., Persengiev, S.P., Verkaik, N.S., Brüggenwirth, H.T., Modesti, M., Giglia-Mari, G., Bezstarosti, K., Demmers, J.A.A., Luider, T.M., Houtsmuller, A.B., van Gent, D.C., 2006. Dynamic assembly of end-joining complexes requires interaction between Ku70/80 and XRCC4. Proc. Natl. Acad. Sci. 103, 18597–18602. https://doi.org/10.1073/pnas.0609061103
Mariani, V., Biasini, M., Barbato, A., Schwede, T., 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. Bioinformatics 29, 2722–2728. https://doi.org/10.1093/bioinformatics/btt473
Marmignon, A., Bischerour, J., Silve, A., Fojcik, C., Dubois, E., Arnaiz, O., Kapusta, A., Malinsky, S., Bétermier, M., 2014. Ku-Mediated Coupling of DNA Cleavage and Repair during Programmed Genome Rearrangements in the Ciliate Paramecium tetraurelia. PLOS Genet. 10, e1004552. https://doi.org/10.1371/journal.pgen.1004552
Marnef, A., Cohen, S., Legube, G., 2017. Transcription-Coupled DNA Double-Strand Break Repair: Active Genes Need Special Care. J. Mol. Biol. 429, 1277–1288. https://doi.org/10.1016/j.jmb.2017.03.024
Marnef, A., Legube, G., 2017. Organizing DNA repair in the nucleus: DSBs hit the road. Curr. Opin. Cell Biol., Cell Nucleus 46, 1–8. https://doi.org/10.1016/j.ceb.2016.12.003
Martin, A., Chahwan, R., Parsa, J.Y., Scharff, M.D., 2015. Chapter 20 - Somatic Hypermutation: The Molecular Mechanisms Underlying the Production of Effective High-Affinity Antibodies, in: Alt, F.W., Honjo, T., Radbruch, A., Reth, M. (Eds.), Molecular Biology of B Cells (Second Edition). Academic Press, London, pp. 363–388. https://doi.org/10.1016/B978-0-12-397933-9.00020-5
Martin, M.J., Garcia-Ortiz, M.V., Esteban, V., Blanco, L., 2013. Ribonucleotides and manganese ions improve non-homologous end joining by human Polμ. Nucleic Acids Res. 41, 2428–2436. https://doi.org/10.1093/nar/gks1444
Maruyama, T., Dougan, S.K., Truttmann, M., Bilate, A.M., Ingram, J.R., Ploegh, H.L., 2015. Inhibition of non-homologous end joining increases the efficiency of CRISPR/Cas9-mediated precise [TM: inserted] genome editing. Nat. Biotechnol. 33, 538–542. https://doi.org/10.1038/nbt.3190
Matsukage, A., Bohn, E.W., Wilson, S.H., 1974. Multiple Forms of DNA Polymerase in Mouse Myeloma. Proc. Natl. Acad. Sci. 71, 578–582. https://doi.org/10.1073/pnas.71.2.578
Maurer-Alcalá, X.X., Knight, R., Katz, L.A., 2018. Exploration of the Germline Genome of the Ciliate Chilodonella uncinata through Single-Cell Omics (Transcriptomics and Genomics). mBio 9, 10.1128/mbio.01836-17. https://doi.org/10.1128/mbio.01836-17
Maurer-Alcalá, X.X., Nowacki, M., 2019. Evolutionary origins and impacts of genome architecture in ciliates. Ann. N. Y. Acad. Sci. 1447, 110–118. https://doi.org/10.1111/nyas.14108
McInnis, M., O'Neill, G., Fossum, K., Reagan, M.S., 2002. Epistatic analysis of the roles of the RAD27 and POL4 gene products in DNA base excision repair in S. cerevisiae. DNA Repair 1, 311–315. https://doi.org/10.1016/S1568-7864(02)00007-1

McVey, M., Adams, M., Staeva-Vieira, E., Sekelsky, J.J., 2004. Evidence for Multiple Cycles of Strand Invasion During Repair of Double-Strand Gaps in Drosophila. Genetics 167, 699–705. https://doi.org/10.1534/genetics.103.025411
McVey, M., Khodaverdian, V.Y., Meyer, D., Cerqueira, P.G., Heyer, W.-D., 2016. Eukaryotic DNA Polymerases in Homologous Recombination. Annu. Rev. Genet. 50, 393–421. https://doi.org/10.1146/annurev-genet-120215-035243
Menon, V., Povirk, L.F., 2016. End-processing nucleases and phosphodiesterases: An elite supporting cast for the non-homologous end joining pathway of DNA double-strand break repair. DNA Repair 43, 57–68. https://doi.org/10.1016/j.dnarep.2016.05.011
Mimitou, E.P., Symington, L.S., 2010. Ku prevents Exo1 and Sgs1-dependent resection of DNA ends in the absence of a functional MRX complex or Sae2. EMBO J. 29, 3358–3369. https://doi.org/10.1038/emboj.2010.193
Mimori, T., Akizuki, M., Yamagata, H., Inada, S., Yoshida, S., Homma, M., 1981. Characterization of a high molecular weight acidic nuclear protein recognized by autoantibodies in sera from patients with polymyositis-scleroderma overlap. J. Clin. Invest. 68, 611–620. https://doi.org/10.1172/jci110295
Minchin, S., Lodge, J., 2019. Understanding biochemistry: structure and function of nucleic acids. Essays Biochem. 63, 433–456. https://doi.org/10.1042/EBC20180038
Miyaki, M., Murata, I., Osabe, M., Ono, T., 1977. Effect of metal cations on misincorporation by E. coli DNA polymerases. Biochem. Biophys. Res. Commun. 77, 854–860. https://doi.org/10.1016/S0006-291X(77)80056-9
Modesti, M., Budzowska, M., Baldeyron, C., Demmers, J.A.A., Ghirlando, R., Kanaar, R., 2007. RAD51AP1 Is a Structure-Specific DNA Binding Protein that Stimulates Joint Molecule Formation during RAD51-Mediated Homologous Recombination. Mol. Cell 28, 468–481. https://doi.org/10.1016/j.molcel.2007.08.025
Mojumdar, A., Adam, N., Cobb, J.A., 2022. Multifunctional properties of Nej1XLF C-terminus promote end-joining and impact DNA double-strand break repair pathway choice. DNA Repair 115, 103332. https://doi.org/10.1016/j.dnarep.2022.103332
Moon, A.F., Garcia-Diaz, M., Bebenek, K., Davis, B.J., Zhong, X., Ramsden, D.A., Kunkel, T.A., Pedersen, L.C., 2007. Structural insight into the substrate specificity of DNA Polymerase μ. Nat. Struct. Mol. Biol. 14, 45–53. https://doi.org/10.1038/nsmb1180
Moon, A.F., Pryor, J.M., Ramsden, D.A., Kunkel, T.A., Bebenek, K., Pedersen, L.C., 2014. Sustained active site rigidity during synthesis by human DNA polymerase μ. Nat. Struct. Mol. Biol. 21, 253–260. https://doi.org/10.1038/nsmb.2766
Moon, J., Kitty, I., Renata, K., Qin, S., Zhao, F., Kim, W., 2023. DNA Damage and Its Role in Cancer Therapeutics. Int. J. Mol. Sci. 24, 4741. https://doi.org/10.3390/ijms24054741
Moran, J.V., Wilson, T.E., 2022. Reverse transcriptase meets DNA, again: Possible roles for transposable elements in host DNA repair. Cell 185, 3643–3645. https://doi.org/10.1016/j.cell.2022.09.012
Moscato, B., Swain, M., Loria, J.P., 2016. Induced Fit in the Selection of Correct versus Incorrect Nucleotides by DNA Polymerase β. Biochemistry 55, 382–395. https://doi.org/10.1021/acs.biochem.5b01213
Moser, M.J., DiFrancesco, R.A., Gowda, K., Klingele, A.J., Sugar, D.R., Stocki, S., Mead, D.A., Schoenfeld, T.W., 2012. Thermostable DNA polymerase from a viral metagenome is a potent RT-PCR enzyme. PloS One 7, e38371. https://doi.org/10.1371/journal.pone.0038371
Nakane, S., Nakagawa, N., Kuramitsu, S., Masui, R., 2012a. The role of the PHP domain associated with DNA polymerase X from Thermus thermophilus HB8 in base excision repair. DNA Repair 11, 906–914. https://doi.org/10.1016/j.dnarep.2012.09.001
Nakane, S., Nakagawa, N., Kuramitsu, S., Masui, R., 2012b. The role of the PHP domain associated with DNA polymerase X from Thermus thermophilus HB8 in base excision repair. DNA Repair 11, 906–914. https://doi.org/10.1016/j.dnarep.2012.09.001
Nakane, S., Nakagawa, N., Kuramitsu, S., Masui, R., 2009. Characterization of DNA polymerase X from Thermus thermophilus HB8 reveals the POLXc and PHP domains are both required for 3′–5′ exonuclease activity. Nucleic Acids Res. 37, 2037–2052. https://doi.org/10.1093/nar/gkp064
Nelson, D.L., Lehninger, A.L., Cox, M.M., 2008. Lehninger Principles of Biochemistry. W. H. Freeman.
Nelson-Sathi, S., Sousa, F.L., Roettger, M., Lozada-Chávez, N., Thiergart, T., Janssen, A., Bryant, D., Landan, G., Schönheit, P., Siebers, B., McInerney, J.O., Martin, W.F., 2015. Origins of major archaeal clades correspond to gene acquisitions from bacteria. Nature 517, 77–80. https://doi.org/10.1038/nature13805
New, J.H., Sugiyama, T., Zaitseva, E., Kowalczykowski, S.C., 1998. Rad52 protein stimulates DNA strand exchange by Rad51 and replication protein A. Nature 391, 407–410. https://doi.org/10.1038/34950
Newman, J.A., Cooper, C.D.O., Aitkenhead, H., Gileadi, O., 2015. Structure of the Helicase Domain of DNA Polymerase Theta Reveals a Possible Role in the Microhomology-Mediated End-Joining Pathway. Struct. England1993 23, 2319–2330. https://doi.org/10.1016/j.str.2015.10.014
Nichols, M.H., Corces, V.G., 2021. Principles of 3D compartmentalization of the human genome. Cell Rep. 35, 109330. https://doi.org/10.1016/j.celrep.2021.109330

Nick McElhinny, S.A., Havener, J.M., Garcia-Diaz, M., Juárez, R., Bebenek, K., Kee, B.L., Blanco, L., Kunkel, T.A., Ramsden, D.A., 2005. A Gradient of Template Dependence Defines Distinct Biological Roles for Family X Polymerases in Nonhomologous End Joining. Mol. Cell 19, 357–366. https://doi.org/10.1016/j.molcel.2005.06.012
Nick McElhinny, S.A., Ramsden, D.A., 2003. Polymerase mu is a DNA-directed DNA/RNA polymerase. Mol. Cell. Biol. 23, 2309–2315. https://doi.org/10.1128/MCB.23.7.2309-2315.2003
Nowacki, M., Zagorski-Ostoja, W., Meyer, E., 2005. Nowa1p and Nowa2p: novel putative RNA binding proteins involved in trans-nuclear crosstalk in Paramecium tetraurelia. Curr. Biol. CB 15, 1616–1628. https://doi.org/10.1016/j.cub.2005.07.033
Nuin, P.A.S., Wang, Z., Tillier, E.R.M., 2006. The accuracy of several multiple sequence alignment programs for proteins. BMC Bioinformatics 7, 471. https://doi.org/10.1186/1471-2105-7-471
Ochi, T., Blackford, A.N., Coates, J., Jhujh, S., Mehmood, S., Tamura, N., Travers, J., Wu, Q., Draviam, V.M., Robinson, C.V., Blundell, T.L., Jackson, S.P., 2015. PAXX, a paralog of XRCC4 and XLF, interacts with Ku to promote DNA double-strand break repair. Science 347, 185–188. https://doi.org/10.1126/science.1261971
Okazaki, R., Arisawa, M., Sugino, A., 1971. Slow joining of newly replicated DNA chains in DNA polymerase I-deficient Escherichia coli mutants. Proc. Natl. Acad. Sci. U. S. A. 68, 2954–2957. https://doi.org/10.1073/pnas.68.12.2954
Ollis, D.L., Brick, P., Hamlin, R., Xuong, N.G., Steitz, T.A., 1985. Structure of large fragment of Escherichia coli DNA polymerase I complexed with dTMP. Nature 313, 762–766. https://doi.org/10.1038/313762a0
Onuchic, J.N., Luthey-Schulten, Z., Wolynes, P.G., 1997. THEORY OF PROTEIN FOLDING: The Energy Landscape Perspective. Annu. Rev. Phys. Chem. 48, 545–600. https://doi.org/10.1146/annurev.physchem.48.1.545
Orengo, C.A., Todd, A.E., Thornton, J.M., 1999. From protein structure to function. Curr. Opin. Struct. Biol. 9, 374–382. https://doi.org/10.1016/S0959-440X(99)80051-7
Pannunzio, N.R., Li, S., Watanabe, G., Lieber, M.R., 2014. NHEJ Often Uses Microhomology: Implications for Alternative End Joining. DNA Repair 17, 74–80. https://doi.org/10.1016/j.dnarep.2014.02.006
Park, J., Baruch-Torres, N., Iwai, S., Herrmann, G.K., Brieba, L.G., Yin, Y.W., 2022. Human Mitochondrial DNA Polymerase Metal Dependent UV Lesion Bypassing Ability. Front. Mol. Biosci. 9.
Pascal, J.M., 2008. DNA and RNA ligases: structural variations and shared mechanisms. Curr. Opin. Struct. Biol. 18, 96–105. https://doi.org/10.1016/j.sbi.2007.12.008
Pellegrini, L., 2023. The CMG DNA helicase and the core replisome. Curr. Opin. Struct. Biol. 81, 102612. https://doi.org/10.1016/j.sbi.2023.102612
Pervez, M.T., Babar, M.E., Nadeem, A., Aslam, M., Awan, A.R., Aslam, N., Hussain, T., Naveed, N., Qadri, S., Waheed, U., Shoaib, M., 2014. Evaluating the accuracy and efficiency of multiple sequence alignment methods. Evol. Bioinforma. Online 10, 205–217. https://doi.org/10.4137/EBO.S19199
Pettersen, E.F., Goddard, T.D., Huang, C.C., Meng, E.C., Couch, G.S., Croll, T.I., Morris, J.H., Ferrin, T.E., 2021. UCSF ChimeraX: Structure visualization for researchers, educators, and developers. Protein Sci. Publ. Protein Soc. 30, 70–82. https://doi.org/10.1002/pro.3943
Pignataro, M.F., Herrera, M.G., Dodero, V.I., 2020. Evaluation of Peptide/Protein Self-Assembly and Aggregation by Spectroscopic Methods. Molecules 25, 4854. https://doi.org/10.3390/molecules25204854
Pitcher, R.S., Brissett, N.C., Doherty, A.J., 2007a. Nonhomologous End-Joining in Bacteria: A Microbial Perspective. Annu. Rev. Microbiol. 61, 259–282. https://doi.org/10.1146/annurev.micro.61.080706.093354
Pitcher, R.S., Brissett, N.C., Picher, A.J., Andrade, P., Juarez, R., Thompson, D., Fox, G.C., Blanco, L., Doherty, A.J., 2007b. Structure and Function of a Mycobacterial NHEJ DNA Repair Polymerase. J. Mol. Biol. 366, 391–405. https://doi.org/10.1016/j.jmb.2006.10.046
Pitcher, R.S., Tonkin, L.M., Green, A.J., Doherty, A.J., 2005. Domain Structure of a NHEJ DNA Repair Ligase from Mycobacterium tuberculosis. J. Mol. Biol. 351, 531–544. https://doi.org/10.1016/j.jmb.2005.06.038
Pollard, T.D., De La Cruz, E.M., 2013. Take advantage of time in your experiments: a guide to simple, informative kinetics assays. Mol. Biol. Cell 24, 1103–1110. https://doi.org/10.1091/mbc.e13-01-0030
Pommier, Y., 2006. Topoisomerase I inhibitors: camptothecins and beyond. Nat. Rev. Cancer 6, 789–802. https://doi.org/10.1038/nrc1977
Povirk, L.F., 1996. DNA damage and mutagenesis by radiomimetic DNA-cleaving agents: bleomycin, neocarzinostatin and other enediynes. Mutat. Res. Mol. Mech. Mutagen., Mutagenicity of Anticancer Drugs 355, 71–89. https://doi.org/10.1016/0027-5107(96)00023-1
Prasad, R., Beard, W.A., Strauss, P.R., Wilson, S.H., 1998. Human DNA Polymerase β Deoxyribose Phosphate Lyase: SUBSTRATE SPECIFICITY AND CATALYTIC MECHANISM*. J. Biol. Chem. 273, 15263–15270. https://doi.org/10.1074/jbc.273.24.15263
Prasad, R., Widen, S.G., Singhal, R.K., Watkins, J., Prakash, L., Wilson, S.H., 1993. Yeast open reading frame YCR14C encodes a DNA beta-polymerase-like enzyme. Nucleic Acids Res. 21, 5301–5307. https://doi.org/10.1093/nar/21.23.5301

Preer, J.R., Preer, L.B., Rudman, B.M., Barnett, A.J., 1985. Deviation from the universal code shown by the gene for surface protein 51A in Paramecium. Nature 314, 188–190. https://doi.org/10.1038/314188a0
Prescott, D.M., 1994. The DNA of ciliated protozoa. Microbiol. Rev. 58, 233–267.
Prostova, M., Shilkin, E., Kulikova, A.A., Makarova, A., Ryazansky, S., Kulbachinskiy, A., 2022. Noncanonical prokaryotic X family DNA polymerases lack polymerase activity and act as exonucleases. Nucleic Acids Res. 50, 6398–6413. https://doi.org/10.1093/nar/gkac461
Pryor, J.M., Conlin, M.P., Carvajal-Garcia, J., Luedeman, M.E., Luthman, A.J., Small, G.W., Ramsden, D.A., 2018. Ribonucleotide incorporation enables repair of chromosome breaks by nonhomologous end joining. Science 361, 1126–1129. https://doi.org/10.1126/science.aat2477
Pryor, J.M., Waters, C.A., Aza, A., Asagoshi, K., Strom, C., Mieczkowski, P.A., Blanco, L., Ramsden, D.A., 2015. Essential role for polymerase specialization in cellular nonhomologous end joining. Proc. Natl. Acad. Sci. 112, E4537–E4545. https://doi.org/10.1073/pnas.1505805112
Qi, Z., Redding, S., Lee, J.Y., Gibb, B., Kwon, Y., Niu, H., Gaines, W.A., Sung, P., Greene, E.C., 2015. DNA Sequence Alignment by Microhomology Sampling during Homologous Recombination. Cell 160, 856–869. https://doi.org/10.1016/j.cell.2015.01.029
Raia, P., Delarue, M., Sauguet, L., 2019. An updated structural classification of replicative DNA polymerases. Biochem. Soc. Trans. 47, 239–249. https://doi.org/10.1042/BST20180579
Ramachandran, G.N., Ramakrishnan, C., Sasisekharan, V., 1963. Stereochemistry of polypeptide chain configurations. J. Mol. Biol. 7, 95–99. https://doi.org/10.1016/s0022-2836(63)80023-6
Ramadan, K., Shevelev, I.V., Maga, G., Hübscher, U., 2002. DNA Polymerase λ from Calf Thymus Preferentially Replicates Damaged DNA*. J. Biol. Chem. 277, 18454–18458. https://doi.org/10.1074/jbc.M200421200
Ramsden, D.A., Asagoshi, K., 2012. DNA polymerases in nonhomologous end joining: Are there any benefits to standing out from the crowd? Environ. Mol. Mutagen. 53, 741–751. https://doi.org/10.1002/em.21725
Ramsden, D.A., Carvajal-Garcia, J., Gupta, G.P., 2022. Mechanism, cellular functions and cancer roles of polymerase-theta-mediated DNA end joining. Nat. Rev. Mol. Cell Biol. 23, 125–140. https://doi.org/10.1038/s41580-021-00405-2
Raoult, D., Audic, S., Robert, C., Abergel, C., Renesto, P., Ogata, H., La Scola, B., Suzan, M., Claverie, J.-M., 2004. The 1.2-megabase genome sequence of Mimivirus. Science 306, 1344–1350. https://doi.org/10.1126/science.1101485
Read, R.J., Adams, P.D., Arendall, W.B., Brunger, A.T., Emsley, P., Joosten, R.P., Kleywegt, G.J., Krissinel, E.B., Lütteke, T., Otwinowski, Z., Perrakis, A., Richardson, J.S., Sheffler, W.H., Smith, J.L., Tickle, I.J., Vriend, G., Zwart, P.H., 2011. A New Generation of Crystallographic Validation Tools for the Protein Data Bank. Structure 19, 1395–1412. https://doi.org/10.1016/j.str.2011.08.006
Remmert, M., Biegert, A., Hauser, A., Söding, J., 2011. HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nat. Methods 9, 173–175. https://doi.org/10.1038/nmeth.1818
Reynolds, P., Anderson, J.A., Harper, J.V., Hill, M.A., Botchway, S.W., Parker, A.W., O'Neill, P., 2012. The dynamics of Ku70/80 and DNA-PKcs at DSBs induced by ionizing radiation is dependent on the complexity of damage. Nucleic Acids Res. 40, 10821–10831. https://doi.org/10.1093/nar/gks879
Riballo, E., Woodbine, L., Stiff, T., Walker, S.A., Goodarzi, A.A., Jeggo, P.A., 2009. XLF-Cernunnos promotes DNA ligase IV–XRCC4 re-adenylation following ligation. Nucleic Acids Res. 37, 482–492. https://doi.org/10.1093/nar/gkn957
Ribes-Zamora, A., Mihalek, I., Lichtarge, O., Bertuch, A.A., 2007. Distinct faces of the Ku heterodimer mediate DNA repair and telomeric functions. Nat. Struct. Mol. Biol. 14, 301–307. https://doi.org/10.1038/nsmb1214
Robert, F., Barbeau, M., Éthier, S., Dostie, J., Pelletier, J., 2015. Pharmacological inhibition of DNA-PK stimulates Cas9-mediated genome editing. Genome Med. 7, 93. https://doi.org/10.1186/s13073-015-0215-6
Roberts, S.A., Strande, N., Burkhalter, M.D., Strom, C., Havener, J.M., Hasty, P., Ramsden, D.A., 2010. Ku is a 5'dRP/AP lyase that excises nucleotide damage near broken ends. Nature 464, 1214–1217. https://doi.org/10.1038/nature08926
Rodríguez, G., Martín, M.T., de Vega, M., 2019. An array of basic residues is essential for the nucleolytic activity of the PHP domain of bacterial/archaeal PolX DNA polymerases. Sci. Rep. 9, 9947. https://doi.org/10.1038/s41598-019-46349-8
Romain, F., Barbosa, I., Gouge, J., Rougeon, F., Delarue, M., 2009. Conferring a template-dependent polymerase activity to terminal deoxynucleotidyltransferase by mutations in the Loop1 region. Nucleic Acids Res. 37, 4642–4656. https://doi.org/10.1093/nar/gkp460
Rossi, M.L., Ghosh, A.K., Bohr, V.A., 2010. Roles of Werner syndrome protein in protection of genome integrity. DNA Repair, Helicase and translocases required for the maintenance of genome stability 9, 331–344. https://doi.org/10.1016/j.dnarep.2009.12.011
Roth, D.B., 2014. V(D)J Recombination: Mechanism, Errors, and Fidelity. Microbiol. Spectr. 2, 10.1128/microbiolspec.MDNA3-0041–2014. https://doi.org/10.1128/microbiolspec.MDNA3-0041-2014
Rzeszutek, I., Maurer-Alcalá, X.X., Nowacki, M., 2020. Programmed genome rearrangements in ciliates. Cell.

Mol. Life Sci. 77, 4615–4629. https://doi.org/10.1007/s00018-020-03555-2
Saiki, R.K., Gelfand, D.H., Stoffel, S., Scharf, S.J., Higuchi, R., Horn, G.T., Mullis, K.B., Erlich, H.A., 1988. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science 239, 487–491. https://doi.org/10.1126/science.2448875
Sakamoto, A., Iwabata, K., Koshiyama, A., Sugawara, H., Yanai, T., Kanai, Y., Takeuchi, R., Daikuhara, Y., Takakusagi, Y., Sakaguchi, K., 2007. Two X family DNA polymerases, λ and μ, in meiotic tissues of the basidiomycete, Coprinus cinereus. Chromosoma 116, 545–556. https://doi.org/10.1007/s00412-007-0119-3
Sandoval, P.Y., Swart, E.C., Arambasic, M., Nowacki, M., 2014. Functional diversification of Dicer-like proteins and small RNAs required for genome sculpting. Dev. Cell 28, 174–188. https://doi.org/10.1016/j.devcel.2013.12.010
Schärer, O.D., 2013. Nucleotide Excision Repair in Eukaryotes. Cold Spring Harb. Perspect. Biol. 5, a012609. https://doi.org/10.1101/cshperspect.a012609
Schneider, A., Bergsch, J., Lipps, G., 2023. The monomeric archaeal primase from Nanoarchaeum equitans harbours the features of heterodimeric archaeoeukaryotic primases and primes sequence-specifically. Nucleic Acids Res. 51, 5087–5105. https://doi.org/10.1093/nar/gkad261
Schneider, T.D., Stephens, R.M., 1990. Sequence logos: a new way to display consensus sequences. Nucleic Acids Res. 18, 6097–6100.
Schoenfeld, T.W., Murugapiran, S.K., Dodsworth, J.A., Floyd, S., Lodes, M., Mead, D.A., Hedlund, B.P., 2013. Lateral gene transfer of family A DNA polymerases between thermophilic viruses, aquificae, and apicomplexa. Mol. Biol. Evol. 30, 1653–1664. https://doi.org/10.1093/molbev/mst078
Schroeder, H.W., Cavacini, L., 2010. Structure and Function of Immunoglobulins. J. Allergy Clin. Immunol. 125, S41–S52. https://doi.org/10.1016/j.jaci.2009.09.046
Scully, R., Panday, A., Elango, R., Willis, N.A., 2019. DNA double-strand break repair-pathway choice in somatic mammalian cells. Nat. Rev. Mol. Cell Biol. 20, 698–714. https://doi.org/10.1038/s41580-019-0152-0
Seif-El-Dahan, M., Kefala-Stavridi, A., Frit, P., Hardwick, S.W., Chirgadze, D.Y., Maia De Oliviera, T., Andreani, J., Britton, S., Barboule, N., Bossaert, M., Pandurangan, A.P., Meek, K., Blundell, T.L., Ropars, V., Calsou, P., Charbonnier, J.-B., Chaplin, A.K., 2023. PAXX binding to the NHEJ machinery explains functional redundancy with XLF. Sci. Adv. 9, eadg2834. https://doi.org/10.1126/sciadv.adg2834
Sfeir, A., Symington, L.S., 2015. Microhomology-Mediated End Joining: A Back-up Survival Mechanism or Dedicated Pathway? Trends Biochem. Sci. 40, 701–714. https://doi.org/10.1016/j.tibs.2015.08.006
Shi, X., Hong, T., Walter, K.L., Ewalt, M., Michishita, E., Hung, T., Carney, D., Peña, P., Lan, F., Kaadige, M.R., Lacoste, N., Cayrou, C., Davrazou, F., Saha, A., Cairns, B.R., Ayer, D.E., Kutateladze, T.G., Shi, Y., Côté, J., Chua, K.F., Gozani, O., 2006. ING2 PHD domain links histone H3 lysine 4 methylation to active gene repression. Nature 442, 96–99. https://doi.org/10.1038/nature04835
Shock, D.D., Freudenthal, B.D., Beard, W.A., Wilson, S.H., 2017. Modulating the DNA polymerase β reaction equilibrium to dissect the reverse reaction. Nat. Chem. Biol. 13, 1074–1080. https://doi.org/10.1038/nchembio.2450
Shuman, S., Glickman, M.S., 2007. Bacterial DNA repair by non-homologous end joining. Nat. Rev. Microbiol. 5, 852–861. https://doi.org/10.1038/nrmicro1768
Sibanda, B.L., Critchlow, S.E., Begun, J., Pei, X.Y., Jackson, S.P., Blundell, T.L., Pellegrini, L., 2001. Crystal structure of an Xrcc4-DNA ligase IV complex. Nat. Struct. Biol. 8, 1015–1019. https://doi.org/10.1038/nsb725
Sims, R.J., Chen, C.-F., Santos-Rosa, H., Kouzarides, T., Patel, S.S., Reinberg, D., 2005. Human but not yeast CHD1 binds directly and selectively to histone H3 methylated at lysine 4 via its tandem chromodomains. J. Biol. Chem. 280, 41789–41792. https://doi.org/10.1074/jbc.C500395200
Sinha, K.M., Stephanou, N.C., Gao, F., Glickman, M.S., Shuman, S., 2007. Mycobacterial UvrD1 is a Ku-dependent DNA helicase that plays a role in multiple DNA repair events, including double-strand break repair. J. Biol. Chem. 282, 15114–15125. https://doi.org/10.1074/jbc.M701167200
Skipper, L., 2005. PROTEINS | Overview, in: Worsfold, P., Townshend, A., Poole, C. (Eds.), Encyclopedia of Analytical Science (Second Edition). Elsevier, Oxford, pp. 344–352. https://doi.org/10.1016/B0-12-369397-7/00493-3
Smith, R., Sellou, H., Chapuis, C., Huet, S., Timinszky, G., 2018. CHD3 and CHD4 recruitment and chromatin remodeling activity at DNA breaks is promoted by early poly(ADP-ribose)-dependent chromatin relaxation. Nucleic Acids Res. 46, 6087–6098. https://doi.org/10.1093/nar/gky334
Sonneborn, T.M., 1937. Sex, Sex Inheritance and Sex Determination in Paramecium Aurelia. Proc. Natl. Acad. Sci. 23, 378–385. https://doi.org/10.1073/pnas.23.7.378
Srivastava, M., Raghavan, S.C., 2015. DNA Double-Strand Break Repair Inhibitors as Cancer Therapeutics. Chem. Biol. 22, 17–29. https://doi.org/10.1016/j.chembiol.2014.11.013
Staehelin, M., 1973. Isoacceptor tRNA's, in: Bautz, E.K.F., Karlson, P., Kersten, H. (Eds.), Regulation of Transcription and Translation in Eukaryotes, Colloquium Der Gesellschaft Für Biologische Chemie 26.–28. April 1973 in Mosbach/Baden. Springer, Berlin, Heidelberg, pp. 313–321. https://doi.org/10.1007/978-3-642-

65725-2_16
Steitz, T.A., 1999. DNA Polymerases: Structural Diversity and Common Mechanisms. J. Biol. Chem. 274, 17395–17398. https://doi.org/10.1074/jbc.274.25.17395
Stephanou, N.C., Gao, F., Bongiorno, P., Ehrt, S., Schnappinger, D., Shuman, S., Glickman, M.S., 2007. Mycobacterial nonhomologous end joining mediates mutagenic repair of chromosomal double-strand DNA breaks. J. Bacteriol. 189, 5237–5246. https://doi.org/10.1128/JB.00332-07
Stewart, C.R., Casjens, S.R., Cresawn, S.G., Houtz, J.M., Smith, A.L., Ford, M.E., Peebles, C.L., Hatfull, G.F., Hendrix, R.W., Huang, W.M., Pedulla, M.L., 2009. The genome of Bacillus subtilis bacteriophage SPO1. J. Mol. Biol. 388, 48–70. https://doi.org/10.1016/j.jmb.2009.03.009
Stinson, B.M., Moreno, A.T., Walter, J.C., Loparo, J.J., 2020. A mechanism to minimize errors during non-homologous end joining. Mol. Cell 77, 1080-1091.e8. https://doi.org/10.1016/j.molcel.2019.11.018
Talbert, P.B., Henikoff, S., 2017. Histone variants on the move: substrates for chromatin dynamics. Nat. Rev. Mol. Cell Biol. 18, 115–126. https://doi.org/10.1038/nrm.2016.148
Taylor, M.R.G., Špírek, M., Chaurasiya, K.R., Ward, J.D., Carzaniga, R., Yu, X., Egelman, E.H., Collinson, L.M., Rueda, D., Krejci, L., Boulton, S.J., 2015. Rad51 Paralogs Remodel Pre-synaptic Rad51 Filaments to Stimulate Homologous Recombination. Cell 162, 271–286. https://doi.org/10.1016/j.cell.2015.06.015
Temin, H.M., 1985. Reverse transcription in the eukaryotic genome: retroviruses, pararetroviruses, retrotransposons, and retrotranscripts. Mol. Biol. Evol. 2, 455–468. https://doi.org/10.1093/oxfordjournals.molbev.a040365
Traut, T.W., 1994. Physiological concentrations of purines and pyrimidines. Mol. Cell. Biochem. 140, 1–22. https://doi.org/10.1007/BF00928361
Trifinopoulos, J., Nguyen, L.-T., von Haeseler, A., Minh, B.Q., 2016. W-IQ-TREE: a fast online phylogenetic tool for maximum likelihood analysis. Nucleic Acids Res. 44, W232–W235. https://doi.org/10.1093/nar/gkw256
Tsai, M.-D., 2019. Catalytic mechanism of DNA polymerases—Two metal ions or three? Protein Sci. 28, 288–291. https://doi.org/10.1002/pro.3542
Tsai, Y.-C., Johnson, K.A., 2006. A New Paradigm for DNA Polymerase Specificity. Biochemistry 45, 9675–9687. https://doi.org/10.1021/bi060993z
Uchiyama, Y., Kimura, S., Yamamoto, T., Ishibashi, T., Sakaguchi, K., 2004. Plant DNA polymerase λ, a DNA repair enzyme that functions in plant meristematic and meiotic tissues. Eur. J. Biochem. 271, 2799–2807. https://doi.org/10.1111/j.1432-1033.2004.04214.x
Uchiyama, Y., Takeuchi, R., Kodera, H., Sakaguchi, K., 2009. Distribution and roles of X-family DNA polymerases in eukaryotes. Biochimie 91, 165–170. https://doi.org/10.1016/j.biochi.2008.07.005
Uphoff, S., Reyes-Lamothe, R., Garza de Leon, F., Sherratt, D.J., Kapanidis, A.N., 2013. Single-molecule DNA repair in live bacteria. Proc. Natl. Acad. Sci. U. S. A. 110, 8063–8068. https://doi.org/10.1073/pnas.1301804110
Ussery, D.W., 2002. DNA Structure: A-, B- and Z-DNA Helix Families, in: Encyclopedia of Life Sciences. John Wiley & Sons, Ltd. https://doi.org/10.1038/npg.els.0003122
Uversky, V.N., 2011. Intrinsically disordered proteins from A to Z. Int. J. Biochem. Cell Biol. 43, 1090–1103. https://doi.org/10.1016/j.biocel.2011.04.001
Van Der Wal, C., Ho, S.Y.W., 2019. Molecular Clock, in: Ranganathan, S., Gribskov, M., Nakai, K., Schönbach, C. (Eds.), Encyclopedia of Bioinformatics and Computational Biology. Academic Press, Oxford, pp. 719–726. https://doi.org/10.1016/B978-0-12-809633-8.20257-4
van Gent, D.C., Hoeijmakers, J.H.J., Kanaar, R., 2001. Chromosomal stability and the DNA double-stranded break connection. Nat. Rev. Genet. 2, 196–206. https://doi.org/10.1038/35056049
Van Houten, B., 2020. Graphical snapshot of Samuel H. Wilson. DNA Repair, Tribute to Samuel H. Wilson: Shining Light on Base Excision DNA Repair 93, 102934. https://doi.org/10.1016/j.dnarep.2020.102934
Van Houten, B., Santa-Gonzalez, G.A., Camargo, M., 2018. DNA repair after oxidative stress: current challenges. Curr. Opin. Toxicol. 7, 9–16. https://doi.org/10.1016/j.cotox.2017.10.009
Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., Yuan, D., Stroe, O., Wood, G., Laydon, A., Žídek, A., Green, T., Tunyasuvunakool, K., Petersen, S., Jumper, J., Clancy, E., Green, R., Vora, A., Lutfi, M., Figurnov, M., Cowie, A., Hobbs, N., Kohli, P., Kleywegt, G., Birney, E., Hassabis, D., Velankar, S., 2022. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res. 50, D439–D444. https://doi.org/10.1093/nar/gkab1061
Vashishtha, A.K., Konigsberg, W.H., 2016. Effect of Different Divalent Cations on the Kinetics and Fidelity of RB69 DNA Polymerase. Biochemistry 55, 2661–2670. https://doi.org/10.1021/acs.biochem.5b01350
Vashishtha, A.K., Wang, J., Konigsberg, W.H., 2016. Different Divalent Cations Alter the Kinetics and Fidelity of DNA Polymerases. J. Biol. Chem. 291, 20869–20875. https://doi.org/10.1074/jbc.R116.742494
Vogt, A., He, Y., Lees-Miller, S.P., 2023. How to fix DNA breaks: new insights into the mechanism of non-homologous end joining. Biochem. Soc. Trans. BST20220741. https://doi.org/10.1042/BST20220741
Walker, J.R., Corpina, R.A., Goldberg, J., 2001. Structure of the Ku heterodimer bound to DNA and its

implications for double-strand break repair. Nature 412, 607–614. https://doi.org/10.1038/35088000
Wallace, S.S., Murphy, D.L., Sweasy, J.B., 2012. Base Excision Repair and Cancer. Cancer Lett. 327, 73–89. https://doi.org/10.1016/j.canlet.2011.12.038
Wang, J., Konigsberg, W.H., 2022. Two-Metal-Ion Catalysis: Inhibition of DNA Polymerase Activity by a Third Divalent Metal Ion. Front. Mol. Biosci. 9.
Wang, T.S., Eichler, D.C., Korn, D., 1977. Effect of Mn2+ on the in vitro activity of human deoxyribonucleic acid polymerase beta. Biochemistry 16, 4927–4934. https://doi.org/10.1021/bi00641a029
Waters, C.A., Strande, N.T., Pryor, J.M., Strom, C., Mieczkowski, P., Burkhalter, M.D., Oh, S., Qaqish, B.F., Moore, D.T., Hendrickson, E.A., Ramsden, D.A., 2014. The fidelity of the ligation step determines how ends are resolved during Nonhomologous end joining. Nat. Commun. 5, 4286. https://doi.org/10.1038/ncomms5286
Watson, J.D., Crick, F.H., 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171, 737–738. https://doi.org/10.1038/171737a0
Weill, J.-C., Reynaud, C.-A., 2008. DNA polymerases in adaptive immunity. Nat. Rev. Immunol. 8, 302–312. https://doi.org/10.1038/nri2281
Weinfeld, M., Mani, R.S., Abdou, I., Aceytuno, R.D., Glover, J.N.M., 2011. Tidying up loose ends: the role of polynucleotide kinase/phosphatase in DNA strand break repair. Trends Biochem. Sci. 36, 262–271. https://doi.org/10.1016/j.tibs.2011.01.006
Weller, G.R., 2002. Identification of a DNA Nonhomologous End-Joining Complex in Bacteria. Science 297, 1686–1689. https://doi.org/10.1126/science.1074584
Wheeler, D.L., Church, D.M., Federhen, S., Lash, A.E., Madden, T.L., Pontius, J.U., Schuler, G.D., Schriml, L.M., Sequeira, E., Tatusova, T.A., Wagner, L., 2003. Database resources of the National Center for Biotechnology. Nucleic Acids Res. 31, 28–33. https://doi.org/10.1093/nar/gkg033
Whitaker, A.M., Freudenthal, B.D., 2020. History of DNA polymerase β X-ray crystallography. DNA Repair, Tribute to Samuel H. Wilson: Shining Light on Base Excision DNA Repair 93, 102928. https://doi.org/10.1016/j.dnarep.2020.102928
Whitford, D., 2013. Proteins: Structure and Function. John Wiley & Sons.
Wilson, T.E., Lieber, M.R., 1999. Efficient processing of DNA ends during yeast nonhomologous end joining. Evidence for a DNA polymerase beta (Pol4)-dependent pathway. J. Biol. Chem. 274, 23599–23609. https://doi.org/10.1074/jbc.274.33.23599
Woodard, L.E., Wilson, M.H., 2015. piggyBac-ing models and new therapeutic strategies. Trends Biotechnol. 33, 525–533. https://doi.org/10.1016/j.tibtech.2015.06.009
Woodbine, L., Gennery, A.R., Jeggo, P.A., 2014. The clinical impact of deficiency in DNA non-homologous end-joining. DNA Repair 16, 84–96. https://doi.org/10.1016/j.dnarep.2014.02.011
Wright, W.D., Heyer, W.-D., 2014. Rad54 Functions as a Heteroduplex DNA Pump Modulated by Its DNA Substrates and Rad51 during D Loop Formation. Mol. Cell 53, 420–432. https://doi.org/10.1016/j.molcel.2013.12.027
Wright, W.D., Shah, S.S., Heyer, W.-D., 2018. Homologous recombination and the repair of DNA double-strand breaks. J. Biol. Chem. 293, 10524–10535. https://doi.org/10.1074/jbc.TM118.000372
Xiao, T., Hall, H., Kizer, K.O., Shibata, Y., Hall, M.C., Borchers, C.H., Strahl, B.D., 2003. Phosphorylation of RNA polymerase II CTD regulates H3 methylation in yeast. Genes Dev. 17, 654–663. https://doi.org/10.1101/gad.1055503
Xu, Y., Komiyama, M., 2023. G-Quadruplexes in Human Telomere: Structures, Properties, and Applications. Mol. Basel Switz. 29, 174. https://doi.org/10.3390/molecules29010174
Xu, Z., Zan, H., Pone, E.J., Mai, T., Casali, P., 2012. Immunoglobulin class-switch DNA recombination: induction, targeting and beyond. Nat. Rev. Immunol. 12, 517–531. https://doi.org/10.1038/nri3216
Yamtich, J., Sweasy, J.B., 2010. DNA polymerase family X: function, structure, and cellular roles. Biochim. Biophys. Acta 1804, 1136–1150. https://doi.org/10.1016/j.bbapap.2009.07.008
Yang, W., 2014. An Overview of Y-Family DNA Polymerases and a Case Study of Human DNA Polymerase η. Biochemistry 53, 2793–2803. https://doi.org/10.1021/bi500019s
Yano, K., Morotomi-Yano, K., Lee, K.-J., Chen, D.J., 2011. Functional significance of the interaction with Ku in DNA double-strand break recognition of XLF. FEBS Lett. 585, 841–846. https://doi.org/10.1016/j.febslet.2011.02.020
Yoo, S., Dynan, W.S., 1999. Geometry of a complex formed by double strand break repair proteins at a single DNA end: Recruitment of DNA-PKcs induces inward translocation of Ku protein. Nucleic Acids Res. 27, 4679–4686. https://doi.org/10.1093/nar/27.24.4679
Yoo, S., Kimzey, A., Dynan, W.S., 1999. Photocross-linking of an Oriented DNA Repair Complex: Ku BOUND AT A SINGLE DNA END*. J. Biol. Chem. 274, 20034–20039. https://doi.org/10.1074/jbc.274.28.20034
Yoon, J.-H., Basu, D., Sellamuthu, K., Johnson, R.E., Prakash, S., Prakash, L., 2021. A novel role of DNA polymerase λ in translesion synthesis in conjunction with DNA polymerase ζ. Life Sci. Alliance 4. https://doi.org/10.26508/lsa.202000900

Yoshida, T., Claverie, J.-M., Ogata, H., 2011. Mimivirus reveals Mre11/Rad50 fusion proteins with a sporadic distribution in eukaryotes, bacteria, viruses and plasmids. Virol. J. 8, 427. https://doi.org/10.1186/1743-422X-8-427

Yousefzadeh, M., Henpita, C., Vyas, R., Soto-Palma, C., Robbins, P., Niedernhofer, L., 2021. DNA damage-how and why we age? eLife 10, e62852. https://doi.org/10.7554/eLife.62852

Yusa, K., 2015. piggyBac Transposon. Microbiol. Spectr. 3, 10.1128/microbiolspec.mdna3-0028–2014. https://doi.org/10.1128/microbiolspec.mdna3-0028-2014

Zapotoczny, G., Sekelsky, J., 2017. Human Cell Assays for Synthesis-Dependent Strand Annealing and Crossing over During Double-Strand Break Repair. G3 GenesGenomesGenetics 7, 1191–1199. https://doi.org/10.1534/g3.116.037390

Zatopek, K.M., Alpaslan, E., Evans, T.C., Jr., Sauguet, L., Gardner, A.F., 2020. Novel ribonucleotide discrimination in the RNA polymerase-like two-barrel catalytic core of Family D DNA polymerases. Nucleic Acids Res. 48, 12204–12218. https://doi.org/10.1093/nar/gkaa986

Zhang, X., Yang, Z., Khan, S.I., Horton, J.R., Tamaru, H., Selker, E.U., Cheng, X., 2003. Structural basis for the product specificity of histone lysine methyltransferases. Mol. Cell 12, 177–185. https://doi.org/10.1016/s1097-2765(03)00224-7

Zhao, B., Rothenberg, E., Ramsden, D.A., Lieber, M.R., 2020. The molecular basis and disease relevance of non-homologous DNA end joining. Nat. Rev. Mol. Cell Biol. 21, 765–781. https://doi.org/10.1038/s41580-020-00297-8

Zhou, H., Hintze, B.J., Kimsey, I.J., Sathyamoorthy, B., Yang, S., Richardson, J.S., Al-Hashimi, H.M., 2015. New insights into Hoogsteen base pairs in DNA duplexes from a structure-based survey. Nucleic Acids Res. 43, 3420–3433. https://doi.org/10.1093/nar/gkv241

Zhou, V.W., Goren, A., Bernstein, B.E., 2011. Charting histone modifications and the functional organization of mammalian genomes. Nat. Rev. Genet. 12, 7–18. https://doi.org/10.1038/nrg2905

Zhu, H., Shuman, S., 2010. Gap Filling Activities of Pseudomonas DNA Ligase D (LigD) Polymerase and Functional Interactions of LigD with the DNA End-binding Ku Protein. J. Biol. Chem. 285, 4815–4825. https://doi.org/10.1074/jbc.M109.073874

Zhu, H., Shuman, S., 2007. Characterization of Agrobacterium tumefaciens DNA ligases C and D. Nucleic Acids Res. 35, 3631–3645. https://doi.org/10.1093/nar/gkm145

Zhu, H., Shuman, S., 2005. Novel 3′-Ribonuclease and 3′-Phosphatase Activities of the Bacterial Non-homologous End-joining Protein, DNA Ligase D. J. Biol. Chem. 280, 25973–25981. https://doi.org/10.1074/jbc.M504002200

Zimmermann, L., Stephens, A., Nam, S.-Z., Rau, D., Kübler, J., Lozajic, M., Gabler, F., Söding, J., Lupas, A.N., Alva, V., 2018. A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core. J. Mol. Biol., Computation Resources for Molecular Biology 430, 2237–2243. https://doi.org/10.1016/j.jmb.2017.12.007

Résumé

L'eucaryote unicellulaire *Paramecium tetraurelia* est un organisme binucléé, qui a pour particularité de perdre lors de sa reproduction le noyau nécessaire à l'expression de ses gènes (somatique). Celui-ci doit donc être régénéré à partir de son autre noyau (germinal), quant à lui diploïde. Cette régénération passe par de nombreuses réplications du génome, mais surtout par des réarrangements massifs, dont certains consistent en l'introduction programmée de cassures double-brin à des milliers de sites dans le génome, afin d'éliminer des séquences d'insertion qui cassent le cadre de lecture dans de nombreux gènes. Une fois ces cassures introduites, elles sont réparées par un système qui repose sur des protéines impliquées dans la réparation non-homologue, ou NHEJ (Ku70/80, DNA-PKcs, Ligase IV, XRCC4) et sur 4 ADN polymérases. Cependant, il existe une différence majeure entre le NHEJ, qui est connu pour son fort taux d'erreurs, et le NHEJ chez la paramécie qui ne fait quasiment pas d'erreurs. L'objectif des travaux de cette thèse est d'expliquer la fidélité de ce système, en se focalisant sur les ADN polymérases impliquées dans cette réparation chez *Paramecium tetraurelia*.

Dans un premier temps, une approche bio-informatique a été utilisée afin d'émettre des hypothèses sur les raisons de la fidélité de ces enzymes, en étudiant de façon approfondie la classification des ADN polymérases de la famille X. Après une étude enzymatique des ADN polymérases de *Paramecium tetraurelia* ayant permis de montrer leurs similitudes avec les ADN polymérases λ et β ainsi que leur grande fidélité, l'existence de deux mécanismes pouvant expliquer cette fidélité a été démontrée. Pour cela, l'activité enzymatique de mutants de l'ADN polymérase λ a été testée, et leur structure a été obtenue par cristallographie aux rayons X. Un premier mécanisme pouvant être impliqué dans la fidélité, similaire à celui rencontré chez l'ADN polymérase β, se base sur des changements conformationnels locaux au sein du site catalytique de l'enzyme. Le second mécanisme, jusqu'ici non caractérisé, utilise une boucle de 10 résidus pour stabiliser l'ADN au sein du site actif, uniquement en présence d'un nucléotide correct, et est retrouvé chez l'ADN polymérase λ.

Ces nouvelles connaissances sur les bases moléculaires de la fidélité des ADN polymérases de la famille X apportent une meilleure compréhension de la fidélité du NHEJ de *Paramecium tetraurelia*, ce qui pourra permettre d'élargir les connaissances sur le NHEJ et ses implications dans le système immunitaire et dans la carcinogenèse.

Abstract

The unicellular eukaryote *Paramecium tetraurelia* is a binucleate organism, which loses the nucleus required for gene expression (somatic) during reproduction. This nucleus must therefore be regenerated from its other nucleus (germinal), which is diploid. This regeneration involves numerous replications of the genome, but above all massive rearrangements, some of which consist in the programmed introduction of double-strand breaks at thousands of sites in the genome, in order to eliminate insertion sequences that break the reading frame in many genes. Once these breaks have been introduced, they are repaired by a system that relies on proteins involved in non-homologous repair, or NHEJ (Ku70/80, DNA-PKcs, Ligase IV, XRCC4) including 4 DNA polymerases. However, there is a major difference between classical NHEJ, which is known for its high error rate, and NHEJ in *Paramecium*, which makes virtually no errors. The aim of this thesis is to explain the fidelity of this system, focusing on the DNA polymerases involved in this repair in *Paramecium tetraurelia*.

Initially, a bioinformatics approach was used to hypothesize the reasons for the fidelity of these enzymes, by studying in depth the classification of DNA polymerases of family X. Following an enzymatic study of *Paramecium tetraurelia* DNA polymerases, which demonstrated their similarities to λ and β DNA polymerases, as well as their high fidelity, the existence of two mechanisms that could explain this fidelity was demonstrated. To this end, the enzymatic activity of DNA polymerase λ mutants was tested, and their structure obtained by X-ray crystallography. A first fidelity mechanism, similar to that found in DNA polymerase β, is based on local conformational changes within the enzyme's catalytic site. The second mechanism, uncharacterized until now, uses a 10-residue loop to stabilize the DNA within the active site, only in the presence of a correct nucleotide, and is also found in DNA polymerase λ.

These new insights into the molecular basis of X-family DNA polymerase fidelity provide a better understanding of *Paramecium tetraurelia* NHEJ fidelity, which may lead to a broader understanding of NHEJ and its implications in the immune system and carcinogenesis.