FACULTY OF SCIENCE
DEPARTMENT OF CHEMISTRY
BIOCHEMISTRY, MOLECULAR AND STRUCTURAL BIOLOGY DIVISION
Celestijnenlaan 200G box 2403
B-3001 Heverlee, BELGIUM
boris.bauwens@kuleuven.be

Boris Bauwens

MOLECULAR EVOLUTION OF POLYMERASES FOR THE SYNTHESIS OF DEOXYXYLOSE-BASED NUCLEIC ACIDS

April 2019

KU LEUVEN

ARENBERG DOCTORAL SCHOOL
FACULTY OF SCIENCE

# Molecular evolution of polymerases for the synthesis of deoxyxylose-based nucleic acids

**Boris Bauwens**

Supervisor:
Prof. dr. J. Robben

Dissertation presented in partial
fulfilment of the requirements for the
degree of Doctor of Science (PhD):
Biochemistry and Biotechnology

April 2019

# MOLECULAR EVOLUTION OF POLYMERASES FOR THE SYNTHESIS OF DEOXYXYLOSE-BASED NUCLEIC ACIDS

Boris BAUWENS

Supervisor:
Prof. J. Robben

Members of the
Examination Committee:
Prof. A. Voet
Prof. M. De Maeyer
Prof. P. Herdewijn
Prof. M. Froeyen
Prof. V. B. Pinheiro
Prof. Y. Briers

Dissertation presented in partial fulfilment of the requirements for the degree of Doctor of Science (PhD): Biochemistry and Biotechnology

April 2019

Uitgegeven in eigen beheer, Boris Bauwens, Leuven

# Dankwoord

Eerst en vooral wil ik mijn promotor, prof. Johan Robben, bedanken. Bedankt voor uw steun vanaf mijn masterthesis, mijn IWT aanvraag, tot aan het afwerken van dit doctoraat. Bedankt voor alle nuttige feedback op elk document, en om mij de goede richting te wijzen als ik even niet wist waarheen. Bedankt voor de goede samenwerking.

I would also like to thank my jury committee for their valuable feedback (and skepticism), both during my PhD project and the final manuscript.

Mijn dank gaat ook uit naar het IWT/VLAIO, voor de financiering van mijn onderzoek.

Vervolgens wil ik al mijn collega's bedanken die mij gedurende de jaren geholpen hebben. Kasia, Yves en Wouter, bedankt om mij te tonen wat een doctoraat maken inhoudt. Jullie kennis en ervaring met polymerase evolutie heeft voor mij een pad geëffend dat ik zonder jullie nooit zo vlot had kunnen bewandelen. Ik apprecieer elk van jullie optimalisaties en uitleg, en de inspanning die jullie hebben gestoken in mutagenese, ELISA en CST heeft mij veel tijd en moeite gespaard. Iris, jouw naam is ondertussen voor mij een begrip geworden dat zoveel betekent als "diegene die alles weet over het labo". Jouw optimisme en hulpvaardigheid zowel binnen als buiten het labo heb ik altijd gewaardeerd en zal ik nog lang onthouden. Vincent en Oscar, bedankt om mee een aangename sfeer te creëren in onze groep, en voor alle hulp en advies.

Bedankt aan iedereen in de 200G en 200F, voor de sfeer van behulpzaamheid, de leuke kerstfeestjes, de BBQs, de labtrips, aan iedereen die kritische vragen heeft gesteld op de "Friday meetings".

Ik wil ook mijn vrienden en vriendinnen bedanken. De Biochemisten, voor de fijne etentjes en goede raad. Iedereen in de kung fu lessen, voor de ontspanning én inspanning. Dries, voor de fijne babbels en etentjes.

Natuurlijk ook mijn  familie, voor al jullie steun. In het bijzonder mijn ouders, die er altijd voor mij zijn geweest en mijn opleiding mogelijk hebben gemaakt.

Tot slot wil ik jou bedanken, Ilse. Je hebt altijd al een voorsprong gehad met jouw doctoraat, en van die ervaring heb ik dankbaar gebruik kunnen maken. Het is fantastisch om altijd iemand bij me te hebben die mijn (doctoraats)problemen perfect begrijpt en kan helpen oplossen. Zonder jou zou dit doctoraat er waarschijnlijk niet geweest zijn. Bedankt voor je begrip, je hulp, je geduld, je liefde. Je bent geweldig. IOU.

# Table of contents

# List of abbreviations

| | |
|---|---|
| A | adenine |
| AHT | anhydrotetracyclin |
| Amp | ampicillin |
| ANA | arabinonucleic acid |
| bp | basepair |
| BSA | bovine serum albumin |
| C | cytosine |
| CAI | codon adaptation index |
| CD | circular dichroism |
| CeNA | cyclohexenyl nucleic acid |
| CSR | compartmentalized self-replication |
| CST | compartmentalized self-tagging |
| DdDp | DNA-dependent DNA polymerase |
| DIG | digoxigenin |
| DNA | 2’-deoxyribonucleic acid |
| DNAP | DNA polymerase |
| dNMP | 2’-deoxyribonucleotide monophosphate |
| dNT | 2’-deoxyribonucleotide |
| dNTP | 2’-deoxyribonucleoside triphosphate |
| dsDNA | double-stranded DNA |
| DTT | dithiothreitol |
| dxNA | 2’-deoxyxylonucleic acid |
| dxNT | 2’-deoxyxylonucleotide |
| dxNTP | 2’-deoxyxylonucleoside triphosphate |
| *E. coli* | *Escherichia coli* |
| EDTA | ethylenediaminetetraacetic acid |
| ELISA | enzyme-linked immunosorbent assay |
| ePCR | emulsion polymerase chain reaction |
| epPCR | error-prone polymerase chain reaction |
| EtBr | ethidium bromide |

| | |
|---|---|
| $exo^-$ | exonuclease deficient |
| FANA | fluoroarabinose nucleic acid |
| FRET | Förster resonance energy transfer |
| G | guanine |
| GGS | Golden Gate Shuffling |
| GNA | glycerol nucleic acid |
| HNA | hexitol nucleic acid |
| HRP | horse radish peroxidase |
| IDA | iminodiacetate |
| iPCR | inverse polymerase chain reaction |
| Kan | kanamycin |
| kb | kilobases |
| KO | knockout |
| LB | Luria-Bertani broth |
| LNA | locked nucleic acid |
| mRNA | messenger RNA |
| MS | mass spectrometry |
| NMR | nuclear magnetic resonance |
| OD | optical density |
| PAGE | polyacrylamide gel electrophoresis |
| PBST | phosphate buffered saline with Tween 20 |
| PCR | polymerase chain reaction |
| PEG | polyethylene glycol |
| PNA | peptide nucleic acid |
| $PP_i$ | pyrophosphate |
| RdDp | RNA-dependent DNA polymerase |
| RNA | ribonucleic acid |
| ROS | reactive oxygen species |
| rpm | revolutions per minute |
| RT | reverse transcriptase |
| SDS-PAGE | sodiumdodecylsulphate polyacrylamide gel electrophoresis |
| SELEX | systematic evolution of ligands by exponential enrichment |

| | |
|---|---|
| spCSR | short-patch compartmentalized self-replication |
| ssDNA | single-stranded DNA |
| T | thymine |
| TMB | 3,3’,5,5’-tetramethylbenzidine |
| TNA | threose nucleic acid |
| tPhoNA | 5’-O-phosphonomethyl-threosyl nucleic acid |
| tRNA | transfer RNA |
| U | uracil |
| VTS | VersaTile Shuffling |
| WT | wild type |
| XNA | xenonucleic acid |

# Summary

One of the most challenging aims of synthetic biology is the creation of a xenonucleic acid (XNA), the artificial equivalent of DNA and RNA, that is functional in a biological system. When these XNA and their building blocks are sufficiently different to natural DNA or RNA so they do not interfere with an organisms metabolism, they are said to be orthogonal. While the chemical synthesis of many different XNA building blocks is feasible for quite some time, efficiently polymerizing them into longer strands and manipulating them requires dedicated enzymes the development of which has only become possible over the recent years. However, there are no highly active and specific enzymes available so far to polymerize truly orthogonal XNA. To solve this problem, existing polymerases are modified to better accept these XNA building blocks.

In this dissertation, state-of-the-art molecular evolution was used to develop such new polymerases. First, the technique compartmentalized self-tagging (CST) was used to select for variants of the hyperthermophilic DNA polymerase Therminator that have increased activity for incorporating 2'-deoxyxylonucleotides (dxNTs). This Therminator DNA polymerase was already shown to be able to incorporate dxNTs, but stalls after incorporating just two. The likely reason for this is a difference in backbone geometry of dxNA to DNA, caused by an inversion of the orientation of the 3'-OH group, resulting in a more elongated and left-handed helix in deoxyxylonucleic acids (dxNA). The polymerase variants were created by mutagenesis that was targeted to six promising regions of the protein structure. Thus, by applying CST selection and screening selected clones with a polymerase activity assay, several mutants were found that can incorporate and extend dxNTs more efficiently than the original enzyme. Selection for this promiscuous activity thus resulted in mutants that appear to have increased activity.

One of these mutants, identified with a H545R mutation, was analysed in detail using gel electrophoresis at single-nucleotide resolution. This allowed the visualization of all extension products, including stalled intermediates, and confirmed that requiring incorporation up to three or more consecutive dxNTs causes polymerase stalling, but also showed that two of them can be extended with natural DNA. Moreover, if the dxNTs are sufficiently spaced by natural nucleotides, the mutant can incorporate up to four of them, and stabilize this presumably highly distorted DNA/XNA backbone. The H545R mutation likely stabilizes the altered backbone conformation at one of the incorporated dxNT positions that coincides with polymerase stalling. It was concluded that this new DNA polymerase mutant represents a first step towards a polymerase that can produce dxNA.

Furthermore, several techniques were also explored to further modify the Therminator polymerase. Instead of relying solely on amino acid substitutions, recombination of these mutations from different regions is possible using Golden Gate Shuffling. This technique allows for the production of a large number of new mutants by combining previous ones in the same protein. The other shuffling-based technique used in this dissertation, VersaTile Shuffling, also allows insertion of unrelated protein sequences in the evolving protein. These recombination strategies are shown to be able to introduce new variation in selected mutant libraries, although to date no new recombination mutant with increased activity has been found. VersaTile Shuffling in particular could prove to be very useful for modifying larger protein regions at once to introduce a new surface geometry, which might be needed to stabilize and bind an XNA with a geometrically strongly altered backbone structure.

Polymerase mutants, based on the ones that were developed in this thesis using molecular evolution, thus start to pave the way for XNA polymerases that are able to produce XNA with a markedly different backbone structure. Such a DNA-dependent XNA polymerase has short-term application in XNA aptamer development. Furthermore, in the long term, orthogonal nucleic acids and their XNA-dependent XNA polymerases that are unable to interfere with existing biology could thereby provide functional applications and increased biosafety in biotechnology as a novel, independent means of carrying sequence information.

# Samenvatting

Een van de meest uitdagende doelstellingen van synthetische biologie is het creëren van een xenonucleïnezuur (XNA), het synthetische equivalent van DNA en RNA, dat functioneel is in een biologisch systeem. Wanneer dit XNA en hun bouwstenen voldoende verschillen van natuurlijk DNA en RNA zodat ze niet interfereren met het metabolisme van een organisme, worden ze orthogonaal genoemd. Hoewel de chemische synthese van diverse XNA bouwstenen al geruime tijd haalbaar is, vereist efficiënte polymerisatie tot langere strengen en hun manipulatie toegewijde enzymen waarvan de ontwikkeling pas in de laatste jaren mogelijk is geworden. Er zijn echter nog geen specifieke enzymen met hoge activiteit beschikbaar om volledig orthogonaal XNA te polymeriseren. Om dit probleem op te lossen worden bestaande polymerase aangepast om XNA bouwstenen beter te accepteren.

In dit proefschrift werd state-of-the-art moleculaire evolutie gebruikt om zulke nieuwe polymerasen te ontwikkelen. Eerst werd de compartmentalized self-tagging (CST) techniek gebruikt om te selecteren op varianten van het hyperthermofiele DNA-polymerase Therminator met een verhoogde activiteit voor de incorporatie van 2'-deoxyxylonucleotiden (dxNTs). Het was al eerder aangetoond dat dit Therminator DNA-polymerase dxNTs kan inbouwen, maar blokkeert na slechts twee incorporaties. Dit komt waarschijnlijk door het verschil in ruggengraatgeometrie tussen deoxyxylonucleïnezuur (dxNA) en DNA, wat wordt veroorzaakt door een inversie in oriëntatie van de 3'-OH groep, wat resulteert in een meer uitgerekte en linkshandige helix in dxNA. De polymerasevarianten werden gemaakt met mutagenese gericht op zes veelbelovende regio's van de proteïnestructuur. Door het toepassen van CST selectie en het screenen van geselecteerde klonen met een polymerase activiteit assay werden op deze manier verschillende mutanten gevonden die dxNTs efficiënter kunnen inbouwen en verlengen dan het origineel enzym. Selectie voor deze promiscue activiteit resulteerde zo in mutanten die een verhoogde activiteit lijken te hebben.

Eén van deze mutanten, waar een H545R mutatie geïdentificeerd werd, werd in detail geanalyseerd met gelelektroforese met resolutie van één nucleotide. Dit liet de visualisatie toe van alle extensieproducten, inclusief geblokkeerde intermediaire producten, en bevestigde dat de incorporatie van drie of meer opeenvolgende dxNTs het polymerase doet blokkeren, maar toonde ook aan dat twee ervan verlengd kunnen worden met natuurlijk DNA.

Wanneer de dxNTs voldoende gescheiden zijn door natuurlijke nucleotiden kan de mutant daarbovenop tot vier dxNTs inbouwen en deze waarschijnlijk sterk vervormde DNA/XNA ruggengraat stabiliseren. De H545R mutatie stabiliseert waarschijnlijk de veranderde ruggengraatconformatie op één van de ingebouwde dxNT posities die overeenkomen met het blokkeren van het polymerase. Er werd geconcludeerd dat deze nieuwe DNA-polymerase mutant een eerste stap is naar een polymerase dat dxNA kan produceren.

Verder werden verschillende technieken onderzocht om het Therminator polymerase verder aan te passen. In de plaats van enkel te vertrouwen op aminozuur substituties kunnen deze mutaties uit verschillende regio’s ook gerecombineerd worden via Golden Gate Shuffling. Deze techniek laat de aanmaak van een groot aantal nieuwe mutanten toe door bestaande mutatie te combineren in hetzelfde proteïne. De andere op shuffling gebaseerde techniek die in dit proefschrift werd gebruikt, VersaTile Shuffling, laat ook de insertie van niet-gerelateerde proteïnesequenties toe in het evoluerende proteïne. Van deze recombinatiestrategieën is aangetoond dat ze nieuwe variatie kunnen introduceren in geselecteerde mutantenbibliotheken, hoewel tot op heden nog geen nieuwe gerecombineerde mutant werd gevonden met verhoogde activiteit. VersaTile Shuffling in het bijzonder zou zeer nuttig kunnen zijn om grotere eiwitregio’s in hun geheel te wijzigen om een nieuwe oppervlaktegeometrie te creëren, wat nodig zou kunnen zijn om XNA met een sterk gewijzigde ruggengraatstructuur te binden en stabiliseren.

Polymerasemutanten, gebaseerd op de mutanten die werden ontwikkeld in dit proefschrift via moleculaire evolutie, beginnen dus te weg vrij te maken voor XNA-polymerasen die XNA kunnen produceren met een erg verschillende ruggengraatstructuur. Zulk een DNA-afhankelijk XNA-polymerase heeft al kortetermijntoepassingen in de ontwikkeling van XNA aptameren. Bovendien kunnen orthogonale nucleïnezuren en hun XNA-afhankelijke XNA-polymerasen die niet interfereren met bestaande biologie hierdoor functionele toepassingen en verhoogde bio-veiligheid voorzien als een nieuwe, onafhankelijke manier om sequentie-informatie te dragen.

# Aims

From both a fundamental research standpoint as well as biosafety, synthetic biologists are attempting to create novel nucleic acid analogues, called xenonucleic acids or XNA. When such an XNA does not interfere with natural DNA or RNA it is said to be orthogonal. To create such orthogonal XNA and manipulate it enzymatically, dedicated polymerase catalysts are required. The goal of this thesis is to develop a polymerase that builds 2'-deoxyxylose based nucleic acid (dxNA). This nucleic acid has a marked difference in backbone geometry, compared to DNA, in that the double helix is left-handed and more elongated. This nucleic acid was chosen because it is not recognized by natural proteins and does not hybridize with DNA or RNA. This way, sequence information cannot pass from dxNA to DNA or RNA, and the artificial nucleic acid is immune to natural endo- and exonucleases. However, polymerases that can accurately and efficiently polymerize dxNA oligonucleotides are not available.

To develop such a polymerase, we intend to use molecular evolution of a thermostable DNA polymerase, termed Therminator. This strategy (see Figure A1) involves creating a library of polymerase mutants which are then selected for one or more rounds on their ability to incorporate 2'-deoxyxylonucleotides (dxNTs) against a DNA template. These selected libraries are then screened for activity, and promising mutants are analyzed.

The first aim is the creation of mutant libraries (Chapter 2) based on the structural information of an X-ray crystal of DNA polymerase from *Thermococcus sp.* 9°N-7 bound to DNA. Modeling of deoxyribose-to-deoxyxylose substitutions in the primer of the bound DNA shows positions in the polymerase that are likely to cause steric interactions. Mutant libraries are then created by targeted mutagenesis of specific regions of the protein that are in close contact with the bound nucleic acid.

The second aim is the selection of these libraries with compartmentalized self-tagging (CST) for improved dxNT incorporation (Chapter 3). *E. coli* cells transformed with these plasmid libraries and expressing the Therminator mutant polymerases are emulsified and heat lysed. Mutant polymerases set free in each droplet (compartment) are allowed to extend a plasmid-bound primer with dxNTs and to further elongate it, eventually tagging their plasmid by incorporating a biotinylated nucleotide. Selection is based on capture of tagged plasmids by their affinity to streptavidin. As non-specific binding may occur, enrichment is monitored using mock libraries of wild type and a knockout mutant. Selected clones are further screened using a polymerase activity assay based on ELISA to confirm their improved activity. Most promising clones are sequence identified and their enzyme activity analyzed in more detail. This is achieved by letting the mutant polymerases extend primers on templates that differ in the number and position of specific dxNTs to be incorporated.

The third aim is the creation of new mutants based on those isolated by mutant screening (Chapter 4). This involves both mutagenesis to create mutants that are highly similar to selected ones and recombination of these mutants with each other.

The outcome of this strategy would be a DNA-dependent dxNA polymerase, which is a first and essential step in the ultimate development of a dxNA-dependent dxNA polymerase.

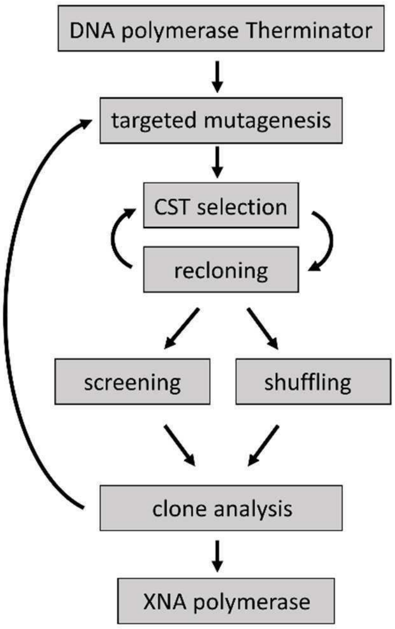


**Figure A1: Overview of the workflow of evolving an XNA polymerase.**

The starting protein is the DNA polymerase Therminator. Mutagenesis of this protein yields different mutant libraries, which are then selected on using compartmentalized self-tagging (CST). After re-cloning the selected mutants, selection can be repeated or the secondary library can be screened for clones with higher activity. These clones are analyzed, and the resulting information is used to guide further mutagenesis. Different mutations can also be recombined, and when multiple combinations exist, these shuffled populations themselves can be subjected to selection rounds.

# 1

# General introduction

In this chapter, a summary of the research field and relevant state of the art technology will be given. First, xenobiology and synthetic biology is discussed from the point of view of unnatural nucleic acids. The concept of orthogonality is discussed in the context of these nucleic acids, which applies when they do not interact with natural nucleic acids and proteins. This also makes them potentially useful as new aptamers that can adapt a functional fold for *in vitro* or *in vivo* applications.

Next, the classification, structure, function and reaction mechanism of DNA polymerases are explained. One of these fairly well characterized DNA polymerases is Therminator, which is the protein that is used as a starting point for molecular evolution.

Molecular evolution, the improving of a biological macromolecule through selection on a mutant library, can be performed in different ways. Self-tagging, self-replication and phage display based strategies, as well as protein design, are discussed.

## 1. Xenobiology and synthetic biology

The field of synthetic biology is an interdisciplinary overlap between biology and engineering. It designs and constructs new biological systems different from those in nature, or redesigns existing ones for new purposes (1, 2). These purposes are both practical and fundamental. New designs and design tools allow our technology to operate on much smaller scales, become more environmentally friendly, or biocompatible for example. In accordance with the words of Richard Feynmann, “What I cannot create, I do not understand”, whether or not these new technologies work depends on our understanding of biology, and therefore simultaneously act as a test of our knowledge.

Xenobiology is simply the study of unusual life, in other words life as we don’t know it on Earth. It is based on the premise that life does not necessarily need to be exactly the way it is on Earth. It could have fundamentally different cellular organization, different metabolic pathways, other proteins and nucleic acids (for example with an expanded genetic code), or have entirely different molecules that carry sequence information or perform the functions necessary for life. Synthesizing these artificial molecules, or combining naturally occurring molecules in novel ways, leads to potentially new forms of life. In this thesis, the focus lies on artificial nucleic acids, more specifically those with altered backbone geometry.

## 1.1. Xenonucleic acids

Nucleic acids are a versatile class of biomolecules that are capable of many functions. They carry the genetic information of all life on Earth and have been accurately doing this for over 3.5 billion years. They can fold into different structures, with double-stranded DNA forming stable helices, and the usually single-stranded RNA forming for example cloverleaf-like tRNA molecules. Even though most catalysis in modern life is performed by protein enzymes, nucleic acids can also perform a large range of complex reactions, such as polypeptide synthesis in ribosomes and mRNA splicing in spliceosomes, or in human-made ribozymes. However, life as we know it has been very conservative with the chemical structure of the nucleic acids it uses: all of them are composed of a sugar-phosphate backbone carrying an evolving sequence of nucleobases. The sugar is always ribose (in RNA) or its derivative 2'-deoxyribose (in DNA), and sequences are constrained by only five canonical bases: adenine, guanine, cytosine, uracil and thymine (A, G, C, U and T) and their derivatives. Increasing this molecular variation by chemically modifying natural nucleic acids can increase their chemical and structural functionality in biotechnology (3, 4). Such synthetic nucleic acids are called xenonucleic acids. Many different types of modifications have been made to produce a large variety of xenonucleotides. The nucleobases can be modified, the natural nucleoside triphosphate can be exchanged by another leaving group, and the sugar-phosphate backbone can be modified or changed altogether into a different molecular scaffold. Since nucleotide leaving group modifications alone do not change the resulting DNA or RNA structure and so do not form an XNA, these are not discussed in detail, but an overview can be found in references (5–8).

### *1.1.1. Base modifications*

Most of the natural variation in nucleic acids already lies in the nucleobases. Besides the usual nitrogenous bases A, G, C, U and T, other variants exist *in vivo* like inosine, hypoxanthine, pseudouridine and dihydrouridine (present in tRNA), 7-methylguanosine (which caps the 5'-end of mRNA) in RNA, and 5-methyl cytosine and 5-hydroxymethyl cytosine in DNA. Most of these modified nucleotides are not incorporated by polymerases as substrates but rather are formed by dedicated enzymes that perform the modification after polymerization.

Synthetic analogues of nitrogenous bases can have several functions. Figure 1.1 shows a representative but non-exhaustive set of explored base modifications. Hundreds of fluorescent nucleobases of various sizes and absorption/emission spectra have been designed (9). Specific molecular tags can be added for purification purposes (like for example in biotinylated cytosine used in this dissertation). In both of these cases, the base-modified nucleotides are incorporated by natural polymerases. Xenonucleotides with modified bases can also act as polymerase inhibitory compounds, for example in antiviral medication against HIV (10). Finally, they can also be used to expand the natural "alphabet" of A:T and G:C base pairs, thereby increasing the potential sequence space. Ideally, these new bases should not pair with canonical bases, and only bind with their own partner. These bases can be self-pairing, resulting in an A:T G:C X:X alphabet (11), or form a new base pair resulting in an A:T G:C X:Y alphabet. An example of the latter has been shown to be able to be propagated stably on a plasmid *in vivo*. These nucleotides (d5SICS and dNaM) carry a 6-methylisoquinoline-1-thione-2-yl and 3-methoxy-2-naphthyl group where the base would normally be (12). Another set of nucleobase-modified nucleotides called DZA (13), where all four bases have small base modifications, could be efficiently used by Taq and Vent exo$^-$ DNA polymerases to elongate with high fidelity and in the kb length range. Yet this compatibility with natural polymerases came with resistance to a large range of restriction enzymes (14). Hoshika et al. (15) have recently reported on an expanded DNA/RNA system with eight bases, four of which are new, called hachimoji DNA and RNA, where the latter can be transcribed from the former using a thermostable T7 RNAP mutant.

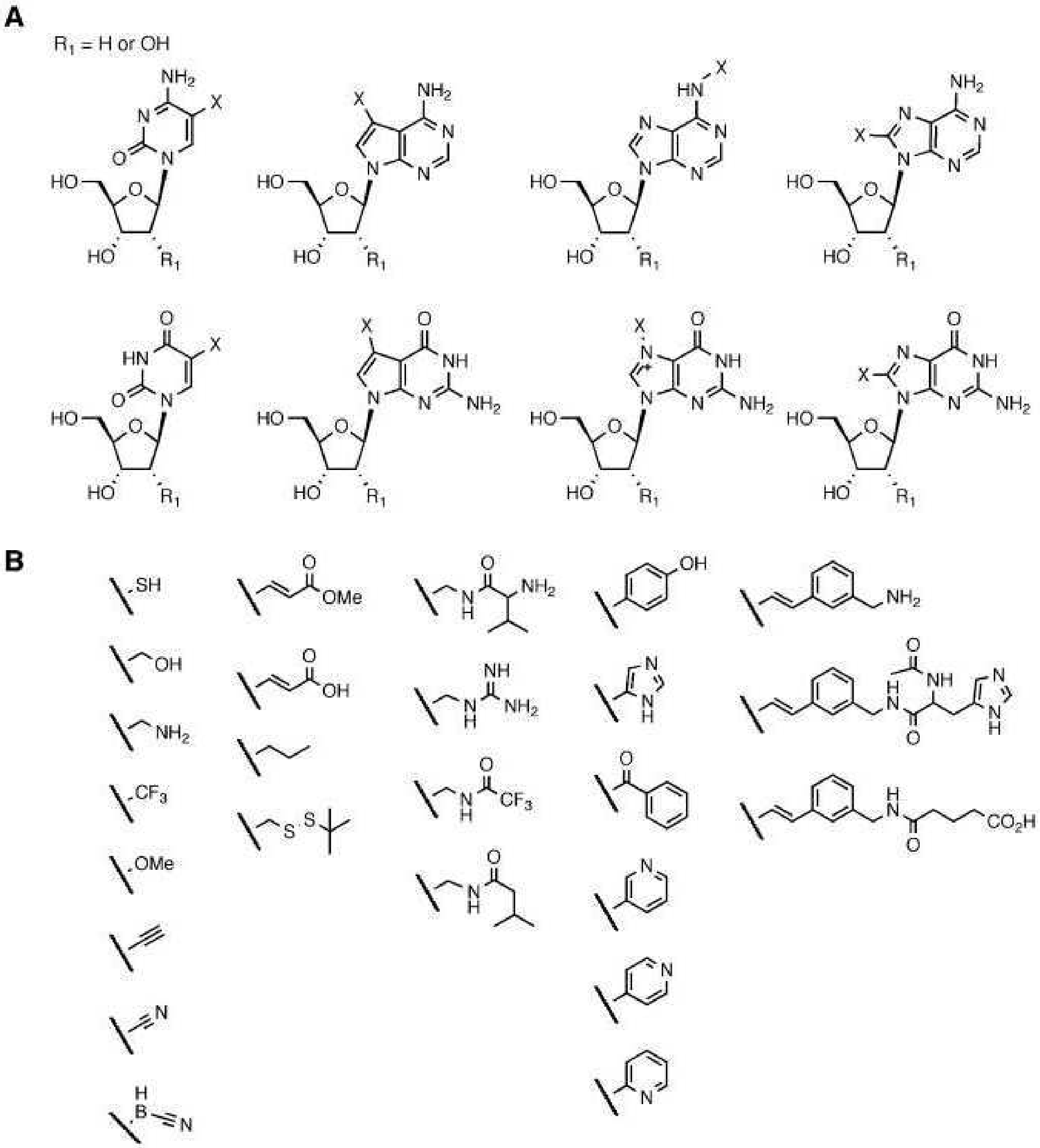


**Figure 1.1: Possible modifications of nucleobases that remain polymerase substrates.**

A) Positions where a modification can be made, marked with an X. B) A representative set of the more than hundred explored base modifications. Adapted from Brudno & Liu, 2009 (8) with permission of the rights holder, Elsevier.

### *1.1.2. Backbone modifications*

Since the natural sugar-phosphate backbone is relatively complex, many modifications have been explored. Similarly to synthetic nucleobases, these can be minor alterations or significant substitutions.

Among the more minor changes, there are phosphate analogues where a non-bridging phosphate oxygen has been substituted by S, Me, OMe, $BH_3$ or Se (8). Atomic substitutions, like oxygen to sulphur in α–thio-dNTPs, can be used to study reaction rates and determine chemical mechanisms of nucleic acid polymerization (16). More significant changes are those that have (deoxy)ribose analogues. These include isomeric sugars such as arabinose (forming arabinose nucleic acid or ANA) or (deoxy)xylose (forming (deoxy)xylonucleic acid or (d)xNA), hexitol (forming hexitol nucleic acid or HNA) and threose (forming threose nucleic acid or TNA). Modifications to the 2' position include fluoroarabinose (forming fluoroarabinose nucleic acid or FANA) and locked nucleic acids (LNA) where the 2'O and 4' position are linked by a -$CH_2$-group. In CeNA, the ring structure is a cyclohexenyl. The most drastic modifications change the nature of the linear backbone altogether, as is the case in peptide nucleic acids (PNA), glycerol nucleic acids (GNA), and amineDNA or aminePNA (see Figure 1.2).

Many of these XNAs can be formed and polymerized synthetically, but only a small number have nucleotides that are good substrates for existing polymerases that can also be extended enzymatically into longer XNA strands (17, 18). Since the enzymatic repertoire is so limited, new dedicated polymerases are required for their manipulation and production of longer strands in sufficient yields.

The xenonucleotides used in this study, 2'-deoxyxylonucleotides (dxNTs), form 2'-deoxyxylonucleic acid (dxNA), which is an XNA with a strongly distorted backbone compared to its stereoisomer DNA (19, 20). Because of this, even one 2'-deoxyxylonucleotide present in a dsDNA helix will cause a structural distortion. While two complementary strands of dxNA will hybridize to each other with an affinity that is similar to that between ssDNAs, they did not bind to complementary strands of DNA (21). This is due to the fact that a dxNA backbone is much more elongated, left-handed instead of right-handed, with a much larger major groove (22) (See Figure 1.3). These properties provide dxNA with an informationally (and genetically) orthogonal potential.

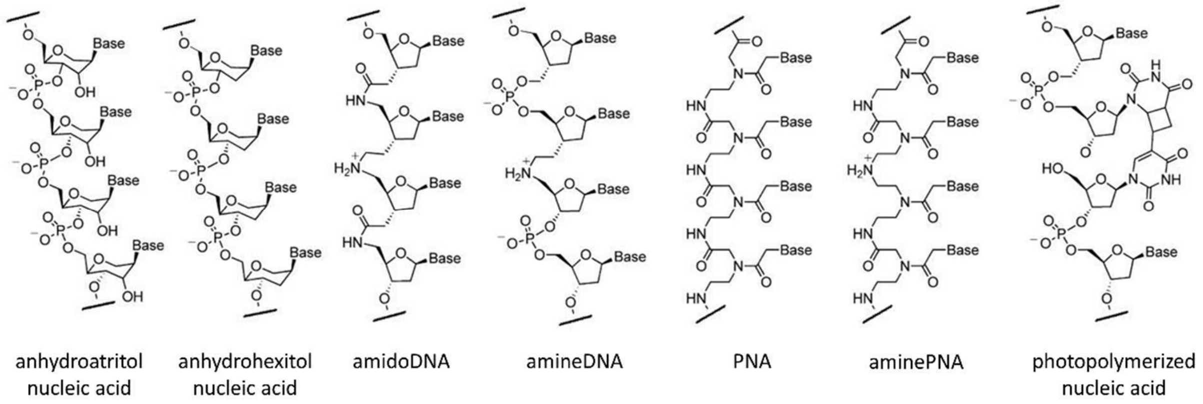

anhydroatritol nucleic acid
anhydrohexitol nucleic acid
amidoDNA
amineDNA
PNA
aminePNA
photopolymerized nucleic acid
Base
OH
HN
$H_2N$

**Figure 1.2: A representative but not exhaustive set of XNAs with modified backbone structures.**

1,5-anhydroatritol nucleic acid and 1,5-anhydrohexitol nucleic acid have a modified sugar analogue, amido and amine DNA differ at the phosphate moiety. Peptide nucleic acid (PNA) and aminePNA lack a sugar phosphate backbone altogether. Photopolymerized nucleic acids are polymerized by covalently bound nucleobases. Adapted from Brudno & Liu, 2009 (8) with permission of the rights holder, Elsevier.

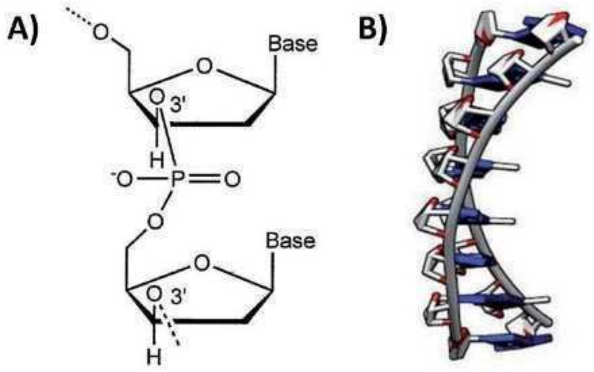


**Figure 1.3: Structure of a dxNA backbone.**

The 3’-OH group of a dxNT is in the β-orientation rather than the α-orientation in deoxyribose (A), which causes the helix of dxNA to be much more elongated and slightly left handed (B). Adapted from Maiti *et al*., 2012 (20) with permission of the rights holder, John Wiley and Sons.

## 1.2. Orthogonality

Synthetic biomolecules are said to be orthogonal to natural biomolecules when this one biological component does not induce side effects in other ones (1). In other words, there are no specific unwanted interactions with other natural biomolecules or metabolic pathways, or vice versa. For example, an XNA that does not hybridize to DNA or RNA and that can only be used as a template by a specific XNA polymerase instead of natural polymerases, is informationally orthogonal since its sequence information cannot flow from the XNA to DNA/RNA. However, if the nucleotide building blocks themselves still interfere with cellular metabolism by having to be phosphorylated by an ATP-dependent kinase *in vivo*, these are not energetically orthogonal. When designing a synthetic nucleic acid for usage *in vivo* as an extra chromosome next to the existing DNA genome, it would be favourable to have it not be independent of the existing metabolism to retain the practical benefit of not having to design an XNA metabolism. If the sequence information in the XNA chromosome is to stay on the chromosome, only a genetic “firewall” is necessary by achieving informational orthogonality, and the nucleic acid and its building blocks just cannot act as unwanted substrates or inhibitors when “plugged in” to the existing metabolism. In the case of *in vivo* XNA systems this means that an orthogonal XNA polymerase is dedicated to XNA and has negligible to no affinity for DNA, RNA or their nucleotides.  The same goes for other nucleic acid enzymes like helicases, primases, ligases, nucleases etc. and eventually even ribosome analogues. To be fully orthogonal, the natural enzymes and nucleic acids should not interact with the XNA either.

Besides nucleic acids, proteins can also be rendered orthogonal. Proteins that contain unnatural amino acids themselves, possibly translated from a natural DNA/RNA system with an altered or expanded codon table (23), are already informationally orthogonal as long as the central “dogma” of molecular biology holds (sequence information only goes from nucleic acid to either nucleic acid or protein, never from protein to nucleic acid).

There are also cases where full orthogonality is unwanted, for example in the case of antiviral nucleotide analogues that are  designed to specifically inhibit the reverse transcriptase of the AIDS virus (HIV-RT) (10). Examples of these modified nucleotides are Zidovudine (also known as azidothymidine or AZT), Abacavir, Lamivudine, Emtricitabine and Tenofovir disoproxil. The ideal anti-viral nucleotide analogues are orthogonal to human cell biology, but interfere with the replication cycle of the virus by inhibiting this key polymerase.

## 1.3. Aptamer technology

Aptamers are relatively short, single-stranded nucleic acids that fold into well-defined three-dimensional structures, and are selected for a specific interaction with a target molecule (24). They have been used as a structural scaffold, sensor, probe, or therapeutic molecule (25). As a scaffold, they are also used to direct delivery of other structures, like nanoparticles with drugs encapsulated in them, to specific cells. They can be seen as the nucleic acid version of antibodies since they have a similar mechanism of action. However, while monoclonal antibodies need to be produced *in vivo*, aptamers can be synthesized *in vitro* and can be the subject of molecular evolution more easily and are less prone to contamination. They are also less immunogenic than antibodies, but because of their smaller size, they also have shorter half-lives due to their susceptibility to renal filtration. Because of their size and strong negative charge, they will not easily cross membranes, so the easiest *in vivo* targets for aptamers are extracellular in the circulatory systems and at the outside of the cellular membrane. If the target is in the cytosol, they need to be able to cross the cell membrane.

One method of obtaining aptamers is systematic evolution of ligands by exponential enrichment (SELEX) (26). Usually, a very large library of up to $10^{15}$ oligonucleotides is made, with a randomized sequence of 20-100 nucleotides long, flanked by specific sequences for enzymatic manipulation. For DNA SELEX, ssDNA strands can be PCR-amplified and separated from the double-stranded DNA products, while ssRNA SELEX libraries are made by *in vitro* transcription using for example T7 RNA polymerase. Selection occurs by affinity chromatography for the intended target (either in pure form or in complex mixtures), resulting in enrichment of sequences with the intended biological activity that can then be re-amplified, and subjected to multiple rounds of selection and mutation to increase the activity. SELEX can be used on XNA aptamers as well in principle, as it has been shown to be possible to evolve an XNA aptamer with two of four modified bases (5-CldUTP and 7-deazadATP) as a high affinity protease inhibitor using this technique (27).

Because of the ubiquitous presence of DNase and RNase, unmodified DNA or RNA aptamers will tend to have short half-lives, which hinders their potential use *in vivo*. For this reason, they have been either made with modified nucleotides like for example 2'-amino pyrimidine RNA nucleotides (28), 2'-fluoro pyrimidine nucleotides (29, 30) or 2'-O-methyl nucleotides

(31), or they are modified after their synthesis with for example polyethylene glycol (PEG) added to the 5’ end (32).

However, while nucleotides that differ drastically from the canonical ones can build more nuclease resistant XNAs, these substrates could in turn be used less efficiently by polymerases and hinder replication and (reverse) transcription steps during the SELEX procedure. On the other hand, applying modifications after selection potentially lowers their affinity and specificity for their target, since these modifications were not present during selection. A modified synthetic nucleic acid that already contains the wanted modifications that make it degradation resistant or add specific functional groups that will be present during the selection step, such as the previously mentioned DZA (14), is more efficiently produced since it is produced enzymatically instead of synthetically. An aptamer can become partially or fully orthogonal if its sequence information cannot be read by the natural replication, transcription or translation machinery, or if it is resistant to nucleases.

## 2. DNA polymerases

### 2.1. DNA replication

The general mechanism of DNA replication in cellular life is universal (33). Double-stranded DNA is unwound by a helicase, the single-stranded DNA is protected by ssDNA-binding proteins and a primase adds an RNA primer to provide a 3’-OH group. This primer is used by the DNA polymerase to extend the DNA, which has a drastically increased processivity due to the presence of the sliding clamp (or β-clamp) which tethers the polymerases to the DNA. Because DNA replication occurs by melting the original duplex, creating two new duplexes each carrying one original and one copied strand, DNA replication is semi-conservative (34).

Bacterial DNA replication is initiated at a specific AT-rich sequence, called origin of replication. DNA strand synthesis is unidirectional in the 5’-3’ direction. However, simultaneous replication of both strands occurs in replication forks, causing one to become the leading strand and the other the lagging strand. The synthesis of the leading strand is continuous, while that of the lagging strand happens in discontinuous, so-called Okazaki fragments.

Several other enzymes are active during DNA polymerization. Primer RNA is removed and replaced by DNA repair polymerases, and the resulting nicks are closed by ligases. The torsion on the helix due to helicase-induced unwinding is relieved by topoisomerases to avoid excessive supercoiling (34).

### 2.2. Classification

DNA polymerases are classified in eight families: A, B, C, D, E, X, Y and RT, based on their sequence homology within these groups (35–37). No sufficient sequence homology between these families is found to produce a complete phylogeny of all DNA polymerases, which could in principle allow for the possibility of convergent evolution of different DNA polymerase families evolving independently. However, when structural information (geometry, secondary structure and physicochemical properties of amino acid residues)  is used instead of linear sequence alignment alone (38), the right-hand shaped polymerases of which such high resolution structural information is available (DNAPs from families A, B and Y, several DNA-dependent RNA polymerases (DdRps), RNA-dependent RNA polymerases (RdRps) and reverse transcriptases (RTs)) are found to share a common core of 57 α-carbon positions, where only the two catalytic Asp residues are retained (38). These lie in a conserved core within the palm and fingers domains. This structure-based analysis also produced clustering consistent with the six families, both between and within these family branches. However, viral RdRps do not follow phylogeny based on coat protein folding, virion architecture and genome structure, which is hypothesized to be caused by viral horizontal gene transfer above the viral family level, since replicative polymerases are relatively independent of the rest of the viral infection and assembly. The same is the case for potentially phage-derived mitochondrial family A DNAPs and alphaproteobacterial family B DNAPs.

Family A contains both replicative and repair DNAPs, and these are found in metazoa, plants, bacteria and phages. Most of them contain 3’-5’ exonuclease domains for proofreading, and a 5’-3’ exonuclease domain to remove RNA from Okazaki fragments during replication.

Family B is made up of mostly replicative polymerases from eukaryotes, archaea and even bacteriophages like T4 and RB69. They usually have strong 3’-5’ exonuclease activity. Therminator DNAP, the polymerase studied in this thesis, is a member of family B. While family B DNAPs replicate chromosomal DNA in eukaryotes, knockout mutation experiments showed that in *Thermococcus kodakaraensis* and *Methanococcus maripaludis*, DNAP of family D, rather than B, is necessary and sufficient for genomic replication (39, 40). The euryarchaeote *Pyrococcus abyssi* however requires both (41). The N-terminal domain has a uracil binding pocket that can be used to scan the template strand for ribonucleotides.

Family C is mostly comprised of replicative DNAPs from bacteria (besides cryptic phages and plasmids). They are large multi-subunit complexes, where the one with catalytic polymerase activity as well as 3’-5’ exonuclease activity is subunit α.

Family D is confined to euryarchaeota. They are active as heterodimers or heterotetramers, where the largest subunit is the catalytically active one and the small one has 3’-5’ exonuclease activity.

Family E so far consists of a single polymerase enzyme from the pRN1 plasmid of the thermoacidophile crenarchaeote *Sulfolobus islandicus*. Besides DNA polymerase activity, it has primase and ATPase activity. The N-terminal half of the protein contains the primase and polymerase functions, and shares no homology with other known primases or DNA polymerases (42).

Family X are present in bacteria, archaea, eukaryotes and several viruses. The eukaryote polymerase β has a larger polymerase domain, but a smaller domain is also involved in base excision repair (BER) which cuts out abasic sites in damaged DNA. Phylogenetic analysis suggests that family X arose in bacteria (most closely resembling those of *Bacillus*), from which polymerase IV then branched off, followed by the branching of sister clades μ and TdT, and that of β and λ after gene duplications. Vertebrates are the only group where the μ, TdT, β and λ polymerases are all present (43). TdT (terminal deoxynucleotidyl transferase) is a unique DNA polymerase in the sense that it polymerizes nucleotides in the absence of a template strand. Its structure is extremely similar to that of the other family X DNAPs, but because of the presence of a peptide loop where a DNA template would normally bind, it can only perform primer extension in the absence of template. It can also use a variety of metal cofactors ($Mg^{2+}$, $Zn^{2+}$, $Co^{2+}$ and $Mn^{2+}$). Its physiological role lies in generating variation in VDJ recombination in vertebrate adaptive immunity by generating *de novo* sequence information at specific sites (44).

Family Y polymerases can bypass damaged DNA (45) and are found in all three domains of life. Their error rate (without 3'-5' proofreading) is around $10^{-4}$ to $10^{-2}$, making them DNA polymerases with very low fidelity as well as low processivity. Once the lesion is bypassed, replicative polymerases with high fidelity and processivity (that stall on these lesions) can displace them and continue replication.

The last family of DNA polymerase are the reverse transcriptases (RT), which differ from the other DNA polymerases in that they use RNA instead of DNA as template. This group contains retroviral RTs that transcribe viral ssRNA genomes to proviral ssDNA, and telomerases, that use an RNA template to elongate telomere 3' ends with short ssDNA repeats (used by regular DNA polymerases to make dsDNA).

### 2.3. Structure

The folded overall structure of DNA polymerases is usually C-shaped, and its domains are named after parts of a human right hand(46). These are the fingers, palm and thumb domain, although other domains can be present, and many functional holoenzymes are heteromeric. The structurally most conserved domain is the palm, with a typical β1-α1-β2-β3-α2-β4 fold.

In Therminator and structurally similar DNAPs the fingers bind and position the incoming nucleoside triphosphate substrate, the palm contains the universally conserved carboxylate residues that catalyse the polymerization reaction, and the thumb binds the downstream primer-template complex. Therminator, being an archaeal family B DNAP, also has an N-terminal domain and its own 3’-5’ exonuclease domain (see Figure 1.4).

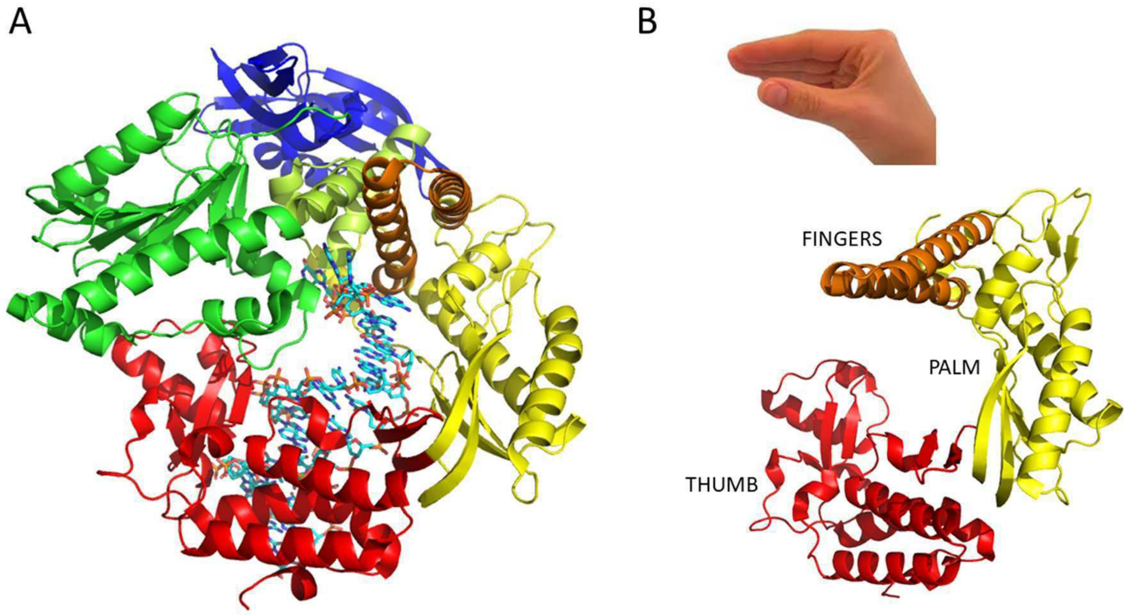


**Figure 1.4: Structure of the Therminator DNA polymerase precursor DNAP 9°N-7 (PDB: 4K8X).**

A) Crystal structure of the polymerase binary complex with the bound dsDNA primer/template complex in the closed state. Blue: N-terminal domain. Green: 3’-5’ exonuclease domain. Yellow: palm domain. Orange: fingers domain. Red: thumb domain. The dsDNA template is shown in light blue with coloured heteroatoms. B) Rotated view of the fingers, palm and thumb domains to show the structural similarity with the human right hand. Protein structure figures were created using PyMol.

### 2.4. Catalytic cycle and reaction mechanism

The catalytic cycle of a polymerase is relatively complex and a simplified version is shown in Figure 1.5 (35). The first step is the polymerase enzyme (E) binding to the primer-template complex (p/t). Next comes the binding of the nucleotide, followed by conformational changes from the open towards the closed state (an aligned state for correct nucleotides or a misaligned state for mismatches (47)). The following step is the chemical reaction that binds the α-phosphate of the nucleotide substrate to the 3’-OH group of the primer, thereby extending it. Pyrophosphate acts as the leaving group, which then dissociates, together with a conformational change to restore the open state. Then the polymerase can either dissociate from the $p_{+1}$/t complex (distributive polymerization) or remain bound and bind the next dNTP substrate. The essence of this mechanism is the same for all polymerases, but the rates of each step can vary strongly between different polymerases, as well as which step is the rate-limiting one or how many and which specific conformational changes occur. This reaction mechanism also does not include reversing some of these steps due to exonuclease activity. Some specific structural details are given for DNA polymerase Therminator.

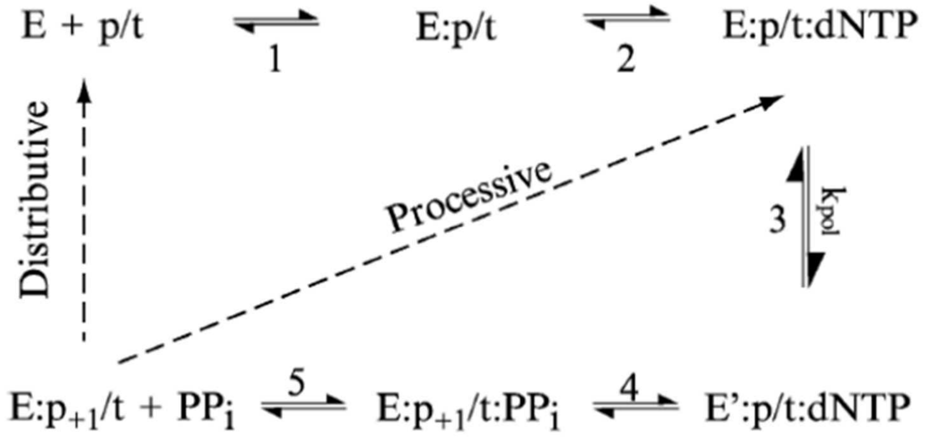


**Figure 1.5: The catalytic cycle of a DNA polymerase.**

Step 1: binding of enzyme E in the open state to primer-template complex p/t. Step 2: binding of the nucleoside triphosphate to this E:p/t complex. Step 3: conformational change of the enzyme to the closed state E'. Step 4: condensation of the nucleotide to the primer 3' end with pyrophosphate ($PP_i$) as leaving group. Step 5: dissociation of the leaving group, followed by either enzyme dissociation, or processive polymerization with the next dNTP after translocation of the $p_{+1}$/t complex. Modified from Rothwell & Waksman, 2005 (35) with permission of the rights holder, Elsevier.

### *2.4.1. Formation of the E:p/t complex*

The binding of the primed DNA duplex happens primarily through the thumb domain. In DNA polymerase Therminator, this thumb has a highly flexible C-terminal end and is relatively large compared to the thumb of many other polymerases. It mainly interacts through positively charged residues and with negatively charged phosphates of the DNA backbone, forming a positively charged cleft through which the DNA can translocate. Closer to the active site there are also several positively charged residues that move into the minor groove of the newly formed helix. The palm domain also interacts with the minor groove. In the Klentaq DNAP, the first single-stranded templating base is separated from the rest of the template bases through insertion of a Tyr residue that stacks aromatically with them (35).

### *2.4.2. Formation of the E:p/t:dNTP complex*

The binding of the dNTP occurs through binding with and movement of the fingers domain. The fingers domain consists of two α-helices. The region of the helix closest to the active site (the O-helix in DNAP Klentaq (35)) contains several positively charged residues that interact with the phosphates of the dNTP. The fingers domain is hinged and can move towards and away from the palm domain, forming the catalytically active complex when in the closed state and the dNTP is bound. This movement of the bound nucleotide covers a distance of roughly 10-15 Å. However, in family B DNAP RB69, kinetic studies suggest that this is not necessarily the case, and dNTPs could also diffuse directly into the active site, where only those nucleotides that fit due to Watson-Crick base pairing and the correct sugar moiety (deoxyribose versus ribose for example) can stay bound long enough for the rate limiting step to occur (48). Discrimination of incorrect nucleotides can happen during binding. Most replicative polymerases have a much larger $K_D$ for incorrect nucleotides, while for example repair polymerases tend to have lower fidelity, although they still strongly favour correct nucleotides. Therefore, mismatching dNTPs will bind less strongly or dissociate more easily. Discrimination between RNA and DNA nucleotides in RNAPs and DNAPs usually happens with a “steric gate” residue, which will either need an interaction with a 2’-OH group for RNA, or takes its place in the case of 2’-deoxyribose which sterically hinders incoming NTPs in a DNA polymerase. This discrimination is especially necessary in DNAPs, since *in vivo*, NTPs are about an order of magnitude more abundant than dNTPs. The other recognition is the fitting of the proper base pairs, adenine against thymine with two hydrogen bonds, and guanine against cytosine with three hydrogen bonds. The pattern of H-bond donor and acceptor groups on the nucleobases is important to sample and bind dNTPs in the correct orientation long enough for polymerization to occur. However, a large part of the binding enthalpy also comes from aromatic stacking of the bases, and hydrogen bonds are not necessary for successful specific recognition between fully hydrophobic bases (49). The size of the base pair also matters: purine-purine or pyrimidine-pyrimidine binding pairs are too wide or too narrow respectively to fit in a normal DNA helix, which is a possible reason for why transition mutations occur more often than transversions (50). In T7 DNAP (47), the binding energy of nucleotide substrates contributes to the formation of either a closed aligned or closed misaligned state, meaning that binding a mismatched dNTP could actively influence the orientation of residues necessary for catalysis, thereby decreasing the incorporation rate and increase the release rate.

### *2.4.3. Catalytic step*

The E’:p/t:dNTP complex is the state in which the polymerase reaction occurs. While DNA polymerases vary greatly in protein sequences and overall folds, the parts that catalyse the incorporation of a dNTP on a growing primer strand function in the same way, and the catalytic mechanism is universal (51). The active site residues coordinate the 3’-hydroxyl group of the primer, the α-phosphate from the nucleotide and two catalytic $Mg^{2+}$ ions (see Figure 1.6).

Among the most important catalytic residues are the two universally conserved aspartates. The reaction itself is a nucleophilic attack by the 3’-OH on the nucleotide α-phosphate, which has a pentacoordinated transition state that is stabilized by both $Mg^{2+}$ ions (in turn correctly oriented by the aspartates), releasing the pyrophosphate as the leaving group. Hydrolysis of pyrophosphate makes this reaction irreversible and avoids product inhibition. *E. coli* DNAP IV (family Y) has recently been shown to have an intrinsic pyrophosphatase activity (52), and since PCR using many different DNAPs readily occurs without the need for additional pyrophosphatases, it is likely that this is a general mechanism in DNA polymerases to drive the polymerization reaction in the forward direction. One of the $Mg^{2+}$ ions also activates the 3’-OH for the nucleophilic attack, while the other chelates and stabilizes the pyrophosphate group.

Studying the reaction mechanism has seen a wide variety of techniques and strategies (16). Chemical quenching methods were used first to outline the basic reaction pathway. Later, the non-covalent steps, like conformational changes, have been studied using fluorescence-based approaches.

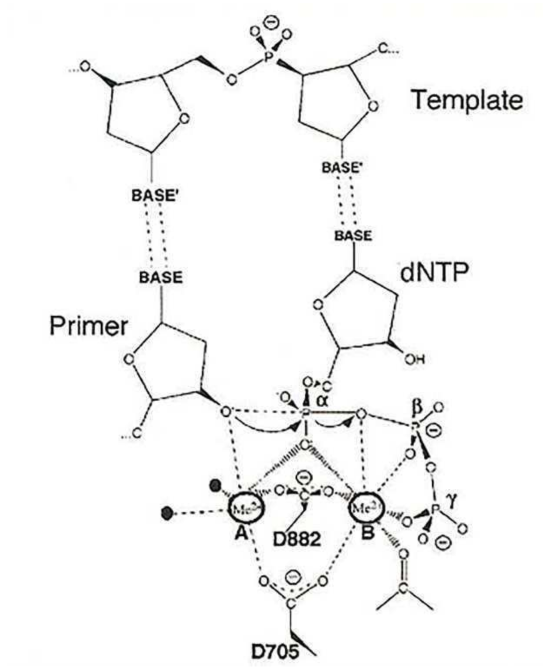


**Figure 1.6: Schematic overview of nucleotide polymerization reaction mechanism.**

Amino acids are numbered according to the Klenow fragment. A and B are the catalytic $Mg^{2+}$ ions. Black dots are coordinated water molecules. Adapted from Brautigam & Steitz, 1998 (51) with permission of the rights holder, Elsevier.

The covalent steps can be studied using steady-state, single-turnover, or burst kinetics. Steady-state experiments are easier to carry out, and require less enzyme and no fast kinetics instruments, but results are harder to interpret since the rate-limiting step is potentially a mixture of different rates. The $k_{cat}/K_M$ ratio does allow for comparisons between the efficiency of different substrates.

In single-turnover experiments, polymerase is in excess over DNA, and the observed rate $k_{pol}$ is that of the slowest step up to the moment of phosphoryl transfer (53). These rates are usually in the order of tens to hundreds of incorporations per second, requiring rapid-quench-flow technology. Evaluating whether or not the chemical step is rate limiting can be done using the sulphur elemental effect (from an α-thio-phosphate), comparing pulse-chase and pulse-quench yields (a difference implies a fast chemical step flanked by two slower steps) or the solvent deuterium isotope effect (which suggests that in many polymerases, the transition state is accompanied by two proton transfers).

In burst kinetics, the concentration of DNA and enzyme is comparable, with only two or three times more DNA (53). The first primers are extended at a faster rate, followed by slower dissociation and re-association, and if these rates differ about 20-fold, this is a biphasic reaction. Observing this type of kinetics is then evidence for a rate-limiting step after chemistry, while not observing it means the limiting step happens at or before the chemical step.

In contrast to chemical quenching, fluorescence studies can also show steps that are not rate-limiting, they can be independent of product formation, and are ideal for studying structural rearrangements of substrates and enzymes. Examples of fluorescent probes are 2-aminopurines or a variety of FRET probes. 2-AP is quenched by neighbour stacking, so when 2-AP is on the template +1 position, there is a large increase in fluorescence, which is consistent with the next base being flipped out. Using Förster resonance energy transfer (FRET) with a fluorescent base analogue FRET donor/acceptor within the primer-template complex and a FRET partner bound to the moving part of the fingers domain (54–56), it was shown that in Klentaq and PolI(KF) the fingers close before the chemical and rate-limiting steps. Fluorescence-based probes are also useful for determining if DNA is in the bound or unbound state.

### *2.4.4. Dissociation of the pyrophosphate leaving group and translocation*

After the condensation of the nucleotide resulting in the elongation of the primer, a next conformational change allows the release of the pyrophosphate group and translocation of the DNA duplex by one base pair for the next incorporation event. Different polymerases have different affinities for the $PP_i$ product and therefore the potential of product inhibition. In the Klenow fragment, the affinity for $PP_i$ is only five times lower than for the p/t complex (57), while it is nearly a 1000 times lower in T7 DNAP for example (58). Pyrophosphatases therefore make the reaction irreversible. Given that the surface of the thumb domain that interacts with the p/t complex is positively charged, it is hypothesized that the nucleic acid helix is relatively free to move along its length axis within this electrostatic field, thereby making room for a new nucleoside triphosphate (35).

## 3. Protein design and molecular evolution

### 3.1. Mutant libraries

New proteins can be generated by modifying existing ones, or by designing them from scratch. In both cases, a directed, rational approach can be taken using theories of structure-function relationships, where specific mutations to an existing protein are made and then tested experimentally, or a protein is designed from scratch. Another strategy is using molecular evolution, where a large population of random variants are selected for a certain function. These can be mutated variants of an existing protein, or consist of a fully randomized library.

While the total sequence space of all potentially existing proteins is enormous, natural proteins only represent about 1000 to 1400 folds or topologies (59–61). Natural evolution has therefore only sampled a tiny part of all possible proteins. Of course, natural proteins are extremely rich in functional sequences compared to the total sequence space, the majority of which is probably not functional. Still, certain unexplored "islands of function" may exist, as shown by the finding of four different ATP-binding proteins in a single experiment with folds that do not correspond to any known protein (62). These proteins were found in a library with a size of 6x$10^{12}$, one of which was improved by further rounds of mutagenesis (63). Another example is a study that restored fd phage infectivity after deletion of the g3p minor coat protein, with a much smaller randomized library size of $10^{6}$. It even proved possible to evolve a functionally similar protein after seven generations of selection from a single random starting point (64).

Designing *de novo* proteins therefore needs a good understanding of protein structure-function relationships. Molecular evolution on the other hand requires no insight, but is constrained by having to sample close by in the sequence space of existing proteins. If a protein with the wanted function does not exist within a realistically close distance, it simply is not practical to evolve it.

Still, the advantage of molecular evolution lies in the high throughput of sequences that can be tested, which increases every time screening techniques allow for larger libraries to be sampled. In this way, brute force can both cover all the rational options and also stumble on solutions that a rational approach would have missed, which is popularly expressed by Leslie Orgel's second rule: "Evolution is cleverer than you are" (65).

Molecular evolution has been used to evolve several polymerases with increased efficiency for XNA production, usually from DNA templates (17, 66). The first step in molecular evolution is usually to start with an enzyme that is likely or known to have promiscuous activity for the wanted reaction or function (67–69). This promiscuous protein is then mutated, for which many different methods exist that can differ greatly in how targeted they are. An alternative is searching for proteins such as antibodies that are high-affinity binders for transition-state analogues (70), although these tend to have low catalytic rates compared to regular enzymes.

One of the untargeted approaches is the addition of mutagenic chemicals to transformed cells that contain the gene of interest (71). These are usually dNTP mimics. Since they will interact with metabolism and replication of the whole cell and not just the gene of interest, these are very toxic and can also mutate other genes. They can also have a very skewed mutational spectrum, causing a heavily biased library. A slightly less invasive technique is using strains that are naturally mutagenic, so called mutator strains, such as *E. coli* XL1-RED (72). These lack mismatch repair enzymes, causing them to have up to thousand times higher background mutation rates, and the absolute mutational load can be controlled by the amount of generations. The plasmid is then isolated and cloned in a strain with higher fidelity for further experiments. A more targeted approach is error-prone PCR (epPCR), where only the gene of interest is mutated, by performing PCR using lower fidelity polymerases, unequal nucleotide concentrations, or mutagenic cofactors such as the presence of $Mn^{2+}$ in addition to the catalytic $Mg^{2+}$ (73). For example *Taq* polymerase has an error rate of $5.5x10^{-4}$, which can already be sufficient for mutagenizing long fragments. Mutational load can also be increased by increasing the amount of PCR cycles and starting with low amounts of template. epPCR also produces mutation spectra that are less biased. In error-prone rolling-circle PCR, a circular template is replicated by ϕ29 polymerase into many linear tandem repeats in one step and used directly to transform, which increases efficiency (74).

Specific regions within the gene of interest can be rationally targeted, for example with synthetic degenerate oligonucleotides (75, 76). This allows for high local diversity, without introducing so many off-target mutations that the gene would likely become non-functional. This strategy was used is this thesis, by performing PCR with high fidelity enzymes but with primers that are partially or fully randomized in specific, well-chosen regions. This method allows very tight control over the position and type of mutations that will occur. A necessity for good focussed libraries is a good understanding of the structure-function relationship of DNA polymerases (77).

Variation does not always need to be introduced as new mutations to existing sequences. New functions can be found be recombining existing sequences (78). These can be sequences that have not yet been under selection, or mutant clones that have been found by screening. In case of the latter, when mutations exert their beneficial effect at least in part in isolation, a superior search strategy can be to combine multiple of these beneficial mutations in one clone. This mirrors evolution as it works in nature. When fitness landscapes are sufficiently flat, adaptive genotypes can be found by multiple small mutational steps over multiple generations. Only fitness peaks that are separated from the starting template by wide fitness valleys will necessarily require large mutational steps which are more unlikely. Combining independent beneficial mutations also happens in nature and can accelerate the evolution of traits, for example by homologous recombination or horizontal gene transfer (the movement of genetic material between unrelated organisms, as opposed to the vertical inheritance of traits from parent to offspring). All these processes allow the combination of beneficial mutations without the need for independently "inventing" them.

### 3.2. Selection strategies

Once the variation is created, it is necessary to separate the variants that have more desired functionality from those that are unchanged, less active, or inactive. Very small library sizes allow for screening, where technically each variant can be tested individually. However, the chance of this library to contain very good variants is low. The larger the library is, the larger the odds of the wanted variants being present, but screening becomes practically unfeasible. Selection can then be used by introducing a process where genotypes that code for more functional phenotypes are allowed to amplify, hence enrich over the others, which can then be analysed. It is important that the function (phenotype) remains physically linked with its coding variant (genotype), since selection acts on the phenotype but analysis and further manipulation requires the genotype. This genotype linkage can be achieved through physical binding, such as through compartmentalization, for example in water-in-oil emulsions, or in phage display. A schematic overview of the discussed selection strategies is given in Table 1.1.

### 3.2.1. *Compartmentalized self-replication (CSR)*

CSR is an emulsion-based method of selection that is designed specifically for polymerase evolution (79). A schematic overview is given in Figure 1.7. It is based on a feedback loop in which a polymerase has to amplify its own gene. Polymerases that are more active in the chosen selective environment will produce more copies of their own gene, increasing the proportion of active variants in the population. The reactions take place in water microdroplets surrounded by an oil phase. Each water droplet contains at most one *E. coli* cell expressing the polymerase gene from a plasmid, and all the necessary components for the amplification reaction that follows a cell lysis step, such as primers and nucleotides. The resulting amplified fragments are then re-cloned, and the cycle can be repeated multiple times to keep enriching the library in active variants, until only a small number of the best mutants remain, which can then be characterized. There are several traits that can be selected for. Among these are the reaction temperature optimum, resistance to inhibitors, increased tolerance for DNA lesions in ancient DNA (which allowed amplification of paleo DNA from 47,000-60,000 year old cave bear DNA using an evolved version of Taq DNAP) (80). Even in high temperature regimes in these PCR-like reactions, the droplets remain stable and do not exchange large macromolecules like the plasmids or expressed proteins, preserving the genotype-phenotype linkage. Because of the heat lysis and PCR-like reaction conditions, this technique is usually limited to thermostable polymerases that do not denature together with the native *E. coli* proteins. It is however also possible to achieve cellular lysis by freeze-thawing, which made the molecular evolution of ϕ29 DNA polymerase possible using isothermal CSR (81).

In the case of XNA research, modified nucleotides can be used to evolve dedicated XNA polymerases. When the products are XNA rather than DNA fragments, reverse transcription to DNA during or after the selection step  is necessary for re-cloning, which is not a trivial step in most cases.

A variant on CSR is short-patch CSR (spCSR) (82). Here, rather than replicating the whole gene, only a fraction of the gene is amplified, more specifically the part that contains the introduced variation. This is useful when selection conditions are too harsh for most polymerases to complete elongation of the full gene, even if some variants exist that do have increased activity compared to the wild type.

Finally, it is also possible to use this method to select on proteins other than polymerases, if the activity of this protein can indirectly be linked to gene amplification. An example is a nucleoside kinase in the presence of dNMPs or dNDPs and a source of phosphates, so that the only kinases that efficiently produce dNTPs necessary for replication of their coding sequence can be amplified (83).

A potential caveat that extends to all methods of selection is that "you get what you select for" (65). This means that any mutation that confers higher fitness will be selected for, even if it is not the activity the protein designers had in mind. For example, if thermal melting of the DNA is a necessary and reaction-limiting step, mutations that lower the GC-content can be selected for. There will also be selection on polymerase solubility and expression efficiency, and when there can be no control for the effective protein concentration in a reaction that is protein-limited, mutants with these traits will also be selected for, even if the activity of an individual protein species is unchanged. And while these traits can at least be desirable side-effects, truncation or deletion mutants that drastically lower the size of the fragment can be much more rapidly amplified and cloned, even if they are completely inactive. If such genetic parasites are produced in the re-cloning step with a selective benefit that outweighs the selective benefit of active variants during the selection step, a library can become unusable since it will keep becoming enriched in inactive variants.

#### *3.2.1. Compartmentalized self-tagging (CST)*

When CSR or even spCSR requires sequence amplification that is too long for the polymerase under the selection conditions, or a reverse transcriptase that is necessary for the re-cloning step is not available, CST can be used (84). As the name suggests, the polymerase will tag its own coding sequence rather than replicating it. This tag can then be used to isolate the active variants using affinity chromatography. This technique is especially useful for DNA-dependent polymerases that can only incorporate a handful of xenonucleotides that are orthogonal to reverse transcription. It is based on the same water-in-oil emulsion techniques of CSR. The major difference is the presence of only one primer, which will carry a tag (see Figure 1.8A). This tag can either be present on the primer itself, or on a nucleotide incorporated by the polymerase. A necessity for this technique is that the primer, once elongation has occurred, has increased affinity for the plasmid, or at least remains hybridized to the plasmid for long enough for the plasmid-primer-tag complex to survive the isolation step, otherwise the genotype/phenotype link is broken. Extraction can be achieved with paramagnetic beads that are coated in a molecule with high affinity for the tag, such as streptavidin for biotin tags (See Figure 1.8B).

### *3.2.1. Phage display*

Using this method, a target gene is inserted into a phage (usually filamentous bacteriophage M13 or fd) or into a derived phagemid vector and fused to a gene coding for a viral coat protein in such a way that when the virion is assembled, the target protein is displayed on the surface of the phage particle (85) (See Figure 1.9). These phagemids can be amplified in *E. coli* $F^+$ cultures together with a helper phage that produces the other necessary proteins for the formation of the complete phage particle. Superinfection does not take place, which assures that the protein variant that is displayed on the coat proteins is coded by the ssDNA inside the phage. Phages displaying more active proteins are then selected for.

Phage display was originally developed as a way to select for short peptide binders (86) and antibodies (87) but can also be used for evolving enzymes (88) such as DNA polymerases (89, 90).

A main advantage of phage display compared to emulsion-based techniques is the large library sizes that can be used (larger than $10^9$) due to the small and highly stable phage particles. The libraries can also be efficiently re-cloned by infection.  More toxic proteins can also be used, since their expression is delayed until phage infection. A prerequisite however is that displayed proteins are efficiently transported through the cell membrane and do not interfere with phage assembly and infection. Selecting on enzymatic activity is also less trivial than selecting for binding affinity. However, the used proteins do not have to be thermostable since no cell lysis or DNA melting is required, and in the case of polymerase evolution using phage display, the genotype-phenotype linkage can be covalent and therefore strong and reliable (90).

**Table 1.1: Comparison of different selection strategies for XNA polymerases**

| Selection technique | Advantages | Disadvantages |
|---|---|---|
| (sp)CSR | - direct selection on wanted activity<br>- thermal lysis deactivates enzymatic background | - requires strong promiscuous activity for long and accurate elongation<br>- necessity of reverse transcription to DNA<br>- (usually) reserved for thermostable polymerases |
| CST | - requires only low promiscuous activity for short extension<br>- no reverse transcription necessary<br>- thermal lysis deactivates enzymatic background | - indirect selection on tagging rather than high fidelity elongation<br>- (usually) reserved for thermostable polymerases |
| Phage display | - high throughput<br>- strong (covalent) genotype-phenotype linkage | - reserved for mesophilic polymerases<br>- necessity of co-display of polymerase and nucleic acid substrate |

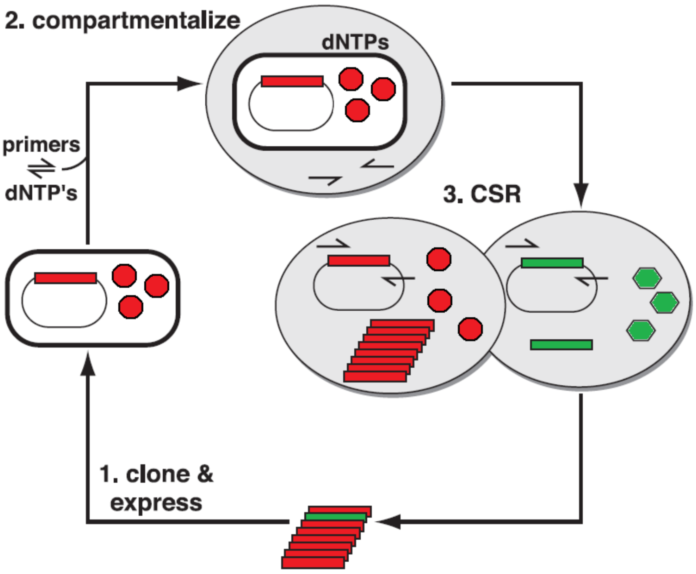


**Figure 1.7: Schematic overview of CSR.**

Each round of selection consists of a cloning and expression step, followed by mixing these cells in a watery phase together with nucleotides, primers, buffer and other necessary or wanted components depending on what is being selected for. This water phase is then emulsified in an oil phase. The red circles represent polymerases that are much more active than the green hexagon polymerases, which therefore are overrepresented in the population. These can then be re-cloned. Adapted from Ghadessy *et al.*, 2001 (79). Copyright 2001 Proceedings of the National Academy of Sciences USA.

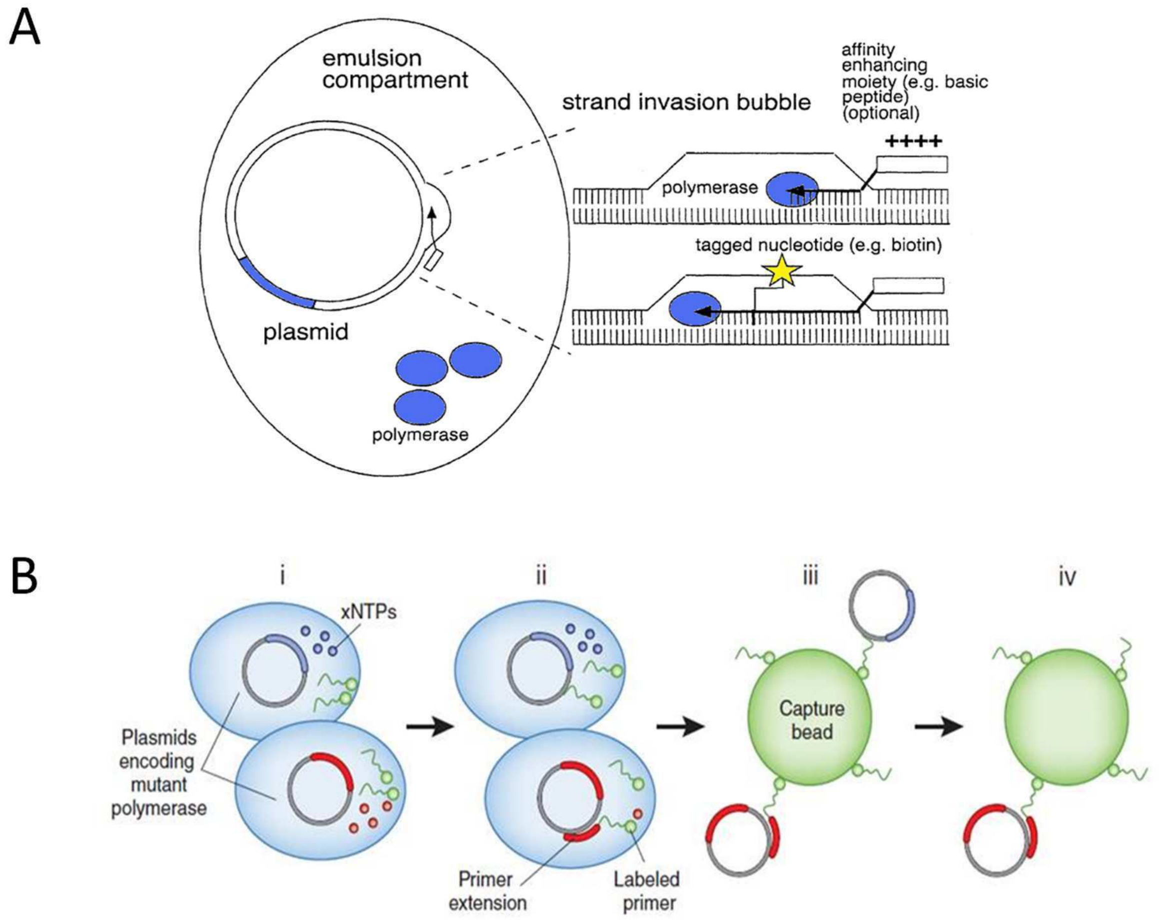


**Figure 1.8: Schematic overview of the CST reaction.**

A) As shown, the primer can either be tagged itself, or make use of tagged nucleotides. The reaction takes place in a water droplet in a water-in-oil emulsion, after cell lysis brings the plasmid template, expressed polymerase variant and nucleotides together. Adapted from Holliger *et al*., 2010 (91). B) Overview of CST and extraction of active variants using paramagnetic beads. The polymerases expressed from the red gene are able to tag the primer by elongating it, the ones from the blue gene are not. The emulsion is then broken and the tagged complexes are captured by the beads. Only plasmid-primer complexes that are tagged and bound together stably remain attached to the bead. These beads are then washed and extracted using a magnet, and the supernatant containing plasmids that code for less active variants is removed. Adapted from Steele & Gold, 2012 (92) with permission of the rights holder, Springer Nature.

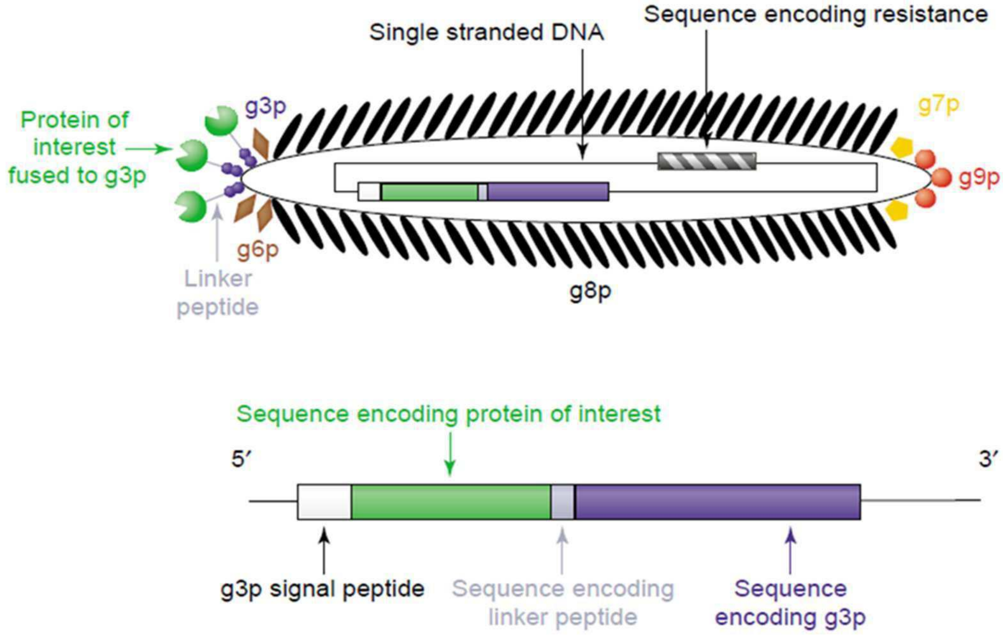

Sequence encoding resistance
Single stranded DNA
Protein of interest fused to g3p
g3p
g7p
g9p
g6p
Linker peptide
g8p
Sequence encoding protein of interest
5′
3′
g3p signal peptide
Sequence encoding linker peptide
Sequence encoding g3p

**Figure 1.9: Schematic of phage display.**

The protein of interest is fused between the signal peptide and the N-terminal end of the g3p gene with a linker peptide, causing it to be displayed on the outside of the phage particle. Figure adapted from Fernandez-Gacio *et al*, 2003 (85) with permission of the rights holder, Elsevier.

# 2

# Isolation of improved XNA polymerases

## 1. Introduction

This chapter summarizes the process of constructing, selecting and screening mutant libraries. The process of molecular enzyme evolution starts with choosing a suitable protein, of which variants are made that may have variable catalytic activities. Mutagenesis was performed to create focused mutant libraries, which are collections of polymerases that are randomly mutated in specific regions of the protein. These regions are chosen because mutations in these regions are predicted to have strong and more direct effects on polymerase activity. From this large pool of mutants, those with increased activity towards desired functions are then selected for using a technique designed specifically for the evolution of DNA polymerases: compartmentalized self-tagging (CST). This emulsion-based technique enriches the library for mutant polymerases based with an improved ability to incorporate 2'-deoxyxylonucleotides (dxNTs) and extend them with natural nucleotides. After selection, a number of clones are screened for this activity, and the most promising ones are sequenced and further characterized. This resulted in the discovery of a specific clone that consistently outperformed other mutants and the wildtype polymerase in its ability to extend DNA primers with up to two dxNTs followed by a larger number of natural nucleotides.

## 2. Materials and methods

### 2.1. General materials

Enzymes were bought from Thermo Scientific (Waltham, Massachusetts), except BsaI and commercial Vent exo$^-$ and Therminator exo$^-$ DNA polymerase which were bought from New England Biolabs (Ipswich, Massachusetts). Oligonucleotides were bought from Integrated DNA Technologies (Leuven, Belgium). 2'-Deoxyxylonucleotides (dxTTP and dxATP) were supplied by the Herdewijn lab of Medicinal Chemistry from the Rega Institute and were synthesized as described (19). Anhydrotetracyclin (AHT) and pASK-IBA2 vector were purchased from IBA Life Sciences (Göttingen, Germany). Electrocompetent E. cloni® cells were purchased from Lucigen (Middleton, Wisconsin). Biotinylated dCTP was purchased from Jena Bioscience (Jena, Germany). For all PCR and polymerase extension reactions a T300 Thermocycler (Biometra, Göttingen, Germany) was used. Kits for plasmid extraction, DNA purification and gel extraction kits were bought from Thermo Scientific. Protein purification PD MiniTrap G-25 columns were bought from GE Healthcare Life Sciences (Buckinghamshire, UK), Ni-NTA resin from Machery-Nagel (Düren, Germany). DNA sequencing was performed by LGC Genomics (Berlin, Germany). Images of protein structures are made with PyMol (Molecular Graphics System, Version 1.3 Schrödinger, LLC).

### 2.2. Agarose gel electrophoresis

DNA fragments were visualized using agarose gel electrophoresis. All gels were 1% (w/v) agarose for fragments in the 700-7000 bp range, and 2% for fragments around 500 bp or less. Gels were prepared by weighing the appropriate amount of agarose, mixing it with TAE buffer (20 mM Tris-HCl, 20 mM acetic acid and 1 mM EDTA), dissolved by heating in a microwave and cast in a mold. Gels were subjected to a constant potential of 120 V until the product was sufficiently separated. Visualization was performed by soaking the gel in an ethidium bromide (EtBr) bath for 15 min and illuminating with UV light. Gels containing DNA fragments to be extracted for cloning were stained with crystal violet instead, to avoid DNA damage and presence of EtBr in downstream reactions.

## 2.3. Library construction

### 2.3.1. Target sequence analysis

The crystal structure of DNA polymerase 9°N-7 (PDB: 4K8X) was used as a guiding structure. The only difference between the 9°N-7 and Therminator DNA polymerases are the A485L terminator mutation (93) and presence of an N-terminal hexahistidine tag (see below). The 4K8X crystal structure is a binary complex of a short dsDNA helix bound to the enzyme. Using the molecular visualization program PyMol, protein residues within a 6 Å shell around the nucleic acid were defined as potential mutation sites. These roughly 100 residues, which are dispersed over the linear sequence, were then narrowed down to six regions according to functional and structural constraints rather than mere distance to nucleic acid atoms. Additional residues were targeted in between them to create contiguous regions. The targeted regions include amino acids in the active site, sequences of amino acids that constitute regions that directly bind the nucleic acid backbone or interact with the minor or major groove, and amino acids in the vicinity of the incoming nucleotide substrate. The latter is not present in the 4K8X structure, but its position was inferred from a DNA polymerase RB69 structure (PDB: 3NCI) that is structurally homologous (only 24% sequence identity, 25% gaps). This was done by a superposition of both protein structures, creating a model of DNA polymerase 9°N-7 with a bound dNTP. This model was created by Dr. Mathy Froeyen using the Dali server (http://ekhidna2.biocenter.helsinki.fi/dali/).

### 2.3.2. Synthetic gene

The synthetic gene that codes for Therminator exo$^-$ includes the A485L terminator mutation, the D141A and E143A mutations that make it exonuclease-deficient (later referred to as exo$^-$), and is fused with an N-terminal hexahistidine tag. The coding sequence is 2394 bp long and codes for a 798 amino acid protein of 90 kDa.

The coding sequence was optimized for expression in *E. coli* by Dr. Wouter Delespaul using the Optimizer program (http://genomes.urv.es/OPTIMIZER/) and the online rare codon analysis tool by GenScript (http://www.genscript.com/cgi-bin/tools/rare_codon_analysis). The codon adaptation index (CAI) is an index of how well the used codons matches the proportional composition of the tRNA pool for highly expressed genes (94). The CAI of the original sequence, if it were to be expressed in *E. coli*, was 56%. Codon optimization brought this to 100%. The GC-content changed from 56.1% to 49.9%.

Using the online tool GeneDesign (http://54.235.254.95/gd/), 51 unique restriction sites were introduced without changing the amino acid sequence to aid in later sequence manipulation and analysis. This lowered the CAI to 83% and GC-content to 46.6%.

The synthetic gene for exonuclease-deficient DNA polymerase Therminator was ordered from GeneArt Life Technologies (Thermo Scientific, Waltham, Massachusetts). It was delivered on a pMK-RQ plasmid and was re-cloned into the pRSET-B vector using PCR amplification with primers with BamHI and NcoI recognition sites for seamless cloning into a digested pRSET-B plasmid (see Table 2.1, primers TermPol_FWD_BamHI and TermPol_REV_NcoI). Since this plasmid is relatively large the gene was later re-cloned in a pASK-IBA2 derivative plasmid using similar primers with the same restriction sites for BamHI and NcoI, from the pNIC-Bsa4 plasmid (see Table 2.1, primers pASK_FWD_BamHI, pASK_REV_NcoI, Therm_pNIC_FWD_NcoI and Therm_pNIC_REV_BamHI).

### *2.3.3. Mutagenesis*

Six regions were defined that are more or less functionally distinct, consisting of either one or two contiguous stretches of amino acids to be randomized. Partial randomization of these regions was achieved using degenerate primers (75), which are primers that are partially randomized at these specific regions. Within these randomized regions, each base has a 9% chance to be substituted by any of the three other bases, 3% chance each. This would result in roughly one substitution per five targeted residues, meaning that the average amount of mutations per clone also scales with the region length.

Primer sequences used in inverse PCR for library generation are given in Table 2.1. They have a 5' six nucleotide long non-complementary “handle”, a BsaI restriction site, followed by a sequence that binds a region of the Therminator gene. Each base represented by a lowercase letter has a 9% chance of substitution into any of the three other bases (3% chance each).

Three of the libraries are made with just one spiked primer, the others with two. Region N spans residues 383-390 and 400-411. Region F spans residues 486-501. Region A spans residues 537 to 547, however catalytic residues D540 and D542 were left unchanged. Region M spans residues 588 to 596 and 605 to 617. Region P spans residues 662 to 680. Region T spans residues 706 to 713 and 729 to 745. Libraries were generated by inverse PCR (iPCR). Each library was made in 200 µL reaction volumes with 4 ng template DNA, 0.5 µM of each primer, 0.2 µM dNTPs and 4µL Phire Hot Start II Polymerase.

**Table 2.1: List of used oligonucleotides in cloning, library construction and polymerase activity assays.**

| **Oligonucleotide** | **Sequence (5' to 3')** |
|---|---|
| TermPol_FWD_BamHI | ATC TAG GAT CCG ATG ATC CTG GAC ACC GAC TAC ATC ACC |
| TermPol_REV_NcoI | CTT ATC CCA TGG TTA TTT TTT ACC TTT AAC TTT CAG C |
| pASK_FWD_BamHI | ATA TAT ATG GAT CCG ACC TGT GAA GTG AAA AAT GGC GC |
| pASK_REV_NcoI | ATT ATT ATC CAT GGG TAT ATC TCC TTC TTA AAG TTA AAC |
| Therm_pNIC_FWD_NcoI | TAT TAT TAC CAT GGC TCA CCA TCA TCA TCA TCA TTC TTC TGG |
| Therm_pNIC_REV_BamHI | TTA AAT AAG GAT CCT TAT TAT TTT TTA CCT TTA AC |
| Therm_N_FWD | CATTAGGGTCTCATGTGGGACAACatcgtttacctgGACttccgttctctgtacccgtctATCATCATCACC |
| Therm_N_REV | CATTAGGGTCTCACACAGACCACGTTCCGGTTCtttaacgtaaccgccggcgtaaccACCACGACGG |
| Therm_F_FWD | CATTAGGGTCTCAGACTACCGTCAGcgtTTAattaaaatcctggctaactctttctacggttactacggttacgctaaaGCTCGTTGG |
| Therm_F_REV | CATTAGGGTCTCAAGTCCAGCAGTTTTTTTTCCAGCGGGTCAAC |
| Therm_A_FWD | CATTAGGGTCTCAAGTTctgtacgctGACaccGACggtctgcacgctaccATCCC |
| Therm_A_REV | CATTAGGGTCTCAAACTTTGAAACCGAATTTCTCCTCCAGTTCACGGATA |
| Therm_M_FWD | CATTAGGGTCTCAATCACCacccgtggtctggaaatcgttcgtcgtgactggtctgaaATCGCTAAAG |
| Therm_M_REV | CATTAGGGTCTCATGATTTTACCTTCTTCGTCGATaacagcgtatttttttttggtaacgaaGAAACC |
| Therm_P_FWD | CATTAGGGTCTCACTAGTTatccacgaacagatcacccgtgacctgcgtgactacaaagctaccggtccgcacgtcGCTGTTGCT |
| Therm_P_REV | CATTAGGGTCTCACTAGTTTTTCCGGCGGAACTTCGTATTTAGACAG |
| Therm_T_FWD | CATTAGGGTCTCACGTTACGACgctgaatactacatcgaaaaccaggttctgcctgcagttgaacgtatcctgAAAG |
| Therm_T_REV | CATTAGGGTCTCAAACGGTGTTTGGTCGGGTCGAATTCGTCGGCCGGGATGGCccggtcaccgatacgaccagaaccTTTCAGAAC |
| ELISA_primer | BiotinTEG-GAA CAG ATC ACC CGT GAC CTG |
| ELISA_probe | DIG-GTG CGG ACC GGT AGC |
| ELISA_linker | GCT ACC GGT CCG CAC CGT TCT GCG TTC CTG TCT GTT C |
| ELISA_oligo_1T | GTT CTG CGT TCC TGT CTG TTC GCC CTT **A**TT CCC TTC AGG TCA CGG GTG ATC TGT TCG TGG |
| ELISA_oligo_2T | GTT CTG CGT TCC TGT CTG TTC GCC CT**A** **A**TT CCC TTC AGG TCA CGG GTG ATC TGT TCG TGG |
| ELISA_oligo_3T | GTT CTG CGT TCC TGT CTG TTC GCC C**AA** **A**TT CCC TTC AGG TCA CGG GTG ATC TGT TCG TGG |
| ELISA_oligo_2T1T | GTT CTG CGT TCC TGT CTG TTC GCC CTT **A**TT C**AA** TTC AGG TCA CGG GTG ATC TGT TCG TGG |
| ELISA_oligo_2T2T | GTT CTG CGT TCC TGT CTG TTC GCC CT**A** **A**TT C**AA** TTC AGG TCA CGG GTG ATC TGT TCG TGG |
| ELISA_oligo_1T1T | GTT CTG CGT TCC TGT CTG TTC GCC CTT **A**TT **A**CC TTC AGG TCA CGG GTG ATC TGT TCG TGG |

PCR (30 cycles) was performed in two 100µL aliquots at annealing temperatures 5°C below the $T_m$ of the spiked primer with the lowest $T_m$. Each PCR reaction mixture was then purified with a PCR purification kit, which yielded around 18 µg DNA per library. The eluted DNA was then digested using BsaI (40U) and DpnI (10U) for 4h at 37°C with an extra 1U of BsaI added after the first 2h. The DNA was then purified again using the same PCR purification kit, and 5 µg DNA was ligated with T4 DNA ligase (50U) overnight at room temperature in a total reaction volume of 5 mL. The reaction products were then concentrated using a Speedvac concentrator and purified by PCR purification, where ligation yielded around 3.9 µg. Sample purity (as measured by $A_{260}/A_{280}$ using Nanodrop spectrophotometer) remained high (close to 1.8) for most samples, except for the purified ligation products, where the $A_{260}/A_{280}$ ratio was often larger than 2.5. This purity was sufficient for successful transformation.

## 2.4. Selection and screening

### *2.4.1. Compartmentalized self-tagging (CST)*

Electrocompetent E. cloni® cells were transformed with plasmid libraries and plated on LB agar with ampicillin (Amp). Colonies grown overnight were harvested by adding LB medium and gently scraping them from the plates. The suspension (~30 mL, ~30 $OD_{600}$) was mixed with glycerol (60%) in a 3:1 ratio, flash-frozen in 100 µL aliquots and stored at -80 °C. One aliquot was used to inoculate a 4 mL starting culture grown overnight (37°C, 200 rpm). An aliquot of 50 µL overnight culture was inoculated in 4 mL LB + Amp medium and grown at 37°C, 200 rpm to 0.7 $OD_{600}$, induced with AHT, and further grown at 25 °C overnight. Cultures were diluted to 1 $OD_{600}$, 100 µL cell suspension was centrifuged at 12000 *g* and supernatant was removed. These cells were added to the aqueous phase before emulsifying.

Selection was performed using CST as described by Pinheiro *et al*. (84), but in the absence of misincorporation-promoting $Mn^{2+}$ ions, and labelling by incorporation of biotinylated dCTP during extension. An oil phase (mineral oil with 4.5% Span 80, 0.4% Tween 80 and 0.05% Triton X-100) and aqueous phase (1x Thermopol buffer (New England Biolabs), $MgSO_4$ (1 mM), biotinylated dCTP (20 µM), equimolar dATP/dGTP mix (200 µM), dxTTP (200 µM), primer (0.38 µM), BSA (1 mg/mL), DTT (1 mM), glycerol (10% v/v)) were emulsified in a 4:1 ratio by adding the aqueous phase dropwise to the oil phase while stirring with a small magnet, according to the protocol in reference (95).

The elongation reactions were run for 10 min at 95 °C, 15 min at 58 °C and 15 min at 72 °C. The primer for the first round of CST required minimal incorporation of GAAAAG**dxT**GC-biotin for capture. In the second round the selective pressure was slightly increased by requiring incorporation of AAAAAGGGAA**dxT**AAGGGC-biotin. While promiscuous activity of the wild type has been shown to allow for up to two consecutive dxNT incorporations, the resulting stalling could prove to be a high difficulty, and therefore the first selection round was kept relatively low, with selection for DNA production, followed by a single dxNT incorporation, and short extension with DNA. The next round increased the selective pressure by increasing the extension length past the incorporated dxNT, thereby selecting for increased translocation rather than increasing the amount of dxNT.

After CST, the emulsion was broken by centrifugation at 13000 *g* for 5 min and two water-saturated diethyl ether extractions. Primer-plasmid complexes were purified using spin columns (GE Healthcare S400 microspin columns) and incubated with magnetic beads (streptavidin magnetic beads, 4 mg/mL, 1 µm diameter, New England Biolabs) for at least 1h at room temperature on a rolling platform. The beads were washed twice with a high salt binding buffer (0.5 M NaCl, 20 mM Tris-HCl, 1 mM EDTA, pH 7.5), once with a lower salt buffer (0.15 M NaCl, 20 mM Tris-HCl, 1 mM EDTA, pH 7.5), and eluted in elution buffer (25 µL of 10 mM Tris-HCl, 1 mM EDTA, pH 7.5). Re-cloning was done by PCR amplification of the targeted half of the Therminator gene using 2 µL of the eluted fraction as template, and PCR amplification of the non-variable half of Therminator plus the rest of the plasmid backbone from the original wild type, followed by restriction digestion and ligation of both fragments to reconstitute the original plasmid.

### *2.4.2. Polymerase activity assay*

After selection, libraries enriched in mutants with improved incorporation activity were screened using a previously described polymerase activity assay (17). Colonies from the re-cloned library were inoculated in 400 µL selective medium in a 96-well plate with deep, rounded wells and were induced to express the polymerase with anhydrotetracyclin (AHT) at 25 °C and 200 rpm overnight. Lysates were prepared for each mutant by heating 20 min at 80 °C, centrifuging to remove the cellular pellet, and mixing the supernatant in 1X Thermopol buffer (20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton X-100 at pH 8.8). The elongation reaction uses the cell lysate as a source of polymerase, and a biotinylated primer (ELISA_primer in Table 2.1) was elongated with a template (ELISA_oligos in Table 2.1, where bases templating for dxT are shown in bold) that requires incorporation of the sequence AAGGGAA**dxT**AAGGGC using dATP, dGTP, dCTP and dxTTP, followed by an non-T containing sequence used for detection by hybridization with a digoxigenin-containing DNA probe (ELISA_probe in Table 2.1) through an intermediary linker (ELISA_linker in Table 2.1). Polymerase mutants that succeed in incorporation of dxTTP opposite of the templating adenine and elongation beyond the resulting distortion produce elongated primers that can bind this DNA probe (see Figure 2.1). After capture of the primers to a streptavidin-coated 96-well plate, duplex denaturation and washing with PBST (125 mM NaCl, 16.6 mM $Na_2HPO_4$, 8.43 mM $NaH_2PO_4$, 0.2% (v/v) Tween 20, pH 7.2), the samples are incubated with an anti-digoxigenin antibody fused to horseradish peroxidase (HRP) (dilution 1:10,000). After washing away non-specifically bound probes and antibodies with PBST, the chromogenic substrate 3,3’,5,5’-tetramethylbenzidine (TMB, 100 µL) is added which turns blue in the presence of HRP. The $OD_{450}$ nm is an indication of the amount of fully elongated primers, and stronger signals correlate with more active polymerase containing lysates.

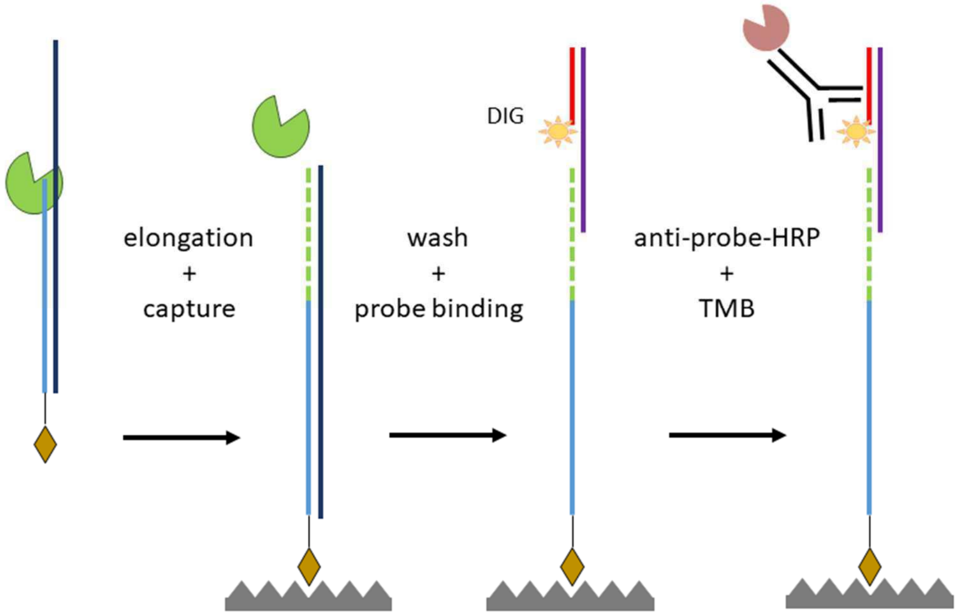


**Figure 2.1: Overview of XNA polymerase activity assay.**

A biotinylated primer (light blue) is hybridized with a template (dark blue), and the lysate containing expressed polymerase is added, together with buffer and the nucleotide substrates dATP, dGTP, dCTP and dxTTP. When full extension occurs, a probe sequence with a DIG-probe (red) can bind through an intermediary linker (purple) when the primer is elongated (green dashed line). The single-stranded part of the extended primer will contain the dxT nucleotides and needs to be produced and fully extended for the probe to bind. Visualization occurs by the horse radish peroxidase (HRP) fused to an anti-DIG antibody that acts on the chromogenic substrate TMB, where the rate of colouring depends on the amount of successfully extended primers.

### *2.4.3. Sodiumdodecylsulphate polyacrylamide gel electrophoresis (SDS-PAGE)*

Cell cultures that expressed DNA polymerases were subjected to SDS-PAGE to analyse the amount of expressed protein, and in case of protein purification it was used to show the presence of the DNA polymerase in different fractions and to estimate the protein concentration in the purified eluted product.

Samples for SDS were diluted to account for cell density and running gels were composed of 10% acrylamide. Sample and gel preparation and visualization were performed according to the Sambrook laboratory manual protocol.

## 3. Results

### 3.1. Construction of large, high quality targeted libraries

Focused libraries were designed based on the crystal structures of the binary complex of the precursor protein 9°N-7. The position of the incoming dNTP was predicted based on the ternary complex of a crystal structure of polymerase RB69. This means that rather than mutagenizing the whole gene, like when using error-prone PCR amplification, mutations can be restricted to a specific region of interest. This also increases the total mutational load, which means that a larger part of sequence space can be probed, albeit in only a part of the enzyme. The underlying rationale for this is that it is more likely to find mutations that change specificity and catalytic activity when they are in close proximity to the DNA, as was the case for most of the mutations conferring XNA polymerase activity for six different XNAs that were selected for using CST by Pinheiro et al. (17).

The synthetic gene expressed well from the pNIC28-Bsa4, pRSET-B and pASK-IBA2 plasmids. However, given the small size of the latter two plasmids (which increases transformation efficiency), and the most consistent protein expression levels from the pASK-IBA2 vector, the latter vector was used for all selection experiments after which screening and mutant analysis took place.

Six libraries were constructed with iPCR, each of a different region of the polymerase in close contact with the bound nucleic acids (see Figure 2.2). Based on dilution series of E. cloni® cells transformed with these libraries, library sizes were estimated. They ranged between 3x$10^4$ and 1.2x$10^7$ independent clones.

Region N contains part of the nucleotide-binding pocket, as well as a less structured part that lies close to the template backbone (yellow in Figure 2.2). Mutant phenotypes are expected to influence nucleotide and template binding. The corresponding library contained around 9.5x$10^6$ clones.

Region F is located at the basal part of one of the finger domain alpha helices (orange in Figure 2.2), and mutants are hypothesized to mainly influence nucleotide binding and recognition, as well as dynamics of the finger domain as a whole. The F library size was approximately 1.2x$10^7$.

Region A contains most of the active site residues (purple in Figure 2.2), and mutations in this region will likely affect catalysis directly. The A library size was approximately 8x10$^{5}$.

Region M mostly interacts with the minor groove of the duplex DNA, as well as part of the active site (red in Figure 2.2). The M library size was approximately 1.2x10$^{7}$.

Region P contacts a large stretch of the primer backbone in the thumb domain (magenta in Figure 2.2). The P library size was approximately 3.7x10$^{6}$.

Region T is also a part of the thumb domain, but is situated more towards the C-terminus, and mostly interacts with the upstream template backbone (pink in Figure 2.2). Mutants in these last two regions will likely affect nucleic acid binding strength, specificity and translocation. The T library size was approximately 3x10$^{4}$.

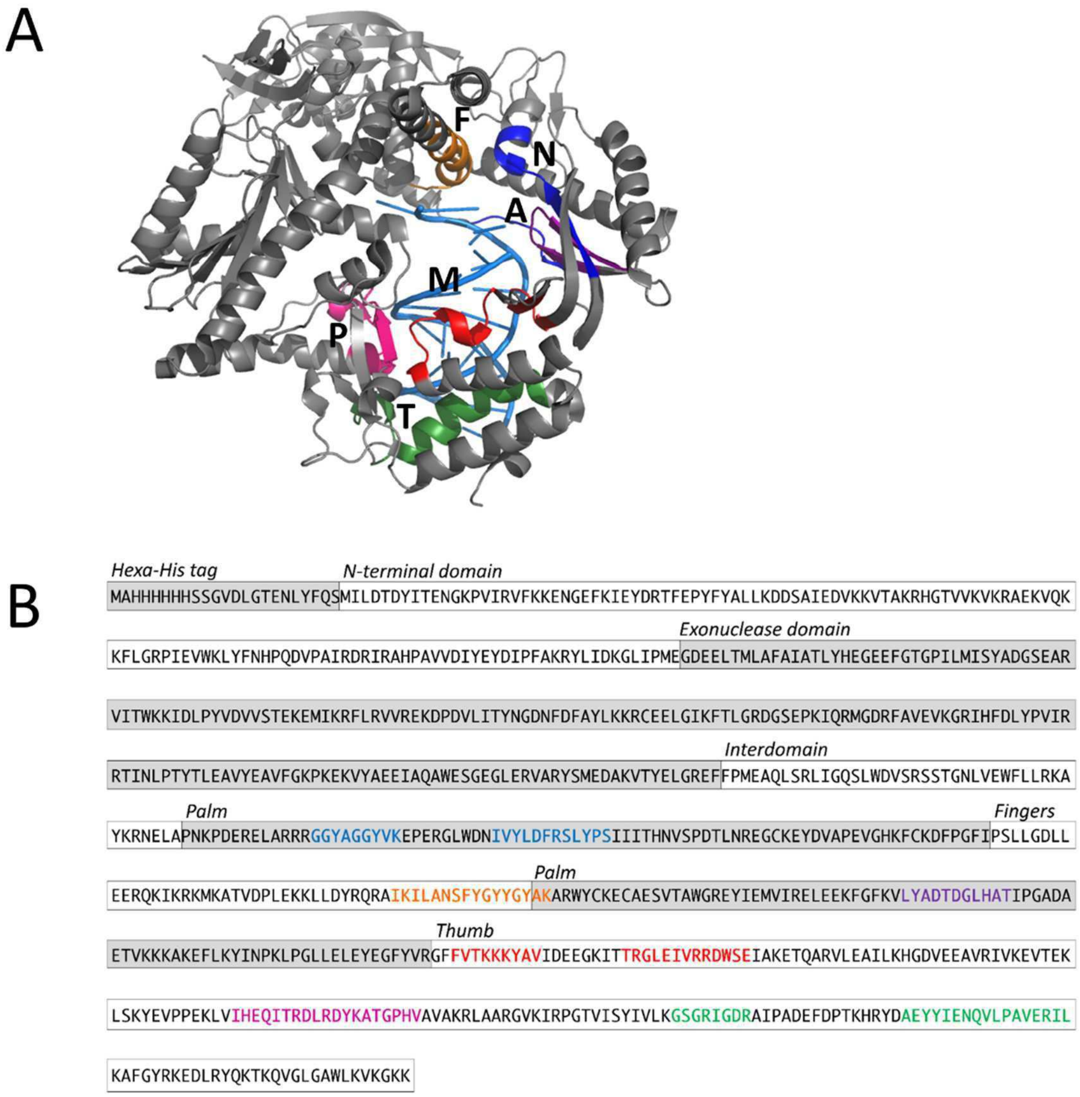

A
F
N
A
M
P
T
B
Hexa-His tag
N-terminal domain
MAHHHHHHSSGVDLGTENLYFQSMILDTDYITENGKPVIRVFKKENGEFKIEYDRTFEPYFYALLKDDSAIEDVKKVTAKRHGTVVKVKRAEKVQK
Exonuclease domain
KFLGRPIEVWKLYFNHPQDVPAIRDRIRAHPAVVDIYEYDIPFAKRYLIDKGLIPMEGDEELTMLAFAIATLYHEGEEFGTGPILMISYADGSEAR
VITWKKIDLPYVDVVSTEKEMIKRFLRVVREKDPDVLITYNGDNFDFAYLKKRCEELGIKFTLGRDGSEPKIQRMGDRFAVEVKGRIHFDLYPVIR
Interdomain
RTINLPTYTLEAVYEAVFGKPKEKVYAEEIAQAWESGEGLERVARYSMEDAKVTYELGREFFPMEAQLSRLIGQSLWDVSRSSTGNLVEWFLLRKA
Palm
Fingers
YKRNELAPNKPDERELARRRGGYAGGYVKEPERGLWDNIVYLDFRSLYPSIIITHNVSPDTLNREGCKEYDVAPEVGHKFCKDFPGFIPSLLGDLL
Palm
EERQKIKRKMKATVDPLEKKLLDYRQRAIKILANSFYGYYGYAKARWYCKECAESVTAWGREYIEMVIRELEEKFGFKVLYADTDGLHATIPGADA
Thumb
ETVKKKAKEFLKYINPKLPGLLELEYEGFYVRGFFVTKKKYAVIDEEGKITTRGLEIVRRDWSEIAKETQARVLEAILKHGDVEEAVRIVKEVTEK
LSKYEVPPEKLVIHEQITRDLRDYKATGPHVAVAKRLAARGVKIRPGTVISYIVLKGSGRIGDRAIPADEFDPTKHRYDAEYYIENQVLPAVERIL
KAFGYRKEDLRYQKTKQVGLGAWLKVKGKK

**Figure 2.2: Targeted regions in Therminator DNA polymerase.**

A) Crystal structure of Therminator DNAP showing the targeted regions. Non-variable parts of the protein are shown in grey. The template dsDNA is shown in light blue. Region N is shown in dark blue, region F in orange, region A in purple, region M in red, region P in magenta and region T in green. B) Linear sequence of Therminator DNAP, including the N-terminal His tag and TEV protease cleavage site (ENLYFQS) fusion. Targeted regions have the same colours as in the crystal structure.

To test whether the produced mutants had the expected number of substitutions, up to 16 randomly picked clones were sequenced per library, and the amount of nucleotide and amino acid substitutions was compared to the predicted averages. Most libraries had average amounts of mutations close to these expected values. Figure 2.3 shows such a histogram of clones from library F, where the expected average amount of nucleic acid mutations across the region was 4.6, resulting in 3.6 amino acid mutations (since some of the nucleic acid mutations are silent). Given that these histograms tended to largely cluster around expected values, it was concluded that the primer composition was made to expectations and the libraries were of expected quality.

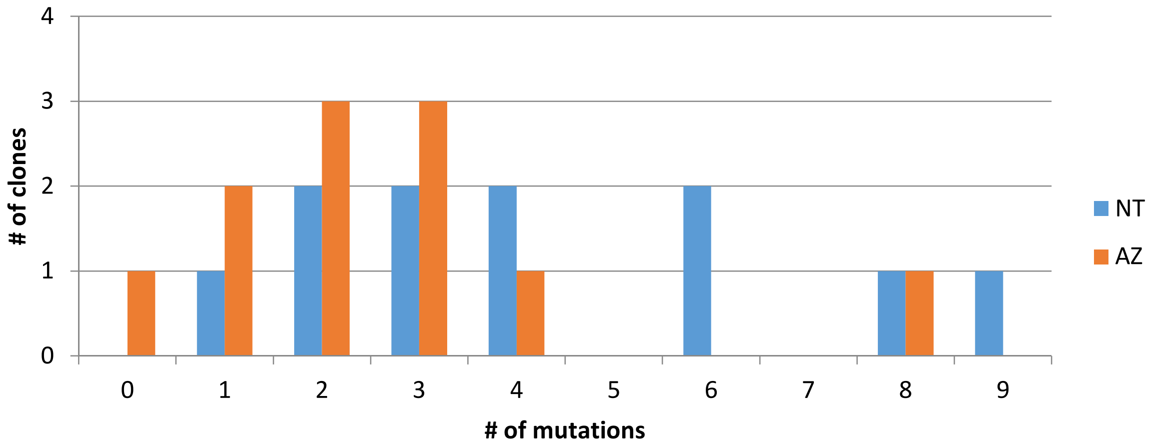


**Figure 2.3: Mutation histogram of 11 clones from library F.**

The expected frequency value for the F library is calculated to be 4.6 for nucleotide substitutions (blue) and 3.6 for amino acid substitutions (orange).

### 3.2. Enrichment of mutants with higher activity by CST

The effectiveness of the CST procedure was evaluated using mock libraries in parallel with mutant libraries. Such mock libraries were composed of various proportions of “wild type” DNA polymerase Therminator (WT) and a knockout mutant (KO). Tested ratios of WT:KO were 1:1, 1:10 and 1:100. The knockout mutant was created by Dr. Yves Peeters by substitution of the two catalytic aspartates to alanine, which introduces a new SacII restriction site at that position. After one and two rounds of CST, clones were randomly sampled with colony PCR (see Figure 2.4 for the 1:100 ratio mock library after two rounds). If there was no selection, the ratio would be very similar to the initial WT:KO ratio. Slight deviations could be simply due to genetic drift or sampling error. Large changes should be due to effective selection, and the more biased the ratio, the stronger the selective pressure. In Figure 2.4, 20 clones showed the wild type pattern, versus nine knockouts and one anomalous clone that is most likely a deletion mutant. Two CST rounds therefore increased the frequency of active variants from 1% to 67% when in competition with inactive variants. After the first round, the wild type frequency was 44% (data not shown). This was taken as a proof of concept that active polymerase genes can be significantly enriched even after few selection rounds.

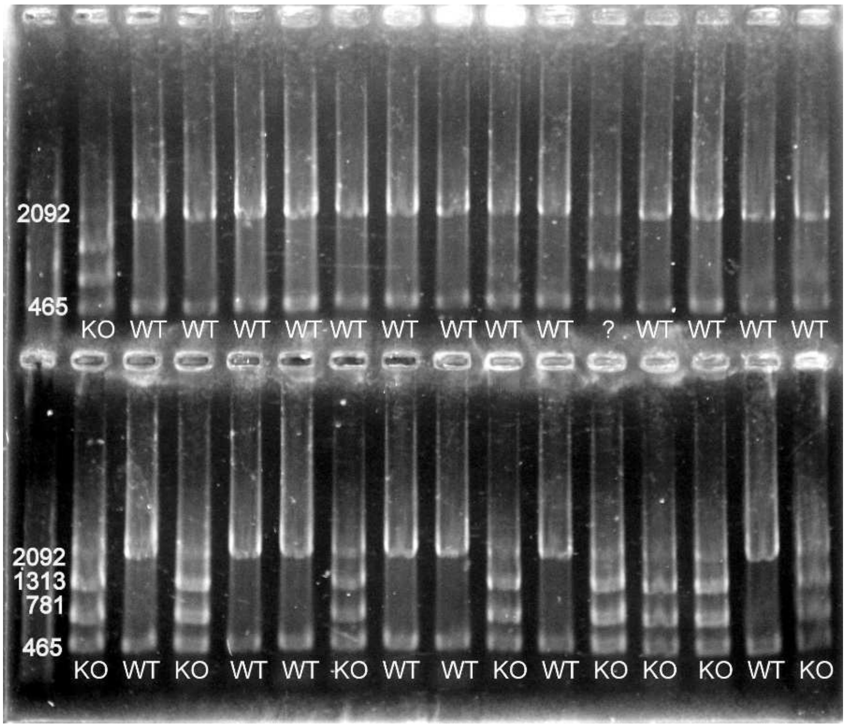

2092
465
KO WT WT WT WT WT WT WT WT WT ? WT WT WT WT
2092
1313
781
465
KO WT KO WT WT KO WT WT KO WT KO KO KO WT KO

**Figure 2.4: Electrophoretic analysis of sampled clones after two rounds of CST in a 1:100 mock library of wild type and knockout Therminator.**

Wild type clones are marked with WT, knockout with KO. When the Therminator gene is amplified it forms a 2557 bp fragment. Wildtype fragments have one SacII site, and digestion will result in a 2092 bp and 465 bp fragment. Knockout mutants with an extra restriction site result in fragments of 465 bp, 781 bp and 1313 bp, thus easily distinguishing it from the wild type. A question mark indicates an anomalous clone.

To monitor library quality during CST, the successive plasmid libraries were analysed and compared with regards to plasmid size, which was a good indicator of correct re-cloning. Plasmids libraries from different selection rounds were compared by electrophoretic analysis of both plasmids that are either undigested or digested with BamHI and NcoI. Figure 2.5 gives an example of libraries A and M after zero to four CST rounds. This revealed that some libraries, after two or more rounds of selection, contain increasing numbers of deletion mutants, to the point where almost no plasmids of normal lengths were present. These are likely to be genetic parasites, such as shorter plasmids that clone more easily with higher transformation efficiency, even if they code for polymerases that are less or not active, or deletion mutants that were not present during selection but are artefacts formed during re-cloning. Similar problems were seen for library T after three CST rounds (data not shown). Because of this, subsequent rounds of selection were always followed by overall library analysis, using plasmid length distributions as a proxy for overall library quality.

Mutant screening was therefore only done for libraries that still exhibited large fractions of intact plasmids, to avoid inactive deletion mutants being present in most of the screening.

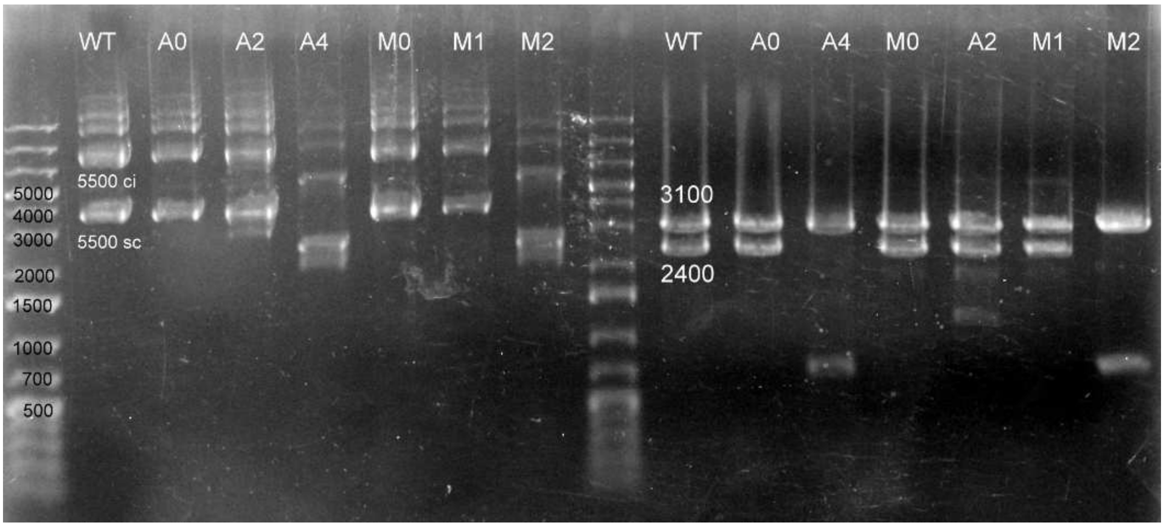


**Figure 2.5: Gel electrophoresis of native and digested plasmid libraries of A and M libraries after various rounds of CST.**

A0, A2 and A4 are of library A after 0, 2 and 4 CST rounds respectively. M0 to M2 are of library M from 0 to 2 rounds of CST. The seven samples on the left are undigested plasmids, those on the right are digested with BamHI and NcoI. Lanes marked WT are wildtype controls. Complete plasmids of 5500 bp should produce fragments of 3100 and 2400 bp. Shorter fragments indicate deletion mutants, which started becoming extremely prevalent in library A after 4 rounds, and already after two rounds in library M. The ladder at the left and centre of the gel is the GeneRuler 1 kb plus ladder.

### 3.3. Screening of enriched libraries using ELISA

High quality libraries that underwent one or two rounds of CST were analysed with an ELISA based DNA polymerase activity assay. Results are shown for library M after one round of CST, and library A after two in Figure 2.6A and 2.6B. Initial screening was performed with a template requiring a single dxTTP incorporation after seven natural nucleotides, which is then further extended with 27 dNTs. The sequence that needs to be made in the screening procedure is very similar to the sequence with which the CST primer was extended. Because variation in the output signal was relatively large and because of the risk of false positives, the best performing clones were tested again in threefold from different colonies of each clone and grown and induced separately (see Figure 2.6C). More demanding templates were also tested, but only the two least demanding ones showed a signal. Reactions requiring the incorporation of one dxTTP gave a much larger signal than when two dxTTP were needed. The clone that consistently outperformed all other mutants and wild type was clone E2.

Negative controls without added nucleotides, or any of the necessary oligonucleotides, or active polymerase, all consistently showed no signals above background. However, negative controls where one nucleotide out of four was omitted (either dTTP or dxTTP) gave very high signals, often higher than reactions where dATP, dCTP, dGTP and dxTTP were all present. This is seen as evidence for the ability of DNA polymerases, both promiscuous and exo$^-$ variants as well as high fidelity ones with repair activity, to bypass template sequences for which the matching nucleotide is missing, presumably by extending downstream of a misincorporation. Negative controls in other related experiments have shown that it takes around three or four consecutive template bases lacking their cognate partner nucleotides to completely stop a polymerase from advancing over such a template. (Dr. Yves Peeters, personal communication) The presence of residual cellular dNTPs could also act as a potential low concentration source of dTTP.

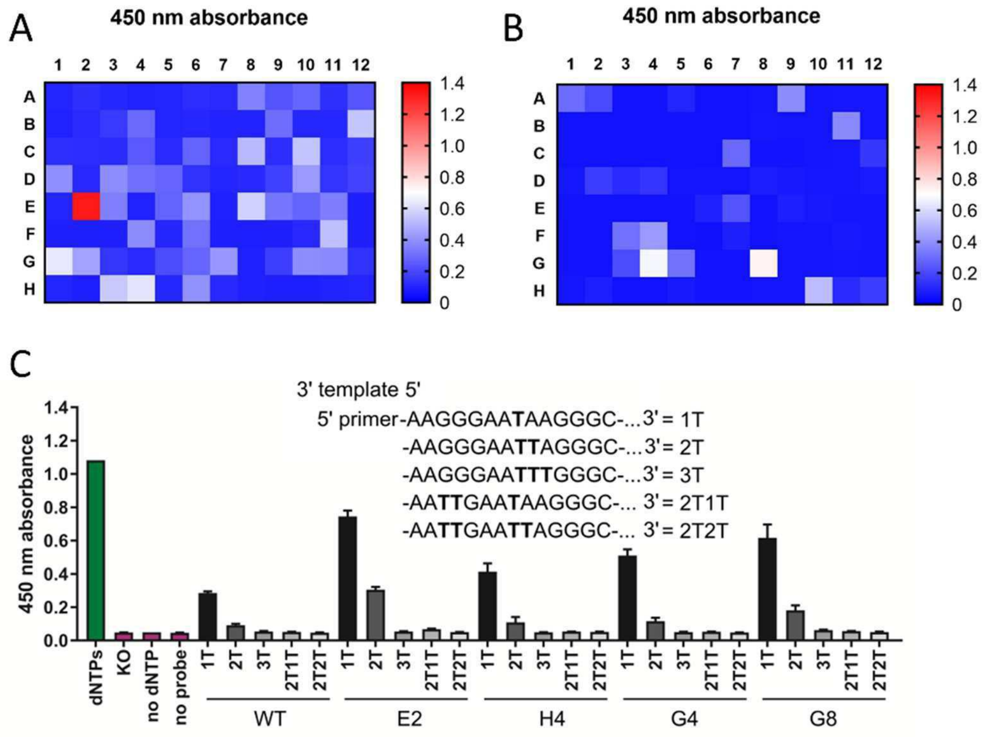


**Figure 2.6: ELISA-based DNA polymerase activity assays of two libraries after CST and some of their best performing clones.**

A) Activities in 96-well plate of 92 randomly picked clones from library A after two CST rounds. Two wild type (wells A2 and B2) and two knockout polymerase clones (wells A1 and B1) were used as positive and negative controls. The intended wild type control did not give a signal above background, but wells A8, D3, D10 and G7 are clones carrying only silent mutations, effectively serving as wild type controls on this plate. The majority of the sampled clones are either inactive or less active than the wild type. Only three clones (E2, G1 and H4) show a drastic increase in activity compared to the wild type. B) A similar 96-well plate screening but with library M clones after one CST round. Eight clones (including G4 and G8) have higher signal than the wild type. Wells A1 and A2 are wild type controls, B1 and B2 are knockout controls. C) ELISA of the four most promising clones from these two screens with five different templates. The sequences that should be produced by the polymerase are shown, where A, G and C are dNTs and T is a dxT. The more dxNTP needs to be used, the lower the efficiency resulting in a decreased ELISA signal. Only primers requiring one or two consecutive dxTs could be fully elongated. The template used in A) and B) corresponds to template 1T. Error bars represent the standard error of mean of reactions that were run in triplicate using the same lysate.

### 3.4. Mutant polymerases with increased activity

Several clones present in the 96-well plate screens were sequenced. These include all the clones with drastically higher activity, most of the ones with slightly higher activities, a few that perform comparable to the wild type, and a small number of less or inactive clones. These sequencing results are summarized in Table 2.2. Library A clones underwent two rounds of CST, library M clones underwent one. The variable region of library A spans residues 537 to 547, library M spans residues 588 to 596 and 605 to 617.

While it is not possible to derive a strong structure-function relationship from these data, there are some overall patterns. Mutations tend to fall in two classes (with exceptions). Based on the crystal structure, many overrepresented positions that cause more drastic phenotypic differences are those that either directly contact the nucleic acid and other important amino acid residues, or that point away from the nucleic acid and are instead positioned in between two alpha helices. The terminator mutation A485L that conferred the increased promiscuous activity for xenonucleotides to 9°N-7 DNA polymerase would fall in this latter class.

**Table 2.2: Amino acid mutations in polymerases that were found in ELISA screening.**

| Clone ID | library | Mutations (bold) in targeted region | Activity compared to WT |
|---|---|---|---|
| E2 | A | $^{537}$LYADTDGL**R**AT | increased |
| H4 | A | L**F**ADTDGLHAT | increased |
| G1 | A | LYAD**N**DGLH**V**T | increased |
| E6 | A | LYADTDGL**L**AT | similar |
| C8 | A | L**S**ADTDGL**Y**A**A** | similar |
| F11 | A | LYADTDGL**Q**AT | similar |
| B12 | A | LYAD**I**DGLHAT | similar |
| H1 | A | LYAD**P**DGLHAT | inactive |
| G8 | M | $^{588}$FVTKKKYAV-$^{605}$TRGLEIV**H**RDWS**R** | increased |
| G4 | M | FVTKK**N**YA**A**- TRGLE**M**VRRDW**P**E | increased |
| H10 | M | **I**VT**T**KK**L**AV- TRGLEIVRRDWS**D** | increased |
| A9 | M | FVTKKKYAV- TRGLE**M**VRRDWS**V** | increased |
| B11 | M | **Y**VT**R**KKYAV- TRGLEIVRRD**S**SE | increased |
| G5 | M | FV**N**KKKYAV- TRGLE**V**VRRDWSE | increased |
| F3 | M | **L**VTKK**T**YA**E**- TRGLEIVRRDWS**V** | increased |
| iH4 | M | FVTKKK**H**AV- TRGLEIV**S**RDWSE | increased |
| C7 | M | F**I**TKKKYAV- TRG**M**EIVRRDWSE | similar |
| H11 | M | FVTKKKYAV- TR**C**LEIVRRDWSE | decreased |
| D9 | M | F**I**TKKK**D**AV- **S**R**D**LEIV**C**RDWSE | inactive |
| F10 | M | FVT**T**KKYA**F**- **S**RGLEIVRR**G**WSE | inactive |

## 4. Discussion

### 4.1. CST modifications

One of the modifications to the published CST protocol (84) was the omission of $Mn^{2+}$ from the aqueous phase, leaving only $Mg^{2+}$ at a relevant concentration. $Mn^{2+}$ was added in the original protocols because it has been shown to stabilize non-canonical nucleotides in the active site to undergo incorporation more easily (96, 97). This could help a polymerase with low promiscuous activity to incorporate it more efficiently. $Mn^{2+}$ was left out of the reaction mixture in later CST rounds in this work for two reasons. The first is that any evolving system that is aided by the environment can become dependent on that help. In the absence of $Mn^{2+}$, the polymerases are required to rely solely on mutations that change their function, rather than an added metal ion.  While in the long term an XNA polymerase that uses its own dedicated catalytic metal ions compared to the natural $Mg^{2+}$ could become more functionally orthogonal, it is likely that the intermediate evolutionary stages that are still transitioning between DNA and XNA as substrates are hindered more than helped by addition of $Mn^{2+}$. The second reason is that there is no direct evidence that the presence of $Mn^{2+}$ helps incorporation of dxNTPs, and could also aid misincorporation, which in turn would aid in bypassing through misincorporation, “cheating” selection, resulting in more false positives. This assumption was later tested with a PAGE experiment where the wildtype and E2 clone were tested in presence of $Mn^{2+}$ (see Figure 2.7). No increase in dxTTP incorporation was seen for either polymerase in the presence of 1 mM $MnCl_2$ and extension was completely absent at 10 mM. In the absence of dxTTP however, misincorporation was drastically increased and the reaction looks inhibited in either case at 10 mM. This justifies the exclusion of a mutagenic catalyst from polymerase selection experiments in general.

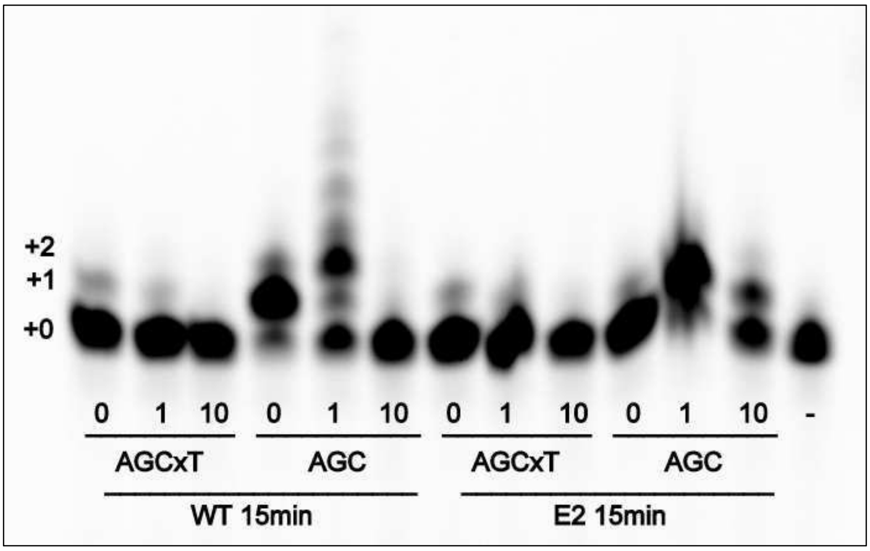


**Figure 2.7: PAGE of wild type and E2 polymerases and the effect of Mn$^{2+}$.**

0, 1 and 10 are concentrations of $MnCl_2$ in mM. AGC are natural nucleotides present in the reaction, xT means dxTTP is present. The used template is template “2D” (see Figure 3.5A). A negative control (unextended fluorescent primer) is labelled with a minus sign. Reactions were quenched and sampled after 15 min.

The initial selective pressure during CST was set relatively low. It required the incorporation of six dNTs, followed by a single dxNT, and minimal extension by only two dNTs. Since the known promiscuous activity of the wild type was very low (up to two consecutive incorporations), the rationale behind choosing a CST primer which requires only one dxNT incorporation was to improve both incorporation and elongation. A polymerase that has an improved incorporation rate for adding a single xenonucleotide will presumably also be improved for all the following ones, but simultaneously selecting on elongation could select against the observed stalling, even if it is extension with dNTs. Mutants that are less prone to stalling and that tolerate translocation of the distorted primer backbone are assumed to be a more useful evolutionary intermediate compared to a polymerase that can very efficiently incorporate a low number of dxNTs and then stalls completely.

### 4.2. Library quality

One of the important aspects of molecular evolution experiments is having a good starting point. This includes choosing a starting protein that has a binding pocket or catalytic activity that already resembles the target substrate or reaction, since it is more likely that less mutations are needed to increase affinity or activity. While DNA polymerase Therminator had already been shown to have a promiscuous activity towards incorporation of dxNTPs (20), this activity is very low compared to other xenonucleotides. The starting activity that needs to be increased therefore needs a lot more improvement to be comparable to, for example, arabinose-, hexitol- or threose-based nucleotides (17, 98). And these xenonucleotides themselves are in turn not efficiently incorporated when compared to dNTP incorporation by DNA polymerases. However, since the reaction mechanism and protein structure are known in sufficient detail, specific regions of the protein could be targeted that have higher chances of containing mutations that increase this promiscuous activity. This choice is also based on general trends in polymerase evolution experiments (17), where mutants with increased specificity or higher catalytic rates carry mutations close to the nucleic acid, both around the active site and in the thumb domain. This rational focusing of mutagenesis to specific promising parts of the protein allowed for higher local mutation rates, thereby exploring a larger part of sequence space in that region since more variation can be introduced. Randomising the entire gene with a mutation rate that produces variants with several substitutions close together in a specific location would most likely burden the protein with unbearable mutational load. Created libraries were analysed before selection by sequencing and were shown to have mutations confined to the desired regions, and at the expected rates. A minority of these clones contained substitutions that create premature stop codons, frameshifted sequences by single base insertions or deletions, or coded for the wildtype sequence, some of the latter with silent mutations but overall, the majority of the clones were those with one or more amino acid substitutions.

Several of these libraries were subjected to CST selection. Mock libraries with a mix of wild type and knockout Therminator genes where the active variant was underrepresented in the population showed that a much more active variant could be increased in allelic frequency by over an order of magnitude by one or two rounds of CST. However, while some libraries underwent up to four rounds of CST, most plasmid libraries were shown to degrade in quality after the second round. This was discovered by gel electrophoresis for re-cloning procedures, where many fragments were shown to be too short, indicating deletion mutants. The exact nature of these deletions was not determined, so it is not possible to know whether these deletion mutants are the result of a specific repeatable cloning artefact or represent spontaneous deletion mutations that occur independently, and if these genetic parasites have been present at undetectably low frequencies in the first library generations or are the unintended result of a change in selection or re-cloning procedures. Since these constructs were so numerous in the last libraries to undergo CST, screening for these libraries would have likely resulted in mainly screening deletion mutants that are probably completely inactive. Therefore, screening was only performed on the secondary libraries that were found to consist mostly of plasmids of the correct size.

To solve this problem of library quality decrease, each step in the re-cloning could be tested and optimized to find out in which step the deletion mutants are formed. This includes potential off-target primer binding that could create smaller PCR product fragments, restriction enzyme star activity and biased or incorrect ligation. Alternatively, if the yield of longer, correct fragments in the reactions is still high enough to contain all library diversity, longer fragments can simply be separated from the shorter ones by gel electrophoresis and gel extraction, in effect adding an extra, strong selection step for correct plasmid size. If this strategy is followed, CST could be performed for many more cycles, thereby reducing library size but increasing the fraction of active variants. Screening will yield more than the handful of clones per 96-well plate after one or two rounds, the chance of hitting improved clones in the library increases, and clones that share some of the same substitutions would show which specific ones are responsible for the beneficial effect.

### 4.3. Library screening and analysis of promising mutants

Since the library size was assumed to be still quite large after one or two rounds of selection, screening would explore only a portion of the selected clones. The advantages of a polymerase activity assay such as this is the feasible scale, the more direct and semi-quantifiable measure of enzymatic activity, and the repeatability.

Several hundred mutants can be sampled on a few 96-well plates. Assuming for the sake of argument that only 0.1% of mutants in a library of $10^5$ independent clones have increased activity (that is, one in a thousand mutations is beneficial), and a 50-fold enrichment took place due to selection, then we expect about 100 beneficial mutants to exist in the starting library that will represent 5% of the population after selection. Using 96-well plates, there should then be about 5 “hits” per plate. This is more or less the observed order of magnitude of clones that score higher than the wild type in these ELISAs.

Screening the activity of a polymerase is most relevant when the reaction strongly resembles the selection conditions during selection. To be a selectable product during CST, biotinylated dCTP needed to be incorporated in the growing primer. Ideally, this only happens after the incorporation of a few natural nucleotides, the dxTTP when the single templating base adenine is encountered, other natural dNTPs, and then at least one biotinylated dCTP. Elongation can then still continue, and it could be argued that the more the primer is extended, the stronger the primer-plasmid binding is and the more biotin moieties will be displayed on the complex, which both enhance affinity chromatography. However, polymerase mutants with very low fidelity might bypass dxTTP use and instead misincorporate dNTPs, or just add many dCTP-biotin to the growing strand. Such mutants do have high fitness in this imposed environment, and will be selected for, but obviously do not have the wanted activity. Such mutants will not perform properly in an ELISA screen, since here the polymerization needs to be sufficiently accurate for an elongated product to be produced that can specifically bind the probe sequence. This means that the activity of a mutant in the ELISA screening is a more direct measure of the desired protein activity, which can also be quantified relative to the wild type and other mutants.

Finally, ELISA measurements can be run in parallel, with replication of the experiment at the level of the lysate, or independently grown colonies from the same clone, which gives a measure of consistency and is capable of validating clones previously thought of as false negatives, or dismissing false positives.

However, this is also one of the major downsides of the ELISA screening. False positives can be clones that during their discovery had higher signals than they should have had. This can be simply the case when during the last colouring steps, a pipetting or washing error occurred causing the well to stain too rapidly. Another possibility is that the expression and purification of the protein occurred exceptionally well. In a reaction where protein concentration is potentially rate-limiting, higher effective expression levels could make an otherwise equally or even slightly less active mutant look like an improved one. False negatives can be caused by the same reasons: too stringent washing, a pipetting error, or a functional clone that did not get expressed properly, will be missed in the screening. For this reason, most clones that showed increased activity were expressed again to see if their improved activity could be repeated, and all of them were sequenced. Some of these clones turned out to be wild type sequences present in the library, which could then serve as an internal control for wild type activity and variation in activity. Some inactive clones were also sequenced, and some of them were quite plausible (premature stop codons, substitution of Thr to Pro between the catalytic Asp residues).

The other pitfall of ELISA is the inherently lower quality of analysis due to the higher throughput. Variation between and within lysates can be substantial, adding uncertainty to the interpretation of the results when differences in activity are within the same order of magnitude as the signal variation. The usage of lysates as a source of the polymerase also comes with the risk of introducing natural nucleotides in the reaction mixture, including the natural counterpart of the dxNT that needs to be incorporated. In our case, even trace amounts of dTTP could then disturb the diagnostic incorporation of dxTTP.

Finally, both the CST and ELISA conditions so far only required incorporation and extension of a single dxTTP. These conditions were chosen because they exert a selective pressure with a relatively low bar. The eventually desired activity is an XNA-dependent XNA polymerase, but since XNA templates are unavailable and the protein obviously stalls upon two incorporations, selection for multiple consecutive incorporations will make the vast majority of mutants effectively incapable of performing the reaction. Finding an active mutant would then presumably take extraordinary luck, which is not a consistently useful search strategy. For this reason, the bar for selection was set rather low. (Ironically, these conditions would then also select against a highly active and specific XNA-dependent XNA polymerase). To investigate whether selected mutants are not simply better at incorporating and extending one dxTTP within a dsDNA strand, ELISA was performed with more demanding templates, requiring two or three consecutive dxNT extensions, or two consecutive ones followed by some natural nucleotides, followed again by one or two consecutive dxNTs. This showed that increased activity of one incorporation was often associated with increased activity for two incorporations, but beyond this, no signal was detected. Since a signal in ELISA can only appear once the probe-complementary sequence is produced, this can mean two things: either the polymerase is stalled completely, or it can incorporate several dxTTPs and maybe even extend them, but not beyond a certain point, stalling before the probe-complement can be finished. The fact that intermediate, shorter products are invisible in ELISA means that other techniques need to be used to visualize the activity in more detail.

For this reason, PAGE was used. PAGE with high amounts of urea can resolve nucleic acids at single base resolution, allowing each potential intermediate to be visualized. This means there are no restrictions to the template sequence that necessitate long extensions. When using fluorescent primers, the detection limit is also very low. Because of the much higher practical investment, it is not feasible to do such experiments on large numbers of mutants, as is the case with ELISA screening. While ELISA has its drawbacks, it was useful for initially finding a small number of the most promising mutants, which could then be investigated with PAGE in more detail. Among these mutants, the G8 clone was quickly shown to have only a narrow set of activities that made it look initially promising, more specifically its faster incorporation of a single dxNT compared to the wild type. This was after all what it was selected for. But it was less active with regards to further extension with more dxNTs, compared to both the wild type and the E2 clone. In contrast, the E2 clone still outperformed all others in both ELISA and initial PAGE measurements. For this reason, all characterization efforts were focused on this E2 clone. The results of these experiments are published as a peer-reviewed research article, which is included in this thesis as Chapter 3.

### 4.4. Mutation types and structure-function relationship

Although only the E2 clone carrying a H545R mutation was thoroughly investigated, a large number of mutants have been found in the initial screenings. This allows us to speculate about the function of specific residues and their tolerance to mutation. While it is not possible to say which specific substitutions confer the phenotypic difference of a mutant with multiple substitutions without studying each substitution mutant independently or in specific combinations, overall trends can be discussed. This is possible because a crystal structure of the DNA polymerase 9°N-7 is available, and most of the mutations are unlikely to cause drastic differences in the peptide backbone structure and the folding of the enzyme. Looking at the specific mutations and the predicted position of those residues in the protein, many of these mutations fall into two classes. They are often either pointed directly towards the bound nucleic acids or other functionally relevant residues, or pointed away from them.

An example of the former is M-region mutant B11, where one of the three consecutive lysine residues (positions 591-593) is mutated to arginine. These three positively charged residues make intimate contact with the DNA minor groove and backbone and form a positively charged cleft, and this substitution is likely to be (nearly) neutral. Functional mutants have also been found with substitutions to Asn or Thr at these positions, indicating at least some freedom to mutate, as long as an overall positive charge or hydrogen bonding potential remains.

The latter are mutations that do not contact the nucleic acid at all, but rather form contacts with other secondary structure elements that themselves are also not interacting with the nucleic acid. This is mostly the case for the many alpha helices surrounding the mutagenized regions (see Figure 2.2). Most of these substitutions involve residues that usually are not that distinct biochemically, but sterically larger or smaller. An example is the I609M mutation that independently occurred in both promising clones G4 and A9. Another example is the A485L terminator mutation itself, which made the 9°N-7 pol more promiscuous. This bias is presumed to be due to sterically forcing secondary structure elements such as these alpha helices apart or closer together, thereby changing distances and relative orientation angles of these helices to each other and the nucleic acid. After all, the changes necessary to accommodate a dxNA/DNA hybrid strand in the nucleic acid-binding regions will still require positive charges and hydrogen bonds to compensate for and bind with the negative phosphate groups in the backbone while it is bound or translocates, and reorienting a secondary structure element within the region itself is presumably not likely to be due to changing a small number of residues in isolation. Rather, a mutation such as the terminator mutation could exert its effect by changing the folding and dynamics of the protein itself. Since this A485L mutation increases this residue in size, steric hindrance could force the finger domain helices away slightly from the exonuclease domain, thereby shifting the dynamic equilibrium between the open and closed state more towards the latter.

A large fraction of the mutations in clones with higher or unchanged activity are substitutions to biochemically similar amino acids, while mutations in inactive clones tend to be more drastic. The most striking example is the inactivating T541P mutation, which is hypothesized to introduce a large backbone structure change right between two catalytic residues. Another particular position is E616, since mutations at this position are common in many clones, with Glu often being changed to the almost identical Asp, but also to the hydrophobic Val. The position of the mutation in the most active clone, H545R, was also shown to tolerate mutation to Leu, Tyr and Gln without losing function, even though these are biochemically distinct.

Several cases of seemingly interchangeable amino acids are the large aromatic ones, or the hydrophobic branched-chain ones (Leu, Ile and Val).

To increase the odds of creating mutants in a library with strong functional differences, it could therefore be useful to take into account more than mere Cartesian distance of residues to the nucleic acid, but specifically target those residues that either have direct points of contact with the nucleic acid, or with other residues in neighbouring secondary structure elements. Residues that are merely solvent exposed will likely be less relevant, as they are constrained within a more or less neutral set of polar amino acids that influence solubility more than binding affinity or protein dynamics. This decrease in positions to undergo mutagenesis within a given region then also directly reduces the sequence space that needs to be searched. This in turn makes sampling from such libraries more efficient. In addition, residues that themselves are too far away for direct contact with the nucleic acid, but are oriented towards alpha helices that are in contact with the nucleic acid on their other side, become potentially interesting targets for further mutagenesis.

# 3

# Characterization of a mutant DNA polymerase with increased activity for deoxyxylonucleotides

This chapter has been published as a peer-reviewed paper: Bauwens B., Rozenski J., Herdewijn P., Robben J., 2018, “A single amino acid substitution in Therminator DNA polymerase increases incorporation efficiency of deoxyxylonucleotides”, ChemBiochem, 19(22), p. 2410-2420. (doi.org/10.1002/cbic.201800411) (99)

## 1. Preface

Listed below are the contributions of the authors of this research article.

Boris Bauwens is lead author of this manuscript and performed all experiments.

Professor Johan Robben supervised Boris Bauwens during the PhD thesis work, including the work on this paper, and corrected the manuscript.

Professor Jef Rozenski performed the oligonucleotide mass spectrometry analyses.

Professor Piet Herdewijn supplied the 2'-deoxyxylonucleoside triphosphates and made his lab infrastructure available at the Rega Institute for Medical Research for PAGE analyses.

Professor Matthy Froeyen is acknowledged for modelling 2'-deoxyxylonucleotide substitutions in a dsDNA helix, constructed a model for the 9°N-7 DNA polymerase with a bound dNTP and with 2'-deoxyxylonucleotide substitutions inserted along the primer strand. This allowed for a structure-based approach for mutagenesis, and a rational framework to interpret mutants after selection.

Professor Arnout Voet is acknowledged for independently modelling a 2'-deoxyxylonucleotide in the +2 position, in combination with the H545R mutation, using MOE (CCG, Canada) which was implemented with the AMBER:EHT force field. These models were subjected to conjugated gradient energy-minimization of the residues in a radius of 10 Å around the R545 guanidinium head. Prof. A. Voet also helped with the interpretation of these results.

Supplementary figures S1-S5 are given in the Addendum.

## 2. Abstract

Deoxyxylonucleic acid (dxNA) is a synthetic polymer which may have the potential of heredity and evolution. Because of its unusual backbone geometry, sequence information stored in dxNA is presumed to be inaccessible to natural nucleic acids or proteins. Despite a large structural similarity with natural nucleotides, 2'-deoxyxylonucleotide (dxNT) incorporation by polymerases is limited. We present the identification of a mutant of DNA polymerase Therminator with increased tolerance to deoxyxylose-induced backbone distortions. Where the original polymerase stops after incorporation of two consecutive dxNTs, the mutant is able to extend incorporated dxNTs with 2'-deoxyribonucleotides (dNTs), and to incorporate up to four dxNTs alternated with dNTs, thereby translocating a highly distorted double helix throughout the entire polymerase. A single His to Arg substitution very close to the catalytic site residues is held responsible for interaction with the primer phosphate groups and for stabilizing nucleotide sugar-induced distortions during incorporation and translocation.

## 3. Introduction

Nucleic acids are a versatile class of molecules that can store genetic information or fold into three-dimensional structures such as aptamers and aptazymes that allow for complex molecular recognition or catalysis, respectively. Xeno nucleic acids (XNA) are their synthetic analogues that can increase this chemical, structural and functional versatility and are potentially evolvable (3, 100). XNA might have a use, e.g., in biosafety, as a “genetic firewall” that prevents sequence information transfer to DNA or RNA (1). Such XNAs are said to be genetically orthogonal. For applications where the XNA polymerase that produces this XNA is present *in vivo*, enzyme orthogonality is also required to avoid interference with the endogenous DNA/RNA systems. XNAs have been produced that differ from natural nucleic acids by modifications of the nucleobases, ranging from small changes to a completely different scaffold (8, 12, 101). It is also possible to modify the universal pyrophosphate leaving group (5) (which would produce natural nucleic acids from xenonucleotides) or the sugar phosphate backbone (8, 102, 103). A large number of substitutes for the (deoxy)ribose sugar and/or the phosphate group has been explored. Even the backbone itself can be fundamentally changed like the glycerol-phosphate backbone of glycerol nucleic acids (GNA) (104) or the *N*-(2-aminoethyl)glycine polyamide backbone of peptide nucleic acids (PNA) (105). One very relevant property of XNA is the potential for structural and/or coding orthogonality *in vivo*, which applies when it does not interact with biological processes like replication, transcription or translation. While there is a large group of potential alternative synthetic nucleotide building blocks to choose from, currently only a fraction of these can be efficiently and accurately incorporated by polymerase enzymes with a broadened substrate spectrum. For example Therminator DNA polymerase can transcribe DNA into threose nucleic acid (TNA) sequences (106) (much more than the closely related Vent polymerase (107)), up to five GNA (108) and up to seven flexible nucleic acid (FNA) nucleotides (109). Evolved variants of TgoT DNA polymerase have allowed for the enzymatic production of 1,5-anhydrohexitol nucleic acids (HNA), arabinonucleic acids (ANA), 2'-fluoro-arabinonucleic acids (FANA), cyclohexenyl nucleic acids (CeNA), 2'-O,4'-C-methylene-β-D-ribonucleic acids or locked nucleic acids (LNA), TNA (17) and 5′-O-phosphonomethyl-threosyl nucleic acids (tPhoNA) (66). Most of the XNAs with altered backbones that have been produced by existing or evolved polymerases maintain canonical base pairing with DNA and/or RNA. Enzymatic production of a structurally orthogonal XNA that does not hybridize to natural nucleic acids has not been achieved so far, presumably because the XNAs with large structural differences are hard to accommodate and translocate in an active site that is optimized for DNA or RNA, even by a polymerase with promiscuous activity. Enzyme engineering that relies on evolution is also limited to polymerases dependent on primer-template complex formation, and assumes that the distance in sequence space between a DNA polymerase and XNA polymerase is not too large to be bridgeable by mutagenesis. One such orthogonal nucleic acid would be 2’-deoxyxylonucleic acid (dxNA) which is identical to DNA except for the ribose C3’ stereochemistry, thereby replacing 2’-deoxyribose by its epimer, 2’-deoxyxylose (Figure 3.1). Because the C3’ oxygen atom is part of the backbone, this inversion drastically changes the backbone structure.

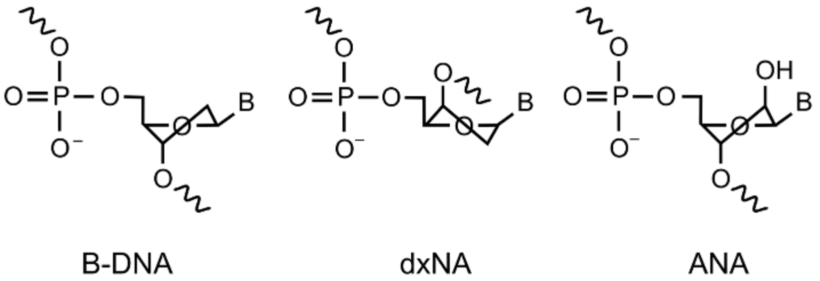


**Figure 3.1: Backbone configuration of 2’-deoxyribose-based DNA, 2’-deoxyxylose-based dxNA and arabinose-based ANA.**

The β-oriented 3’ hydroxyl in dxNA drastically alters the nucleic acid backbone geometry. dxNA has a C3’-endo pucker, while B-DNA has a C24-endo pucker.

It has been reported that *in vivo*, ROS-induced epimerization of 2’-deoxyribose to 2’-deoxyxylose in DNA can hinder replication and induce mutation (110).  Studies based on molecular dynamics (22) and NMR and CD (20, 111) have shown that a fully substituted dxNA strand forms extended, left-handed double helices with an antiparallel strand where Watson-Crick base pairing rules still apply. Complementary dxNA strands anneal with affinities similar to that of DNA, but show no hybridization with either RNA or DNA. Comparing dxNA to multiple other XNA structures (103) shows that dxNA is one of the XNAs with the most drastically altered backbone conformation: the helix is much more elongated, with a very large major groove and the backbone tracing a zig-zag path. The large difference in structure makes dxNA a promising, truly orthogonal, but polymerase-challenging XNA candidate: even sequences with canonical bases would be invisible to natural replication, transcriptional or translational machinery because of a lack of recognition through strand hybridization or backbone interactions. *In vivo* studies using plasmids carrying templating dxNTs (112) have shown that these dxNTs might be read through by cellular polymerases, but the capability to do so drastically lowers with increasing amounts of templating dxNTs. This further strengthens the case that orthogonality is already achieved around two dxNT codons or more.

While chemical synthesis of (oligo)nucleotides is possible, enzymatically manipulating or sequencing dxNA strands is difficult and inefficient. Enzymes like polymerases, primases, ligases, endonucleases etc. that are specific for dxNA would open up the same possibilities that exist for DNA and RNA, but would either need to be designed or obtained through molecular evolution (113). Precisely because of the orthogonal character of these 2'-deoxyxylonucleotides, no enzymes are available that efficiently catalyse their polymerization. Experiments with different polymerases have revealed dxNTs to be very resistant to incorporation and extension, with no more than three consecutive incorporation events on homopolymeric templates (20).

In this study, we attempted to improve on the existing promiscuous activity using molecular evolution of Therminator DNA polymerase, derived from Thermococcus sp. 9°N-7 DNA polymerase by mutation A485L and made exonuclease-deficient by mutations D141A and E143A (93). Using compartmentalized self-tagging (CST) (84), targeted libraries of this thermostable, archaeal family B polymerase were selected for incorporation of one or two 2'-deoxyxylothymidine nucleotides (dxTs), followed by a number of natural nucleotides (dNTs). We report here on the characterization of selected mutants that allowed us to pinpoint a single amino acid substitution improving dxNT incorporation and translocation.

## 4. Experimental section

### 4.1. Mutant libraries

The synthetic gene was subcloned in a pASK-IBA2 plasmid derivative kindly provided by Misha Soskine. Mutant libraries were constructed using inverse PCR with partially degenerate spiked primers. Every nucleobase within a targeted region had a 3% chance of being mutated into one of the other three bases. Seven positions were partially randomized, the two catalytic aspartates were not subject to targeted mutation. Nucleotide substitutions within the active site library (L537 to T547) were calculated to average around 2.4 per mutant, resulting in an average of 1.8 amino acid substitutions within this nine-amino acid region. Library diversity was estimated at 8x10^5. The minor groove binding region library spanned residues F588-V596 and T605-E617.

### 4.2. Compartmentalized self-tagging (CST)

Selection was performed using CST as described by Pinheiro *et al*. (84), but in the absence of misincorporation-promoting $Mn^{2+}$ ions, and labelling by incorporation of biotinylated dCTP during extension. This dCTP was linked to biotin through a 16 atom linker between the biotin COOH group and the cytosine C5 (forming γ-[N-(Biotin-6-amino-hexanoyl-6-aminobutanoyl)]-5-(3-propargylamino)-2'-deoxycytidine-5'-triphosphate). An oil phase (mineral oil with 4.5% Span 80, 0.4% Tween 80 and 0.05% Triton X-100) and aqueous phase (1X Thermopol buffer (New England Biolabs), $MgSO_4$ (1 mM), biotinylated dCTP (20 µM), equimolar dATP/dGTP mix (200 µM), dxTTP (200 µM), primer (0.38 µM), BSA (1 mg/mL), DTT (1 mM), glycerol (10% v/v) were emulsified in a 4:1 ratio by adding the aqueous phase dropwise to the oil phase while stirring with a small magnet.

Electrocompetent E. cloni® cells were transformed with plasmid libraries and plated on LB agar with ampicillin. Colonies grown overnight were harvested by adding LB medium and gently scraping them from the plates. The suspension (~30 mL, ~30 OD) was mixed with glycerol (60%) in a 3:1 ratio and flash-frozen in 100 µL aliquots. One aliquot was used to inoculate a 4mL starting culture grown overnight (37°C, 200 rpm). 50 µL overnight culture in 4 mL LB + Amp medium was grown at 37°C, 200 rpm to 0.7 $OD_{600}$, induced with AHT, and grown at 25°C overnight. Cultures were diluted to 1 $OD_{600}$, 100 µL cell suspension was centrifuged at 12000 *g* and supernatant was removed. These cells were added to the aqueous phase before emulsifying.

The elongation reactions were run for 10 min at 95°C, 15 min at 58°C and 15 min at 72°C. The primer for the first round of CST required incorporation of GAAAAG**xT**GC-biotin. In the second round the selective pressure was increased by requiring incorporation of AAAAAGGGAA**xT**AAGGGC-biotin (See Figure S5).

After CST, the emulsion is broken by centrifugation at 13000 *g* for 5 min and two water-saturated diethyl ether extractions. Primer-plasmid complexes are purified using spin columns (GE Healthcare S400 microspin columns) and incubated with magnetic beads (NEB streptavidin magnetic beads, 4 mg/mL) for at least 1h at room temperature on a rolling platform. The beads are washed twice with binding buffer (0.5 M NaCl, 20 mM Tris-HCl, 1 mM EDTA, pH 7.5), once with low salt buffer (0.15 M NaCl, 20 mM Tris-HCl, 1 mM EDTA, pH 7.5), and eluted in elution buffer (25 µL of 10 mM Tris-HCl, 1 mM EDTA, pH 7.5). Re-cloning is done by PCR amplification of the variable half of the Therminator gene using 2 µL of the eluted fraction as template, and PCR amplification of the non-variable half of Therminator and the rest of the plasmid backbone from the original wild type, followed by restriction digestion and ligation of both fragments to reconstitute the original plasmid.

### 4.3. Polymerase activity assay

After selection, libraries enriched in mutants with improved incorporation activity were screened using a previously described polymerase activity assay (17). Re-cloned plasmid libraries were used to transform E. cloni® cells (Lucigen) using electroporation and plated on LB + Amp agar plates. Colonies were inoculated in a 96-well plate with deep, rounded wells. Cells were grown in 400 µL cultures, induced to express the polymerase with anhydrotetracyclin (AHT) at 25°C and 200 rpm overnight, and lysates were prepared for each mutant by heating 20 min at 80°C, centrifuging and resuspension in 1X Thermopol buffer (20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton X-100 at pH 8.8). The elongation reaction uses the cell lysate as a source of polymerase, and a biotinylated primer is elongated with a template that requires incorporation of the sequence AAGGGAA**xT**AAGGGC using dATP, dGTP, dCTP and dxTTP, followed by an non-T sequence that will hybridize with a DNA probe (through an intermediary linker). Polymerase mutants that succeed in incorporation of dxTTP opposite of the templating adenine and elongating past the resulting distortion produce elongated primers that can bind this DNA probe, which carries a digoxigenin group. After binding the primers to a streptavidin-coated 96-well plate, duplex denaturation and washing with PBST (125 mM NaCl, 16.6 mM $Na_2HPO_4$, 8.43 mM $NaH_2PO_4$, 0.2% (v/v) Tween 20, pH 7.2), they are incubated with an anti-digoxigenin antibody fused to horseradish peroxidase (HRP) (dilution 1:10,000). After washing away non-specifically bound probes and antibodies with PBST, the chromogenic substrate TMB (100 µL) is added which turns blue in the presence of HRP. The $OD_{450}$ nm is an indication of the amount of fully elongated primers, and larger signals correlate with more active polymerase containing lysates.

### 4.4. Polyacrylamide gel electrophoresis (PAGE)

Elongation of 5' Cy5 tagged fluorescent primers was performed with templates that introduce different or increasing requirements for the DNA/dxNA polymerases. Reactions of 20 µL (1x Thermopol buffer, 1 mM $MgSO_4$, 250 nM template, 125 nM primer, 100 µM d(x)NTPs [unless specified otherwise] and 2 µL purified polymerase) were incubated in a T300 Thermocycler (Biometra, Göttingen, Germany) at 72°C, 5 µL samples were taken at indicated times after a short spin using a tabletop centrifuge and quenched in 10 µL quenching buffer (95% formamide, 18mM EDTA, 0.025% SDS and 0.05% bromophenol blue). Denaturing polyacrylamide gels were prepared with acrylamide (15%) and urea (7 M) in TBE buffer (89 mM Tris, 89 mM boric acid, 2 mM EDTA, pH 8) with 0.08% APS and 0.04% TEMED. Samples of 2 µL were loaded (estimated to contain roughly 70 µmol primer per sample). Constant currents were set at 25 mA (single gel) or 50 mA (two gels). Gel dimensions were 20 cm x 30 cm x 1 mm. Gels were imaged on a Typhoon FLA 9500 (GE Healthcare Life Sciences) at 635 nm.

### 4.5. Protein purification

*E. coli* cell cultures containing the pASK-IBA plasmid encoding the Therminator exo$^-$ wild type or mutant polymerase were grown in LB with ampicillin (4 mL) at 37°C, 200 rpm up to an $OD_{600}$ of 0.7. Expression was induced using AHT and 100 mL cultures were grown overnight in LB with ampicillin at 25°C. These were then centrifuged at 3400 *g* in 50 mL falcon tubes for 12 min. The cell pellet was resuspended in 30 mL phosphate buffer (50 mM, pH 7.5) with 1 mM

$MgCl_2$ and 20 µL of DNaseI (1  U/µL), frozen at -80°C for 20 min, thawed at 37°C, and ~10 mg of lysozyme and 60 µL Triton X-100 were added and incubated at room temperature for 30 min. This suspension was then sonicated twice for 90 s and centrifuged at 4°C at 3400 *g*. The supernatant was heated to 80°C for 20 min and centrifuged (15 min at 4°C, 3400 *g*). His-tag carrying polymerases were purified from the supernatant using Ni-NTA resin in PD miniprep G-25 columns. Expression levels were estimated using SDS-PAGE and relative concentrations (mass per volume) were normalized by dilution to roughly 25 µg/mL.

### 4.6. Mass Spectrometry

Elongation reactions of 80 µL each were run for 6 cycles of 30 min at 72°C and 1 min at 30°C, with 20 µM primer (5'CAGGAAACAGCTATGAC3'), 20 µM template (5'CGCTGCCGCAGTCATAGCTGTTTCCTGCC3' for dxTTP, 5'CGCAGCCGCTGTCATAGCTGTTTCCTG CC3' for dxATP), 100 µM d(x)NTPs, Thermopol buffer,) $MgSO_4$ (1 mM) and purified polymerase (8 µL). Samples were purified twice by ethanol purification. Electrospray ionization mass spectra were obtained in negative ion mode on a quadrupole/time-of-flight mass spectrometer (Synapt G2, Waters, Milford, MA) equipped with a standard ionization source. The instrument was tuned to a resolution of 15000 (full peak width at half maximum) and the mass accuracy of the instrument was 1 ppm or better using leucine enkephalin as internal calibrant (lock mass). Masses for the oligonucleotides were obtained by deconvolution of the spectra using the MaxEnt algorithm of the software (MassLynx 4.1, Waters).

### 4.7. High-Performance Liquid Chromatography (HPLC)

HPLC on a C18 reversed-phase column (PepMap 0.5 x 15 mm, LC Packings, Amsterdam, The Netherlands) was performed using a buffer containing N,N-dimethylaminobutane (DMAB, Acros, Geel, Belgium) as ion pairing reagent and 1,1,1,3,3,3-hexafluoro-2-propanol (hexafluoro isopropanol, HFiP, Acros, Geel, Belgium). In brief, the solvent system consisted of acetonitrile 84% (vol/vol) (Fisher Scientific, Loughborough, UK) as organic phase and DMAB 0.05% (vol/vol) with HFiP 1% (vol/vol) in water as aqueous phase (pH 8.0). Oligonucleotides were eluted with a flow rate of 12 µL/min applying a gradient starting at 2% organic phase and increasing by 2% per minute during 15 min. The concentration of the oligonucleotide samples was around 10 µM and between 30 and 480 pmol product was injected per run.

## 5. Results

### 5.1. Library construction, selection and screening

Mutant libraries were designed based on the crystal structure of the Therminator precursor, 9°N DNA polymerase bound to DNA (PDB 4K8X). Six regions of the protein in close contact with DNA were defined and partially randomized in exonuclease deficient (exo$^-$) Therminator, so as to allow for a local high amount of mutations. Of the six libraries, two were used in this study (active site library and minor groove binding region library; see Experimental section). The polymerase catalytic Asp residues themselves were left unchanged. These libraries underwent selection by compartmentalized self-tagging (CST) (84) for the incorporation of dxT. Two modifications of this CST protocol were made: primers were tagged by biotinylated dC incorporation, given that short and destabilizing extension was expected, and no $Mn^{2+}$ was added. The template sequences were chosen such that 9 nucleotides needed to be incorporated (the 7th of these being a dxT) in the first round of selection, and 18 in the second round (the 12th being a dxT), before incorporation of a first biotin-dC. This relatively low selective pressure was meant to initially select for mutants that can build in one or two dxTs into DNA and translocate the dxT-induced backbone distortions along the polymerase. The next logical step would then be to further evolve distortion-tolerant mutants by gradually increasing the dxT incorporation load.

Resulting clones were screened using a DNA-based ELISA of cell lysate for the presence of polymerase variants capable of elongating a primer-template complex. Full elongation required the incorporation and elongation of seven natural non-T nucleotides, a single dxT and 27 natural non-T nucleotides, consecutively. Elongation downstream of the introduced modification was probed with a complementary, fluorescently labelled oligonucleotide. Three of 184 clones scored much higher in signal intensity compared to wild type, the highest of which was clone E2 (Figure 2.6A). Several clones from the active site library and the minor groove binding region library were further tested using purified protein and more demanding templates requiring consecutive and intermittent incorporation of up to four dxTs. The clones E2 and G8 consistently showed higher ELISA activity compared to wild type and the other mutants (Figure 2.6C). DNA sequencing revealed E2 to contain a single amino acid substitution, H545R, and G8 to contain mutations R613H and E617R. Incorporation efficiencies of purified polymerase variants E2 and G8 were further characterized by analysis of primer elongation products with denaturing polyacrylamide gel electrophoresis (PAGE).

### 5.2. Homopolymeric templates

Three polymerases (Therminator, E2 and G8) were tested with different nucleotide types, concentrations and incubation times for comparison with previously published (20) dxNTP incorporation data on Therminator and other polymerases. Both Therminator exo$^-$, for simplicity further denoted as Therminator or wild type, and mutant E2 are capable of incorporating up to two consecutive dxTs and up to three dxAs on homopolymeric polyA and polyT templates, respectively (Figure 3.2). Mutant G8, which had shown promise in the ELISA screening, performed similar to E2 with dxA as a substrate, however incorporated virtually no dxT. Because the E2 mutant performed much better than G8, G8 was not tested further. While both wild type and E2 polymerases extended all primers with at least one dxNT within 5 min, the wild type took 45 min to elongate all primers with a second dxNT. In contrast, E2 only partially added a second dxNT under the same conditions.

Incubation with different equimolar nucleotide concentrations of dNTPs and dxNTPs showed that the lowest concentration of 100 µM was optimal to evaluate incorporation. Increased concentrations of up to 200 and 500 µM did not visibly improve the dxNT incorporation rate, but did increase misincorporation rates in the otherwise more accurate E2 polymerase. Therefore, all further elongation reactions were performed with 100 µM nucleotides. These were kept equimolar to minimize anticipated incorporation bias due to concentration differences.

To investigate effects of dxA and dxT combination on primer elongation by wild type and E2 polymerase, templates were used requiring the incorporation of different combinations of three dxA and dxT nucleotides (Figure 3.3). For all these templates, only up to two dxNTs were incorporated, with no visible difference in incorporation efficiencies between dxA and dxT. The only exception was incorporation against the ATT template where trace amounts of +3 product was possibly formed by E2, quantifiable by digital image analysis. The presence or absence of dGTP and dCTP in the reaction mixture showed no significant difference, neither did TAT or ATA templates (Figure S2).

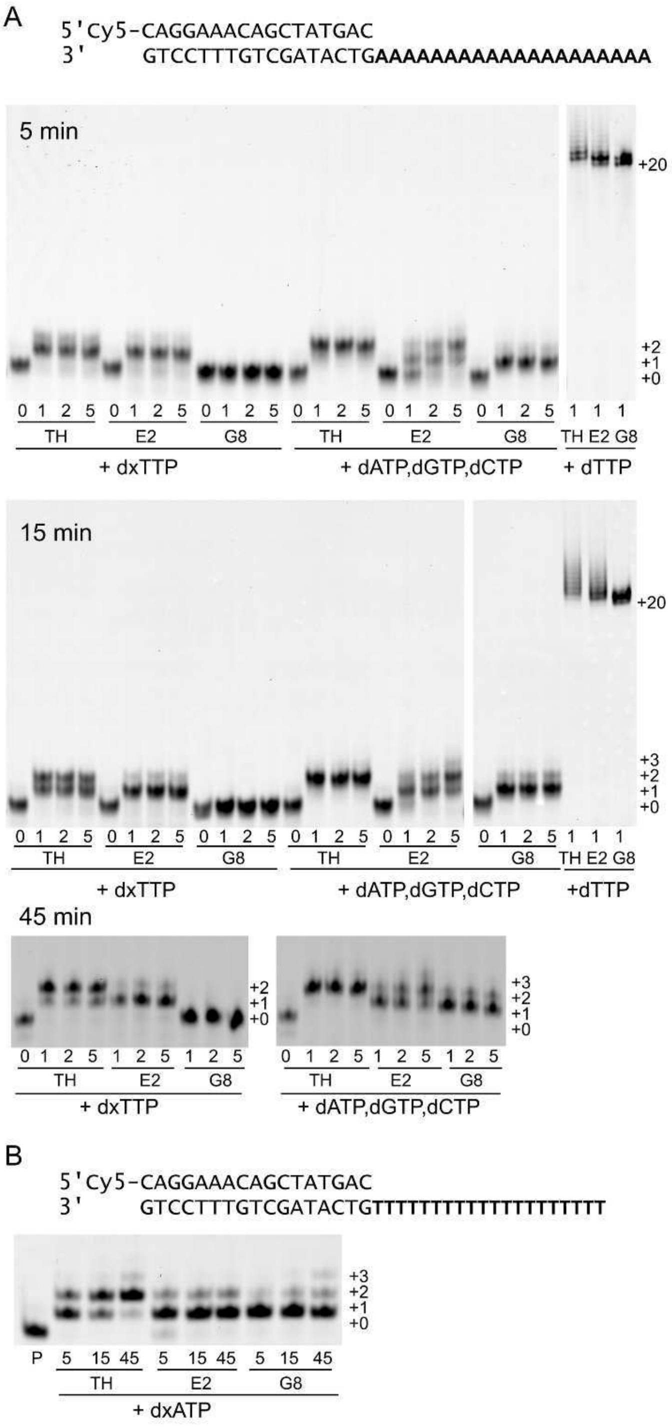


**Figure 3.2: Effects of nucleotide type, nucleotide concentration, and incubation time on primer elongation products from homopolymeric templates by Therminator exo$^-$ (TH) and mutant polymerases E2 and G8.**

**A)** Primer-template complex requiring incorporation of dxNT (template in bold). Elongation products after 5, 15 and 45 min. Lanes 0, 1, 2 and 5 are with dxTTP concentrations of 0, 100, 200 and 500 µM dxTTP. Misincorporation controls are with dATP, dGTP and dCTP but without either dTTP or dxTTP. Positive controls (top right) with 100 µM dTTP. **B)** Primer-template complex requiring incorporation of dxNA (template in bold). Elongation products with 100 µM dxATP. Lanes 5, 15 and 45 are samples taken after 5, 15 and 45 min. P = unextended primer.

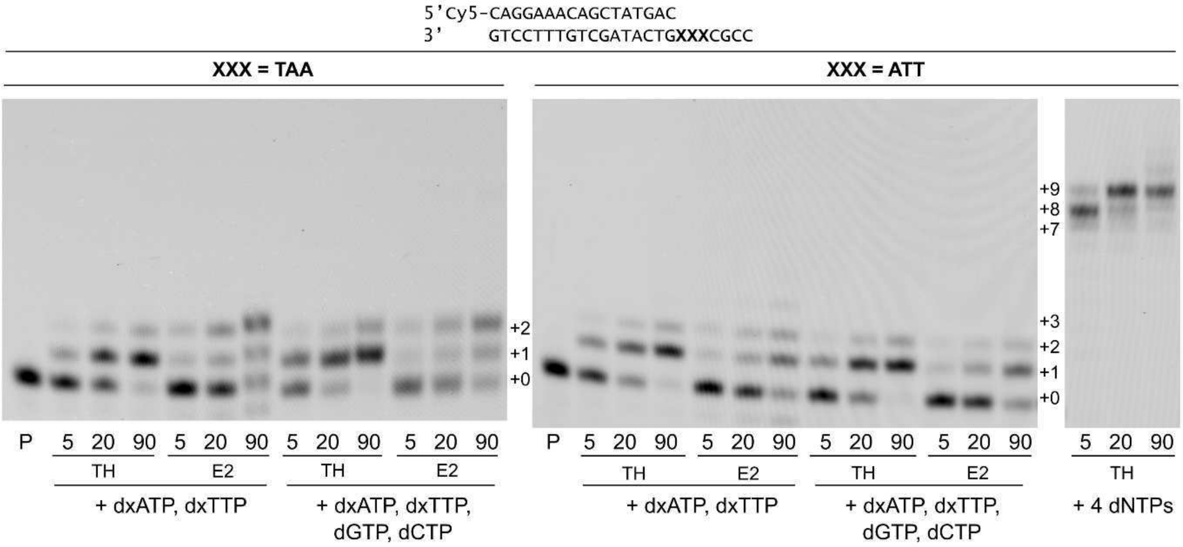

5'Cy5-CAGGAAACAGCTATGAC
3' GTCCTTTGTCGATACTGXXXCGCC
XXX = TAA
XXX = ATT
+2
+1
+0
+9
+8
+7
+3
+2
+1
+0
P 5 20 90 5 20 90 5 20 90 5 20 90
TH E2 TH E2
+ dxATP, dxTTP
+ dxATP, dxTTP, dGTP, dCTP
P 5 20 90 5 20 90 5 20 90 5 20 90
TH E2 TH E2
+ dxATP, dxTTP
+ dxATP, dxTTP, dGTP, dCTP
5 20 90
TH
+ 4 dNTPs

**Figure 3.3: Primer elongation with both dxATP and dxTTP by wild-type Therminator exo$^-$ (TH) and mutant polymerase E2.**

The TAA template (XXX = TAA) was designed to allow incorporation of one dxA and two dxT, the inverse with the ATT template. Reactions ran with dxNTPs in the absence or presence of additional dGTP and dCTP. Samples were taken after 5, 20 and 90 min. Right: positive control with natural dNTPs. P = unextended primer.

### 5.3. Elongated of two incorporated dxTs with dNTPs

Our observation that extension blocks after two dxNTP incorporations brought up the question whether elongation stops because no more nucleotides can be incorporated beyond two consecutive dxNTs, or because the polymerase cannot incorporate three consecutive dxNTs. When templates are used that require not only the incorporation of two initial dxTs but also elongation with non-T dNTs, E2 showed a distinctly higher activity compared to Therminator exo$^-$ (Figure 3.4). It took the wild type up to 45 min to produce small amounts of full-length products, while E2 already showed some complete elongation within 5 min and 38% of primers are fully elongated after 45 min. However, in the absence of dTTP and dxTTP, full-length products were also produced. This indicates that two consecutive mismatches did not stall either polymerase and raises the question whether full-length products observed in the presence of dxTTP were formed after genuine dxT incorporation, or were due to non-T mismatches (see further mass spectral analyses, which show the presence of extended products containing a dxNT).

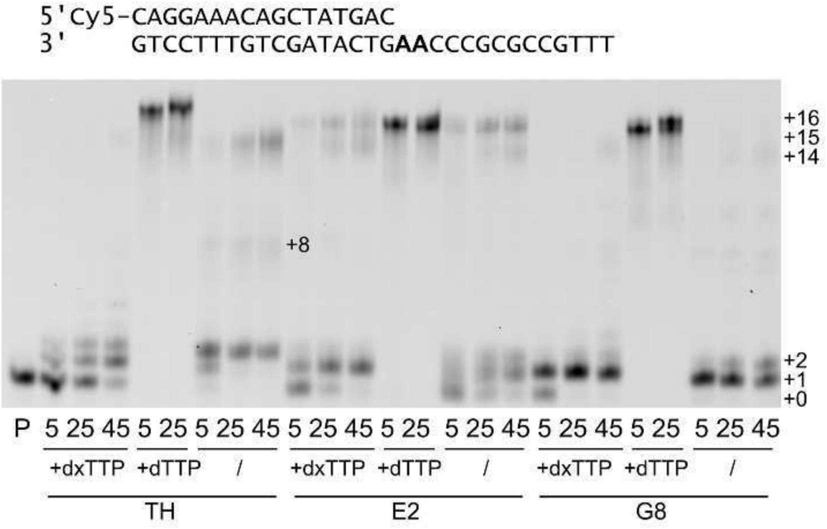

5'Cy5-CAGGAAACAGCTATGAC
3' GTCCTTTGTCGATACTGAACCCGCGCCGTTT
+16
+15
+14
+8
+2
+1
+0
P 5 25 45 5 25 5 25 45 5 25 45 5 25 5 25 45 5 25 45 5 25 5 25 45
+dxTTP +dTTP / +dxTTP +dTTP / +dxTTP +dTTP /
TH
E2
G8

**Figure 3.4: DNA elongation downstream of incorporated dxTTP by wild-type Therminator exo$^-$ and the E2 and G8 mutants.**

Samples were taken after 5, 25 en 45 min. P = unextended primer. All reactions were in presence of dATP, dGTP and dCTP. Either dxTTP, dTTP or no further nucleotides were added. Bold: templating bases requiring incorporation of dxNT.

### 5.4. Elongation with alternating dxNTs and dNTs

Because two consecutive dxNTs impede elongation with a third dxNT while extension with a large number of dNTs is possible, we tested how many local dxNT-induced distortions could be tolerated, and how close they can be spaced. Elongation reactions that required the immediate incorporation of one or two dxNTs, followed by either 1, 2, 4 or 6 natural nucleotides, another dxNT and a stretch of dNTs showed very little end product with the wild type polymerase, while E2 did produce measurable amounts (Figure 3.5). Either dxTTP or dxATP was used as substrate, in combination with the other three normal dNTPs. dxA was incorporated and extended more efficiently, and the amount of full-length product was substantially increased, but the overall patterns were the same. Unextended +1 and +2 products were present in E2 polymerase reactions, but mostly in wild-type reactions, especially with dxTTP as substrate. Yields were higher for the incorporation of a single dxNT compared to two consecutive ones. However, increasing the distance to the next dxNT did not significantly affect full-length product yields, rather the total number of dxNTs in the elongated sequence seems to be the determining factor for production of full-length products.

Intermediate products accumulated at the first two incorporation steps, as well as at the step prior to incorporation of a next dxNT.

To investigate the effects of the number and distance between 2’-deoxyxylose-induced distortions on polymerase activity, templates were used that required the incorporation of between 1 and 5 equally spaced dxNTs within a stretch of 10 nucleotides, followed by further incorporation of dNTs to a total length of +20 (Figure 3.5). Again, dxATP appeared a better substrate than dxTTP. Main intermediate elongation products accumulated prior to incorporation of the second dxNT (roughly one third of the elongation products made by E2 in the presence of dxTTP substrate after 90 min). For templates where the dxNT spacing is close, intermediates were formed at nearly every incorporation step. The wild-type polymerase did not produce full-length products, except small amounts (5%) from the least challenging template requiring incorporation of one dxT, and did not produce measurable amounts of intermediates longer than +5. In contrast, E2 produced full-length products for all but the most demanding template requiring five dxNT incorporations in close succession, although in very low amounts, which are probably in part the result of misincorporation elongation. Nevertheless, synthesis of a dxNT-containing primer strand (as in reaction 3B, Figure 3.5B) implies the possibility of elongating and translocating a large amount of steric distortions. While the data cannot conclusively show that this sequence is produced with high fidelity, it does show that E2 can nevertheless tolerate substantial distortions. In the case where a single dxT was to be incorporated, followed by 19 dNTs (Figure 3.5A, reaction 3D), the majority of primer molecules were fully elongated, showing that in a reaction very similar to that in CST selection, E2 readily incorporates a dxNT into DNA and clearly outperforms the wild type.

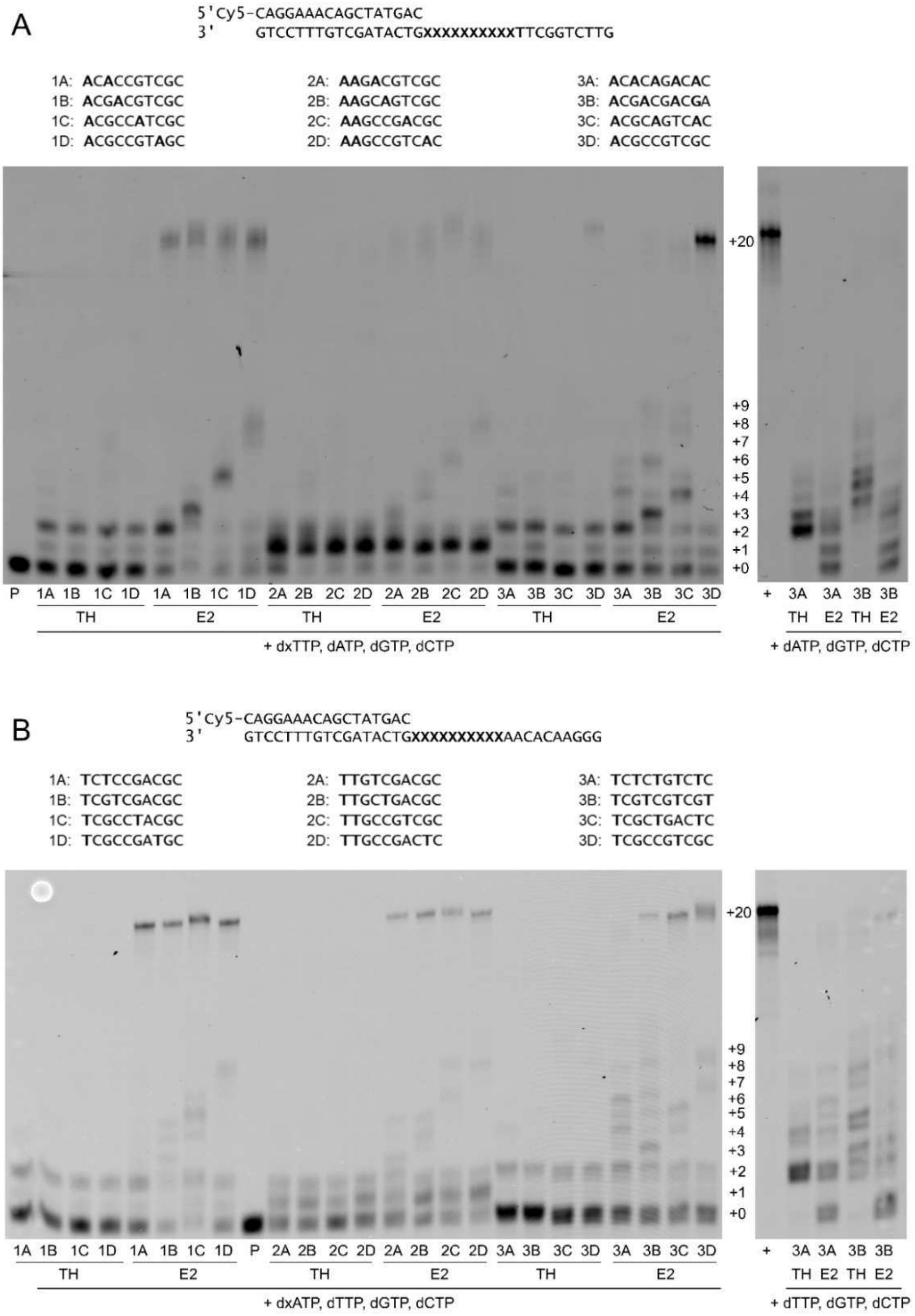


**Figure 3.5: Elongation by polymerases Therminator exo$^-$ (TH) and E2 using templates requiring mixed dxNT and dNT incorporations.**

Samples were taken after 90 min. **A)** Elongation with dxT. Templates 1A, 1B, 1C and 1D require incorporation of one dxT, followed by 1, 2, 4 or 6 dNTs respectively, another dxT, and dNTs to a total length of +20. Templates 2A, 2B, 2C and 2D require the same but with two initial dxTs instead of one. Templates 3A, 3B, 3C and 3D require incorporation of 5, 4, 3 or 1 dxT over 10 bases, followed by 10 dNTs. **B)** Elongation with dxA, similarly to panel A, with the exception that template 3D requires two spaced dxA instead of one. + = positive control with TH and dNTPs. Misincorporation controls (right gels in panels A and B) are performed without dxNTP using templates 3A and 3B.

### 5.5. Terminal transferase activity

Because the full-length products often showed distinct bands that were longer than anticipated, and to find out if these polymerases are able to perform template-independent elongation, elongation reactions were performed with a blunt-ended primer-template complex (Figure 3.6). Reactions with each of the four natural nucleotides were incubated in parallel. Wild-type polymerase extended the majority of the primers with an untemplated nucleotide already within 15 min, the exception being dC, which was added to only half of the primers. In contrast, E2 took 45 min before showing visible +1 products in the presence of dATP or dGTP, which did not increase by much over the next 45 min. Wild-type polymerase, on the other hand, showed partial addition of a second purine nucleotide in the absence of a templating base after 90 min of incubation.

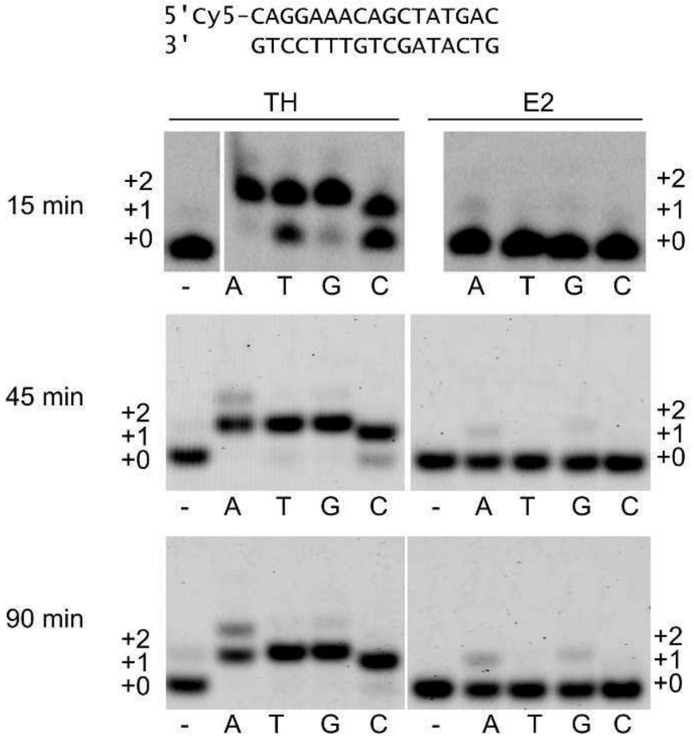


**Figure 3.6: Terminal transferase activity by polymerase Therminator exo$^-$ (TH) and the E2 mutant on a blunt primer-template complex.**

Either dATP, dTTP, dGTP or dCTP (100 µM) was tested as substrate.

### 5.6. Specificity for 2’-deoxyxylose versus arabinose nucleotides

Successful selection on dxNT incorporation can result in increased specificity for dxNT but also a more general broadening of promiscuous activity, while at the same time being constrained by the need to be able to still produce DNA. Based on previous results in our lab, promiscuous activity for ANA by Vent polymerase was shown. We therefore tested incorporation of arabinonucleotides, which have a sugar pucker very similar to DNA, even though the arabinose sugar in the backbone is an epimer of the ribose in RNA nucleotides. We expect that a tolerance mechanism that merely removes steric hindrance created by backbone distortions, which makes DNA-protein interactions less specific (for example by widening the groove through which the formed nucleic acid translocates) would also show improved activity for other xenonucleotides that can be promiscuously incorporated by the wild type. On the other hand, increasing specificity for dxNTPs can come at the expense of activity for other (xeno)nucleotides. A homopolymeric polyT template ($T_{20}$) and a template with five Ts alternated with single non-Ts ($(TV)_5V_{10}$) were used to elongate a primer in presence of arabinose adenosine triphosphate (araATP) and natural nucleotides as a control (Figure 3.7).

Using the template $T_{20}$, the wild-type polymerase readily incorporated 15 to 18 araA within 15 min, while E2 stalled after two and incorporated no more than four nucleotides. The ladder of products extending beyond the expected maximal length may be explained by polymerase slippage by cycles of polymerase dissociation, followed by mismatched reannealing of the primer-template complex, polymerase re-association and further elongation(114, 115).

Reactions requiring either incorporation of five araAs alternated with a non-A dNT, or of araAs only, quickly yielded full-length products (of both expected length and slightly longer ones) by the wild type. On the $(TV)_5V_{10}$ template, E2 first accumulated araA-containing +6 extension products, and produced a majority of full-length products after 90 min only, and could incorporate no more than four consecutive araAs on the $T_{20}$ template.

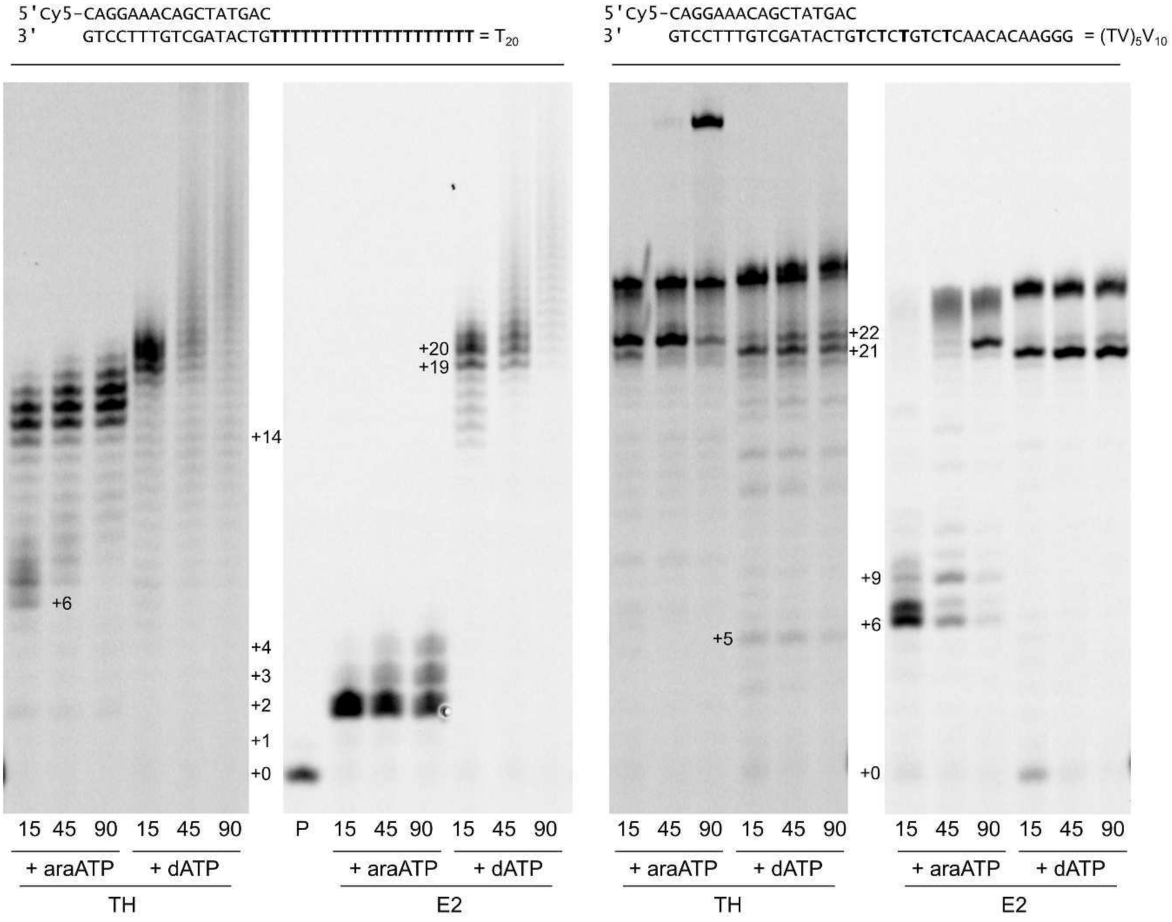


**Figure 3.7: Specificity for arabinose adenosine triphosphate (araATP) versus natural dATP substrates of wild-type polymerase (TH) and the E2 mutant.**

dGTP, dTTP and dCTP were also added when using template $(TV)_5V_{10}$, with V = non-T. Samples were taken after 15, 45 and 90 min. P = unextended primer.

### 5.7. dxNT versus mismatch incorporation

Mass spectrometry was used to distinguish genuine dxNT from mismatch incorporation. Templates required incorporation of one dxNT and nine dNTs. Because dxNTs and their natural isomeric counterparts cannot be distinguished by mass alone, none of the reactions in this study were incubated with both isomers combined. Therefore, masses consistent with the addition of a matching nucleotide in the absence of the natural one were used to indirectly verify of incorporation of the correct dxNT. Elongation products were shorter than expected from PAGE gels, presumably due to the much larger DNA-to-polymerase ratio in the reactions. Nevertheless, elongation products were found where either dxA or dxT was incorporated and extended with the correct natural nucleotide (Figure S3). In the presence of an obligate dxTTP, the Therminator (exo$^-$) polymerase performs purine misincorporations but in lower relative amounts compared to dxT incorporation (Figure S4). Such misincorporations were not observed using E2, or in the presence of an obligate dxATP for either polymerase. Mixtures of only three of the four dNTPs yielded misincorporation and elongation products to up to +7 nucleotides. All four bases were prone to misincorporation, with dA misincorporation dominating in the absence of d(x)TTP (Figure S4). Misincorporated dNTs were also less elongated by E2 compared to wild type. Using wild type in the presence of four natural dNTPs produced mostly full-length product extended by one extra dA (Figure S4).

## 6. Discussion

Despite their chemical similarity to natural nucleotides, dxNTPs are a challenging substrate for DNA polymerases. Their sugar ring conformation in dxNA has an C3’-endo pucker (see Figure 3.1), like RNA and A-DNA, compared to C2’-endo pucker of B-DNA (19). The 3’-hydroxyl is inverted compared to DNA (Figure 3.1), and even though this structural difference with a dNTP is limited, they are still incorporated less efficiently by most tested polymerases (20), including Therminator DNAP and the Therminator M3 mutant. This M3 mutant has a single substitution (L408Q) that confers increased activity for ribonucleotides and C5-modified nucleotides (116), and also shows promiscuous but low activity for dxTTP and dxATP incorporation, comparable to the activities of wildtype Therminator and 9°N exo$^{-}$. We confirmed that the incorporation of a first dxT or dxA by the 9°N and related DNA polymerases is still relatively fast compared to producing full-length products. The latter can take up to 90 min to yield visible bands in PAGE, while substantial amounts of primer extended with at least one dxNT can already be observed in 5 min samples. Once incorporated and further extended by a second dxNT, these nucleotides are expected to introduce a steric distortion in the backbone that affects the position of the phosphate and the relative orientation of the sugar ring and nucleobase, potentially inducing additional distortions in upstream and downstream primer nucleotides through neighbour effects, as well as in the opposite template strand. This is presumably the main difficulty for elongation of a dxNA backbone by a DNA polymerase (more so than incorporating the synthetic nucleotide substrates themselves). These distortions may interact with multiple residues in the DNA polymerase thumb domain during translocation. Since there are many points of contact between the thumb domain and the template and primer (Figure 3.8), and the thumb domain wraps tightly around most of the duplex for roughly one helical turn, blocked translocation rather than incorporation of dxNTs may become the elongation limiting step.

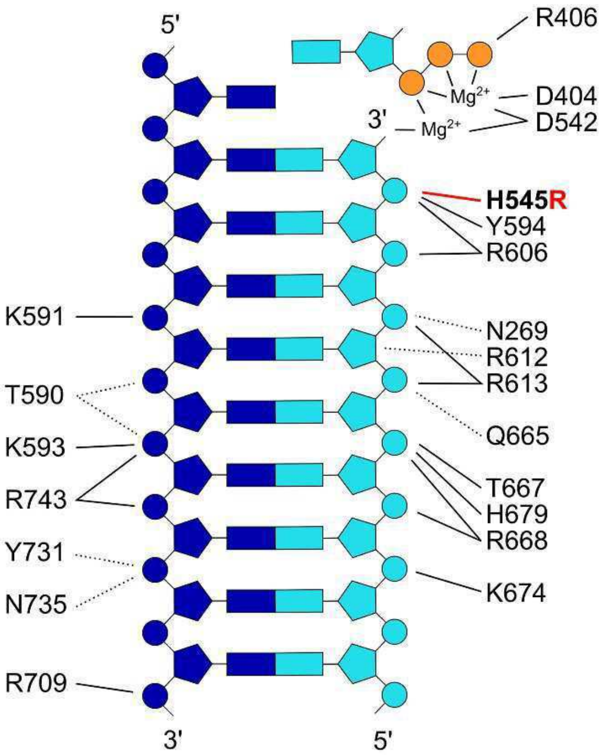


**Figure 3.8: Diagram of interactions between a primer-template DNA duplex and Therminator DNA polymerase based on a 9°N crystal structure (PDB 4K8X).**

Dark blue: template; light blue: primer; orange: phosphates of the nucleotide substrate. The primer 3’-OH and incoming nucleotide are organized by a network of amino acid residues, catalytic $Mg^{2+}$ ions, the triphosphate group and water molecules (not shown). The DNA duplex is tightly bound by the polymerase for roughly one helical turn, mostly by positively charged residues from the thumb domain. Solid lines: hydrogen bonds and ionic interactions. Dotted lines: polar contacts. The side chain of the mutated Arg (red) in the polymerase E2 mutant is positioned to make a potent additional interaction with the primer phosphate backbone.

Bands of elongation products right below the dxNT incorporation band are often observed when a small number of dxNTs are spaced by at least 3 dNTs (Figure 3.5, templates 1C, 1D, 2C, 2D, 3C and 3D), indicating that incorporation and not extension of incorporated dxNTs limits elongation. Stalling prior to dxNT incorporation would suggest poor nucleoside triphosphate binding due to the different sugar ring conformation and/or presence of a 3’-β-OH. However, when the number of incorporated dxNTs increases and their spacing decreases to one or two dNTs (Figure 3.5, templates 3A and 3B), other intermediates become more abundant, even when dNTs are incorporated directly on dNTs. In this case, hampered translocation is also becoming a significant cause of stalling due to increasingly larger helix backbone distortions and potential steric clashes. Intermediates formed at consistent distances from an incorporated dxNT could be explained by steric checkpoints where a backbone distortion is most difficult to translocate.

E2 is capable of incorporating up to four dxA if each is spaced apart by two dNTs, and still producing small amounts of full-length products, implying translocation of the distortions throughout the polymerase. Incorporation of a dxA-dNT tandem series stalled after extension of the fourth dxA, and failed to incorporate a fifth dxA. These results indicate that a double-stranded hybrid dxNA/DNA with at least four closely spaced backbone distortions can be produced. This is remarkable, given the expected strong backbone distortions in this nucleic acid. After all, when in 9 bp DNA duplexes three deoxyribose moieties are replaced by 2'-deoxyxylose, the $T_m$ becomes so low that there is no longer a melting curve for the duplex (21), meaning these strands do not hybridize in solution. Similar results of a polymerase being able to elongate a primer-template complex that is not stable/hybridized in solution were obtained by Tsai *et al*. with Therminator for DNA polymerization on a GNA template (117) and by Liu *et al*. for tPhoNA synthesis and reverse transcription by a TgoT mutant (66). It cannot be deduced from our data whether the distorted primer-template complex is translocated while bound to the polymerase by processive elongation, or whether the polymerase repeatedly dissociates and re-binds, forcing the reannealing of the primer 3' end, although advanced dissociation is expected because of the low enzyme concentration used, the absence of processivity factors and the many instances of forced stalling. In either case, the polymerase has to provide strong stabilization of a deformed duplex.

When incorporating two consecutive dxNTs, formation of the +2 product would imply incorporation of a dxNT as well as its extension with a dxNT. Stalling at +2 is seen for every template that requires either one or two initial dxNTs, which implies that the last phosphodiester group of the primer (Figure 3.8) is specifically affected by a newly incorporated and extended dxNT. We therefore hypothesize it to be a main steric checkpoint. In the crystal structure, this phosphate is in very close contact with residues H545, R606 and Y594 (Figure 3.9). H545 is packed in a hydrophobic pocket with several aromatic residues, and interacts indirectly with the backbone through hydrogen bonding with water molecules and residue Y594. In polymerase variant E2, the H545R substitution appears to tolerate larger backbone distortions, as well as to reduce misincorporation and template-independent elongation, in spite of slower initial incorporation of dxNTs compared to the wild type.

We presume that the positive charge of R545 is positioned very close to this primer phosphate group. Since R545 may undergo aromatic stacking with Y402, this could also change the interactions with other residues and water molecules near or in the active site. Because Arg has a longer side chain, offers better interaction geometry with phosphates and has a strong positive charge, it is presumed that this interaction with the most distortion-sensitive phosphate group stabilizes the +2 product by keeping the phosphate group, and thereby indirectly also the 3' hydroxyl group, closer to the conformation needed for extension, rather than directly improving specificity for dxNT binding and incorporation. The modelling of H545R suggests that it takes the place of two water molecules next to Y402, adopting their polar interactions but bringing the positive charge much closer to the phosphate group. Whereas the wild type can incorporate two consecutive dxNTs efficiently but stalls before further extension, the H545R mutant incorporates dxNTs more slowly and is able to extend them, albeit more readily with natural nucleotides.

Other possible effects of the H545R substitution are a general change in protein dynamics which shifts the equilibrium more towards the closed state or facilitating translocation, or more large scale secondary structure rearrangements that change the DNA-nucleic acid interface or active site geometry, especially in the thumb domain where most of the interactions between the DNA and protein are located. Whatever the mechanism of action, it must be noted that the H545R mutant remains an efficient, accurate and thermostable DNA polymerase.

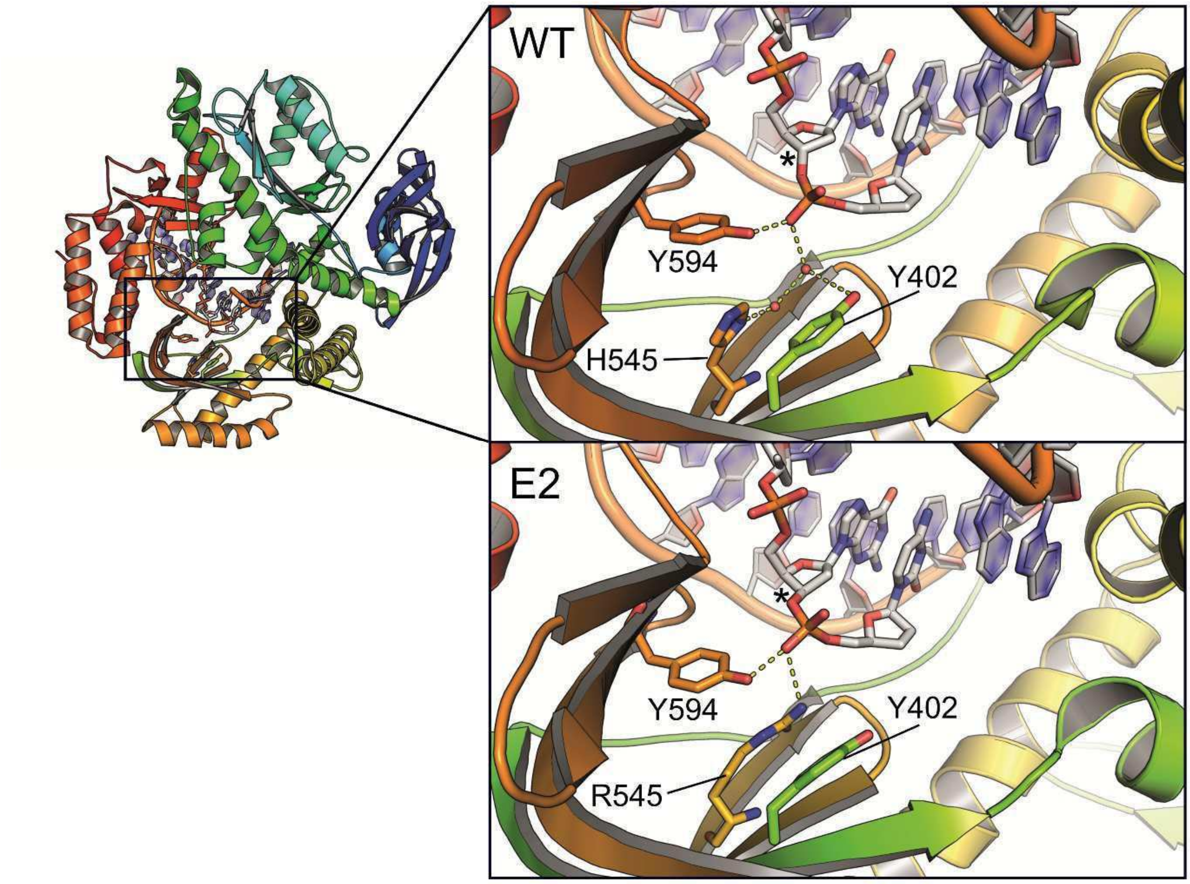


**Figure 3.9: Diagram of the interaction between the primer backbone and residue 545 near the active site of the 9°N DNA polymerase crystal (PDB 4K8X) and the modelled structure of the dxNT complex, rainbow coloured from blue N-terminal domain to red C-terminal thumb domain.**

Top: interaction between the wild type (WT) H545 and 2’-deoxyribose backbone. Bottom: interaction between the E2 mutant R545 and 2’-deoxyxylose backbone. An asterisk indicates the inverted chiral centre of the backbone sugar at position +2. This model was created using MOE (CCG, Canada) implemented with the AMBER:EHT force field. The dxNT was modelled by inverting the stereochemistry of the 3’-carbon-oxygen bond in the crystal structure of the wild type complex (PDB 4K8X). The H545R mutation was introduced followed by conjugated gradient energy-minimization of the residues in a radius of 10 Å around the R545 guanidinium head.

Within archaeal family B polymerases, residue 545 (relative to Therminator) is taken by a varying aromatic residue (His, Tyr or Phe) within an otherwise very strongly conserved active site (Figure S1). The precise aromatic residue varies both within and between clades of 51 family B polymerases over a large phylogenetic distance. This implies there is some mutational freedom at this position, as long as it does not destabilize existing interactions with the conserved residue Y594 and other nearby aromatic residues. When this residue is mutated to Lys (data not shown), polymerase activity assays show it to be still active as a polymerase, but less than the E2 clone, and only for dNTs, implying that merely a positive charge on a long flexible chain is not sufficient to elicit the improved phenotype for dxNT incorporation and elongation. To the best of our knowledge, mutation of this residue has not been reported so far in polymerases evolved to incorporate xenonucleotides.

With regard to fidelity, the H545R mutant appears more accurate than the wild type, at least under the limited tested conditions. The latter readily adds one, and eventually a second, untemplated terminal nucleotide (with a large preference for purines) to a blunt ended DNA duplex, while H545R is nearly inactive. Forced misincorporations (caused by the absence of one or more substrate nucleotide types) happen at lower relative rates in the H545R mutant, and while intermediates accumulate, it takes several consecutive misincorporations to completely stall the polymerases and prevent full elongation. Whether this lower mismatch incorporation rate in the E2 polymerase is due to an overall lower intrinsic rate of the enzyme, or mismatch incorporation rates have been lowered relative to canonical incorporation rates, cannot be deduced from this data since rates of dNT incorporation are comparatively fast and full-length products are formed within 5 min.

While misincorporation was observed by mass spectrometry in the presence of canonical dxNTPs as well (Figure S3), the spectra indicated that canonical dxNT incorporation dominates, especially for the E2 mutant and with dxATP as the modified substrate. Because a large fraction of primers reaches full elongation from templates requiring one dxNT to be incorporated, the majority of these full-length products are therefore presumed to contain a correctly templated and fully translocated dxNT rather than dxNTs acting as extension or elongation terminators. As expected, misincorporation downstream of a correctly incorporated dxNT was not observed, implying that the suboptimal orientation and pucker of the 2’-deoxyxylose sugar group slow down extension but do not negatively affect extension fidelity. Because sugar conformation and mispair geometry were suggested to be independent (118), following nucleotides can still be accurately matched to the template and properly stack on the nucleobase of the previously incorporated dxNT.

The promiscuity of the E2 mutant for other xenonucleotides was tentatively explored by comparing incorporation of arabinose- with 2’-deoxyxylose-based nucleotides. The H545R mutation caused a substantial decrease in incorporation activity for araA both on homopolymeric dT templates and templates with dTs spaced by a few natural nucleotides. This suggests that the tolerated distortions could be more specific to 2’-deoxyxylose, rather than tolerating distortions in general. Evolving towards 2’-deoxyxylose-based nucleotides and away from 2’-deoxyribose therefore presumably comes with a trade-off for arabinose too. This difference between distortion tolerances is consistent with an increased preference for dxNA (and potentially other, structurally similar XNAs), and contradicts a mechanism of merely opening the thumb and thereby avoiding steric clashes by creating a wide, nonspecific tunnel that indiscriminately extends everything it can incorporate. This could be tested further with xenonucleotides that have pseudorotation phase angles which are different from both DNA and dxNA such as FRNA or TNA (103), as well as those that have modified nucleobases rather than backbones.

## 7. Conclusion

In summary, a polymerase was isolated, using CST, that improved elongation of accurately incorporated dxNTs by a single H545R amino acid substitution. Many substitutions in published polymerases that incorporate synthetic nucleotides with sugar modified backbones are found in the thumb domain and, when present in or near the active site, are rarely in direct contact with the nucleic acid. This makes the residue at this position, and the other residues that directly or indirectly interact with it, promising candidates for further targeted mutagenesis to evolve polymerases that not only incorporate modified nucleotides but are also able to extend them, thereby aiming at efficient enzymatic transcription of DNA into long-chain dxNA.

# 4

# Recombination strategies for polymerase molecular evolution

## 1. Introduction

This chapter summarizes the efforts of recombining existing mutations to probe for synergistic epistatic interactions, as well as devising methods for a combinatorial approach on a larger scale for molecular evolution of DNA polymerases.

A small number of substitutions from the best performing clones were combined in single proteins, and new ones were introduced to analyse the mutational freedom at these positions. This however did not result in proteins that are more efficient than either parent clone.

Based on the best performing mutants so far, new libraries for the T region were also made on top of these genetic backgrounds. Polymerase activity assays were performed on randomly picked clones that underwent one round of CST.

Because recombination of promising mutations on a larger scale and in a consistent manner is not feasible by simply performing separate mutagenesis for each one, a method called Golden Gate Shuffling (119) (GGS) was used to allow the shuffling of variation between two or more regions of the polymerase. This is a method where at least nine different DNA fragments can be combined in a fixed order in a single pot reaction. Each of these fragments can be a fixed sequence or a library. The potential for recombination is very large. For example, if one would have a gene with three independently mutated regions that underwent selection, resulting in enriched libraries of 100, 500 and 1000 different mutants from initial libraries of size $10^5$, there are 1600 clones available from the original $3x10^5$. Recombining these using GGS can theoretically produce $5x10^7$ new variants, which is more than the three starting libraries combined, and all of these are already strongly biased towards functionality. This then allows new rounds of selection but from larger and higher quality libraries.

Efficient and well controlled GGS reactions could, in theory, generate large amounts of new variation after selection has strongly reduced it. This new variation will then consist of combinations of mutants that are already more likely to be beneficial. Additionally, it allows the combination of mutations that are too far apart in their sequence to be part of the same library region.

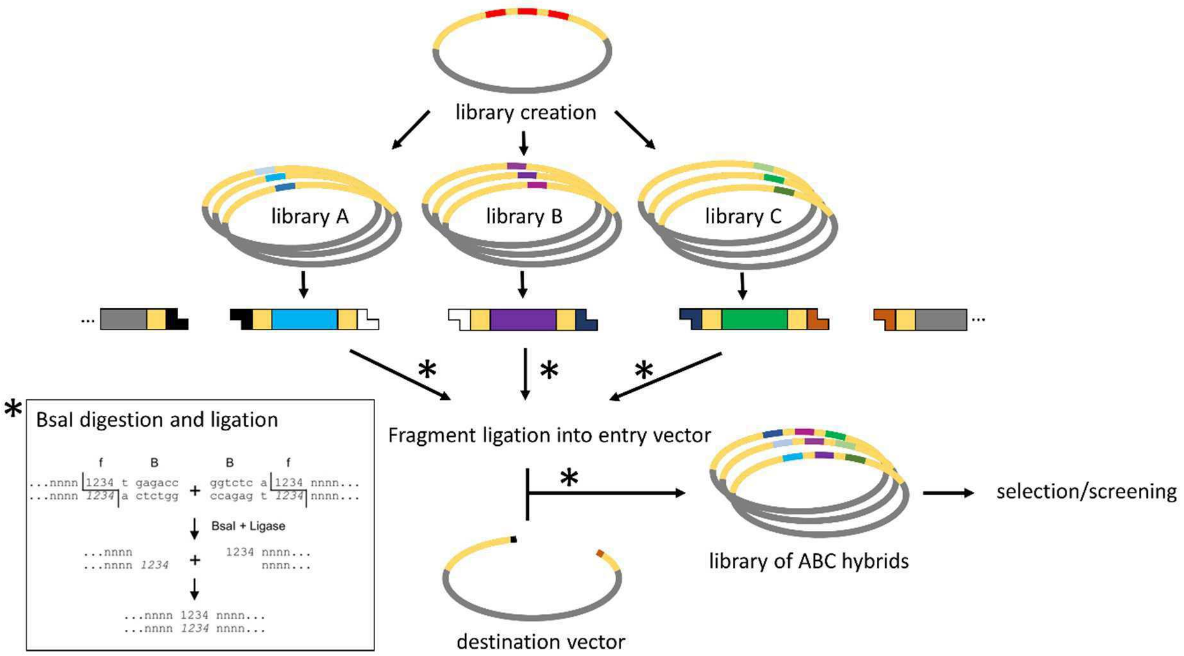


**Figure 4.1: Overview of the Golden Gate Shuffling strategy.**

PCR is performed on mutant plasmid libraries to produce linear DNA fragments flanked by two consecutive unique position marker sequences (labelled “f”) on each side. Simultaneous digestion and ligation with a first type II restriction enzyme allows ligation into an entry vector. Digestion and ligation with the second type II restriction enzyme together with the destination vector (shown with the SacB fragment already cut out) reconstitutes the complete plasmid. The resulting recombined proteins can then be selected on or screened for improved activity. Digestion/ligation insert adapted from Engler et al. (119). Copyright 2009 Public Library of Science.

Finally, a technique called VersaTile Shuffling (120) (VTS) was implemented as a proof of principle method for introducing new variation in a protein evolution experiment. The VTS technology was used in collaboration with Prof. Rob Lavigne (Laboratory of Gene Technology) and Prof. Yves Briers (Laboratory of Applied Biotechnology, Ghent University) in the context of a patent application for this VTS technology.

While GGS is limited by homology at the sites of overlap, VTS allows for recombining completely independent protein fragments in fusions, by introducing flexible, short linkers between the fragments of two amino acids long. Because homology is no longer a constraining factor, any protein fragment can be fused and shuffled in principle. However, in the case of a thermostable DNA polymerase the high reaction temperatures require that the shuffled fragments fold independently and remain stable, without hindering the folding of the rest of the enzyme as well.

This method was used to exchange a peptide loop from the bacteriophage ϕ29 DNA polymerase with the structurally homologous, but sequence-divergent, loop in DNA polymerase Therminator. This loop was chosen because it is known to be responsible for the high processivity and strand displacement activity of ϕ29 polymerase. Grafting this loop in Therminator polymerase, which has a similar fold and DNA binding location, could therefore potentially introduce higher processivity, which in turn could decrease the chance of dissociation during stalling when incorporating dxNTs.

## 2. Materials and methods

### 2.1. Quickchange mutagenesis

Primer couples for mutagenesis were designed to be complementary to each other at their 5' ends. Primers are complementary to the plasmid template, except at positions to be mutated, with at least 10 matching bases upstream and downstream of the introduced mismatch. Non-strand displacing polymerase (Pfu) was used to synthesize nicked circular strands. The parental plasmids were digested by DpnI at 37 °C for one hour to reduce template background. The nicked circular dsDNA carrying the wanted mutations was then used for transformation directly as previously described (see section 2.4).

### 2.2. Golden Gate Shuffling (GGS)

GGS fragments were designed in such a way that each library region is contained in a separate fragment, as well as the remaining non-variable part of the protein. The ends of each fragment lay in a region of the genes for the DNA polymerases Therminator and Vent where both are completely identical, as these are the position markers where the DNA fragments will be ligated. GGS fragments are therefore constrained by the sequence identity of these four nucleotides, and the two codons they are part of. An added constraint is that each Vent region library prepared by Dr. Yves Peeters was also contained in these fragments, even though these Vent and Therminator library sequences do not completely overlap. The fragment-encoded protein segments are shown in Figure 4.2, each in a different colour.

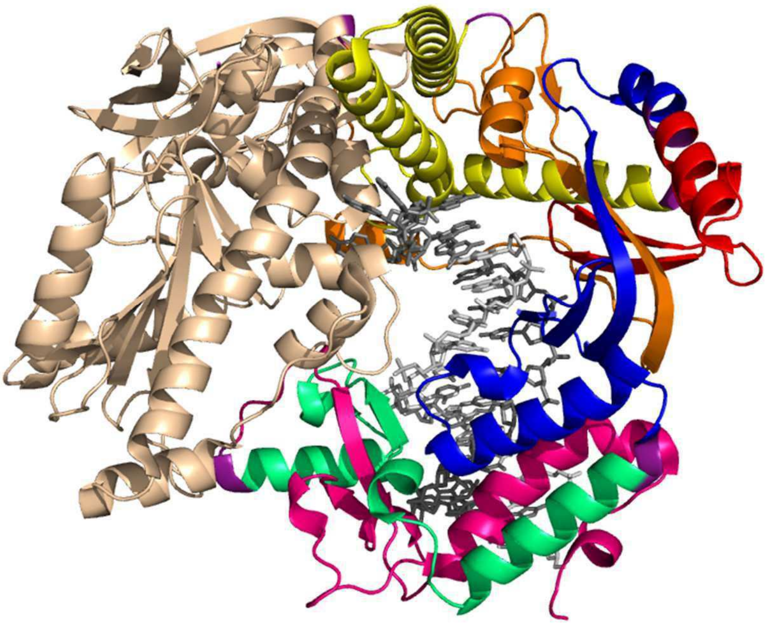


**Figure 4.2: Therminator DNA polymerase protein segments delineated for GGS.**

Protein segments corresponding to GGS fragments are shown in a different colour: beige for the non-variable part of the protein (containing the N-terminal and exonuclease domains), orange for fragment 2 (region N), yellow for fragment 3 (region F), red for fragment 4 (region A), blue for fragment 5 (region M), green for fragment 6 (region P) and pink for fragment 7 (region T). The DNA duplex is shown in grey. The two amino acids that correspond to the junction between fragments are shown in purple. The image was made in PyMol.

DNA fragments were produced by PCR, with each primer overlapping with the complementary primers of the adjacent fragment at the junctions, and containing 5’ BsaI (GGTCTC(1/5)) and SapI (GCTCTTC(1/4)) or BpiI (GAAGAC(2/6)) restriction sites (see Table 4.1, where boxes indicate recognition sequences and unique position markers are underlined). Fragments 1 and 7 required SapI instead of BpiI because of the presence of a BpiI cutting site in those DNA fragments. PCR reactions for fragments 2, 3, 4 and 6 were regular PCR reactions with an annealing temperature of 3 °C above the $T_m$ of the primer with the lower $T_m$ of the primer couple, and 35 cycles. Fragments 1, 5 and 7 did not produce a product in these conditions, for these fragments a touchdown PCR was used with an annealing temperature of 10 °C above the $T_m$ of the primer with the lower $T_m$ of the primer couple, decreasing by 0.5 °C per cycle, with 20 cycles. PCR products were purified with a PCR purification kit, were eluted in 35 µL elution buffer, of which 30 µL was loaded on an agarose gel for gel extraction. This procedure was performed multiple times in parallel for reactions with lower yields.

**Table 4.1: PCR primers for generating DNA fragments for Golden Gate Shuffling.**

| Oligonucleotide | Sequence |
|---|---|
| GGS_Fra1F | TGT GCT CTT CTA GAG GTC TCA CCA TGA TCC TGG ACA CCG AC |
| GGS_Fra1R | TGT GCT CTT CGC TTG GTC TCA GTT ACG TTT GTA AGC TTT ACG CAG |
| GGS_Fra2F | TGT GAA GAC TAT AGA GGT CTC ATA ACG AAC TGG CTC CGA AC |
| GGS_Fra2R | TGT GAA GAC TAG CTT GGT CTC TGA AGT CTT TGC AGA ATT TGT GAC C |
| GGS_Fra3F | TGT GAA GAC TAT AGA GGT CTC ACT TCC CGG GTT TCA TC |
| GGS_Fra3R | TGT GAA GAC TAG CTT GGT CTC ATC ACG GAT AAC CAT TTC GAT G |
| GGS_Fra4F | TGT GAA GAC TAT AGA GGT CTC AGT GAA CTG GAG GAG AAA TTC |
| GGS_Fra4R | TGT GAA GAC TAG CTT GGT CTC ACA GGA ATT CTT TAG CTT TTT TTT TAA CGG |
| GGS_Fra5F | TGT GAA GAC TAT AGA GGT CTC ACC TGA AAT ACA TCA ACC CGA AAC TG |
| GGS_Fra5R | TGT GAA GAC TAG CTT GGT CTC ACA ACG TCA CCG TGT TTC AGG |
| GGS_Fra6F | TGT GAA GAC TAT AGA GGT CTC AGT TGA AGA AGC TGT TCG TAT C |
| GGS_Fra6R | TGT GAA GAC TAG CTT GGT CTC AAC CAC GAG CAG CCA G |
| GGS_Fra7F | TGT GCT CTT CTA GAG GTC TCA TGG TGT TAA AAT CCG TCC GGG TAC |
| GGS_Fra7R | TGT GCT CTT CGC TTG GTC TCA TAC TTT TTA CCT TTA ACT TTC AGC C |

PCR fragments were ligated into an entry vector (pVTSE) by mixing 100 ng of the entry vector, 50 ng of the fragment insert, 10 U BpiI or SapI (depending on the fragment), 3 U T4 DNA ligase and T4 DNA ligase buffer in 20 µL reaction volumes. This reaction was incubated for 2 min at 37 °C and 3 min at 16 °C, which cycled 30 times, followed by 5 min at 50 °C and 5 min at 80 °C. The presence of a restriction enzyme that cuts outside of its own recognition sequence (removing it from the fragment) and a ligase makes ligation of the DNA fragments essentially irreversible. The ligation mixture was used directly to transform *E. coli* TOP10 cells, which were plated on LB agar plates with 5% sucrose for negative selection on entry vectors without insert that still contain the SacB gene, and 100 µg/mL ampicillin to select for plasmid presence. Colonies were then picked and grown for plasmid preparation. These pVTSE and pVTSD vectors were designed for the VersaTile Shuffling protocol (see the following section) but could be adapted for Golden Gate Shuffling.

The entry vectors then each contain one fragment, and can reconstitute the original plasmid by repeating the previous digestion-ligation reaction, but with BsaI as the restriction enzyme, a destination vector (pVTSD), carrying a SacB cassette and a kanamycin (Kan) resistance gene, instead of the entry vector pVTSE, and 50 digestion-ligation cycles instead of 30. The resulting plasmid, when reconstituted, is a slightly modified pASK plasmid with a Kan instead of Amp resistance gene, and is used to transform electrocompetent E. cloni® cells and can be used for protein expression or CST. When needed, the Therminator gene can be re-cloned seamlessly into the original pASK vector using the standard CST re-cloning primers.

### 2.3. VersaTile Shuffling (VTS)

The VersaTile Shuffling (VTS) protocol differs from GGS in the used primers (see Table 4.2, primers are labelled similarly to Table 4.1), which introduce two-amino acid linkers between the fragments to be shuffled. The entry and destination vectors (pVTSE and pVTSD) were designed for VTS but could also be used for GGS. PCR fragments were ligated into an entry vector (pVTSE) by mixing 100 ng of the entry vector, 50 ng of the Tile insert, 10 U BpiI, 3 U T4 DNA ligase and T4 DNA ligase buffer in 20 µL reaction volumes. This reaction was incubated for 2 min at 37 °C and 3 min at 16 °C, which cycled 50 times, followed by 5 min at 50 °C and 5 min at 80 °C. The ligation mixture was used directly to transform *E. coli* TOP10 cells, which were plated on LB agar plates with 5% sucrose for negative selection on entry vectors without insert that still contain the SacB gene, and 100 µg/mL ampicillin to select for plasmid presence. Colonies were then picked and grown for plasmid preparation. Dennis Grimon (Laboratory of Gene Technology) supplied pVTSE and pVSTD vectors and performed the ligation and transformation reactions for both the GGS and VTS protocols.

VTS was used to exchange most of a specific loop (29 amino acids of the TPR2 loop) from the ϕ29 DNA polymerase with a structurally homologous loop in Therminator (3 amino acids long). Tiles 1 and 2 did not yield PCR products in the fragment-generating PCR described above in the GGS protocol. Therefore tile 1 was produced using the same touchdown PCR method as described for GGS tiles 1, 5 and 7, and tile 2 was generated using 4 oligonucleotides (see Table 4.2) that hybridize with each other so the fragment could be produced by filling in the single-stranded gaps.

**Table 4.2: Primer and oligonucleotide sequences used in VTS.**

| Oligonucleotide | Sequence |
|---|---|
| VT_Ti1F | TGT GCT CTT CTA GAG GTC TCA CCA TGA TCC TGG ACA CCG AC |
| VT_Ti1R | TGT GCT CTT CGC TTG GTC TCA GCA CCA GCG TAA CCG TAG TAA C |
| VTS_oligo_1 | TGT GAA GAC TAT AGA GGT CTC GGT GCT CCG GAC GTG ACC GGG AAG |
| VTS_oligo_2 | AAG CCC AAA GCA CCG TTT TCC TTC AGG TAC GGA ACC TTC CCG GTC ACG TCC |
| VTS_oligo_3 | GGT GCT TTG GGC TTT CGT TTG GGT GAG GAA GAG ACT AAG GAT CCG GTG GC |
| VTS_oligo_4 | TGT GAA GAC TAG CTT GGT CTC GCC TGC CAC CGG ATC CTT AGT C |
| VT_Ti3F | TGT GCT CTT CTA GAG GTC TCG CAG GCT GGT ACT GCA AAG AAT GCG CTG |
| VT_Ti3R | TGT GCT CTT CGC TTG GTC TCA TAC TTT TTA CCT TTA ACT TTC AGC C |

The four oligo's overlap with 16, 14 and 18 bp respectively. For this reaction, 15 µL of each oligo (8 µM) was mixed to be equimolar, and hybridized by incubating for 3 min at 90 °C and slow cooling to room temperature. 10 µL of this mixture was added to 5 µL of oligos 1 and 4 (20 µM), and Phusion DNA polymerase (2U), Phusion buffer, and dNTPs (0.2 mM) to a reaction volume of 50 µL. This reaction was then incubated for a standard Phusion PCR reaction (30 cycles, annealing temperature of 58 °C). Plasmids were reconstituted by mixing 100 ng pVTSD, 50 ng of each tile-carrying pVTSE, 1U U BsaI, 3 U T4 DNA ligase and T4 ligase buffer in reaction volumes of 20 µL. Reactions are then cycled 50 times at 37 °C for 2 min and 16 °C for 3 min. The reaction is then incubated at 50 °C for 5 min and at 80 °C for 5 min. The ligation mix is used to transform *E. coli* TOP 10 cells, plated on LB plates with 5% sucrose and Kan (50 µg/mL). Colonies were then picked and grown for plasmid preparation.

## 3. Results

### 3.1. Combining selected mutations with directed mutagenesis

The two clones that had performed best so far were E2 and G8 (see Chapter 3). E2 only carries the H545R substitution, G8 has R612H and E617R. Because residues in positions 545 and 612 are in close or direct contact with the nucleic acid, and selected mutations at positions 545, 612 and 617 all involved substitutions between or towards large, positively charged residues, several combinations of these were made, including substitution to the as of yet unsampled residue Lys. At these three positions, the wild type has amino acids His, Arg and Glu and for ease of comparison it is referred to as “HRE”. Correspondingly, clone E2 is then represented by RRE, and G8 by HHR. Mutants that were generated by mutagenesis are KRE, HKE, HHE, KHR, RKE and RHE. A preliminary polymerase activity assay ELISA performed on these new mutants (same conditions as the assay in Figure 2.6C with templates 1T and 2T) showed that none of them outperformed clone E2.

### 3.2. Combining selected mutations with GGS

As a proof of principle of recombining mutant libraries, GGS was used to combine two small new libraries, consisting of the four best performing clones of region A (clones E2, E6, H4 and G1) and region M (clones G4, G8, H10 and iH4). Initially, this was attempted with a one-pot reaction with all fragments. Every fragments, except for 4 and 5 (which correspond to regions A and M, respectively), contained the wild type sequence. However, this reaction produced a large fraction of deletion mutants, caused by preferential ligation of digested fragment 5 with the destination vector, resulting in the absence of fragments 1 to 4 from the final reconstituted plasmid. To avoid this problem, the digestion/ligation reaction was split so that the destination vector and fragments 1 to 3 were present in one reaction, and 4 to 7 in the other, making ligation of destination vector to fragment 5 ligation impossible, in the hope that the correct ligation would then be favoured. The reactions were then combined and incubation continued. However, this did not yield ligation products that could be successfully used for transformation. For this reason, a new “superfragment” was created. This fragment combines the fragments 1, 2, 3 and 7 in one larger destination vector so that only four fragments instead of seven need to be ligated, and the problematic destination vector to fragment 5 ligation cannot occur. After direct transformation with the ligation product from this new reaction, 82 clones were randomly picked for analysis. Sequencing of 42 of these clones showed 39 correctly recombined plasmids (the rest missing fragment 5 and/or 6), which represents a drastic increase in ligation efficiency.

Since the combination of two small libraries of four clones each can result in 16 combinations, 42 clones were sequenced to isolate each combination. Almost all of them were found, with the exception of the combination of clones E6/H10 and G1/H10.

### 3.3. Introduction of new variation in selected mutants

Two new T region libraries were made, one using solely the E2 clone as the template, the other an equal mix of four genotypes: the G8 clone, the combination of E2 and G4, the combination of E2 and the R612H substitution found in clone G8, and a clone identical to the former but with R612K (to sample the third positively charged amino acid at this position).

ELISA performed on 60 randomly picked clones from the E2 background library and 32 from the mixed background library resulted in three clones that seemed to incorporate a single dxTTP better than the wildtype. These clones were named A1, E11 and C11 and their mutations are shown in Table 4.3. The genetic background of clone A1 is that of clone E2, that of E11 is E2 and G4 combined, that of C11 is that of E2 with one substitution present in the G8 genotype. New substitution mutations in parentheses are side mutations in the non-variable part of the Therminator gene.

However, repeated ELISA using purified protein instead of lysates from these three clones was ambiguous and could not be interpreted with a good degree of certainty. Some substitutions were also present in the non-variable part of the Therminator gene, which means that these were not present during CST but were introduced during or after re-cloning following CST, but if these exert a phenotypic effect, it was then present in the ELISA. Due to time constraints these were not further examined.

**Table 4.3: Substitutions in clones with promising activity in ELISA on randomly picked clones from two new T region libraries after one round of CST.**

| clone | genetic background | new substitutions |
|---|---|---|
| A1 | H545R | (P36T), R709L, D712H |
| E11 | H545R, K593N, V596A, I610M, S616P | (A117T, T349I, Y362H), I710T, I733N, N735K |
| C11 | H545R, R612H | I733F |

### 3.4. VTS of Therminator and ϕ29 DNA polymerases

VersaTile shuffling was used to create a chimeric polymerase, where a short loop from Therminator is exchanged for the much longer terminal protein region 2 (TPR2) loop from the ϕ29 DNA polymerase. This loop is known to be necessary for strand displacement activity and strong processivity and is present in all protein-primed DNA polymerases (121). It does so by forming a very narrow tunnel through which only a ssDNA strand can pass, thus coupling polymerization to strand displacement, and tethering the DNA to the polymerase. The ϕ29 DNA polymerase has also been shown to be able to produce several XNAs (18). Although no appreciable sequence identity exists between the Therminator and ϕ29 polymerases, they are structurally highly similar, especially the relative orientation of the two alpha helices that are connected by this loop (see Figure 4.3). Since this loop is known to interact with DNA, and the single-stranded part of the DNA template enters the polymerase at the same location, it is possible (although not necessarily likely) that this processivity can be (partially) transferred by grafting the TPR2 loop into Therminator.

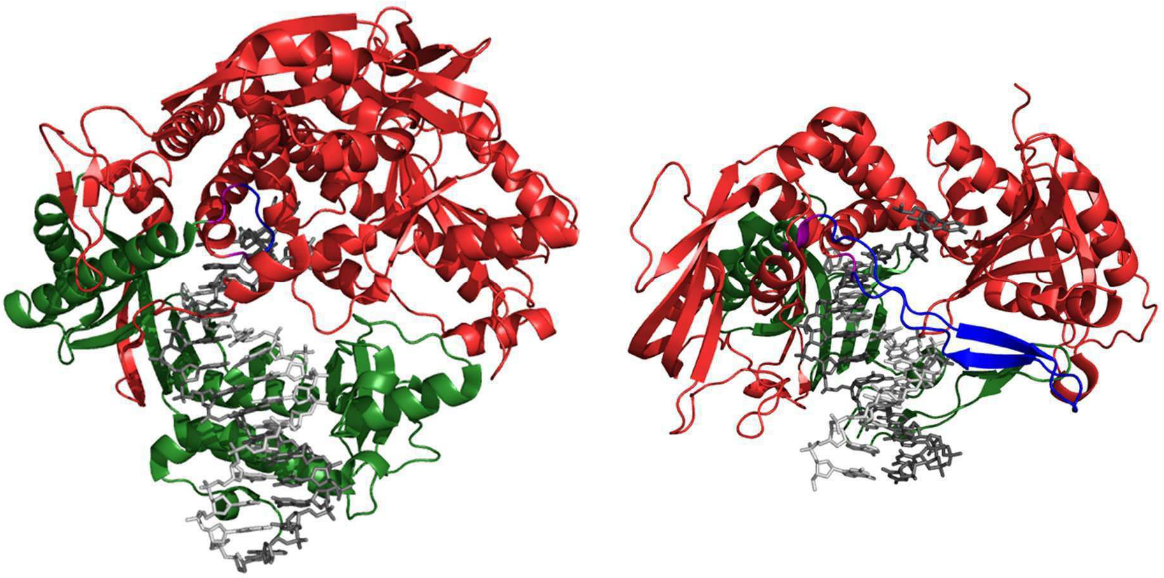


**Figure 4.3: Structures of DNA polymerase 9°N-7 (left, PDB 4K8X) and DNA polymerase ϕ29 (right, PDB 2PYJ).**

Tile 1 is coloured in red, tile 2 in blue and tile 3 in green. Position markers connecting them are purple. The resulting chimeric polymerase combines tiles 1 and 3 from Therminator and tile 2 from ϕ29.

Tiles 1 and 3 (red and green in Figure 4.3) are relatively large, since tile 1 consists of the N-terminal, exonuclease and fingers domains and half of the palm, while tile 3 consists of the rest of the palm and the thumb domain. On the contrary, tile 2 is very small and PCR-based tile generation of tile 2 proved to be difficult to generate and purify. Therefore, it was produced by annealing 4 shorter, partially overlapping oligonucleotides that span the sequence together, which are then filled in and amplified with the standard tile 2 primers. The three tiles were then recombined using the standard VTS protocol, and after transformation two sequenced clones were found that contain the wanted chimeric gene sequence.

Expression of the ϕ29-Therminator chimeric protein did show in SDS-PAGE that it could be expressed and had the expected mass increase. It was also purified and tested in extension reactions on a polyA template with different substrates (see Figure 4.4). Here it showed similar reactions to wild type Therminator and several mutants. Most primer is extended with one dxT, and misincorporation happens quickly as well (most primer is extended with one after 5 min). However, there is no visible trace of further incorporation. Furthermore, while 45 min seems sufficient to extend the primer against the polyA template completely, after 5 min there is a nearly complete set of reaction products of +3 and higher. This indicates that the normal DNA polymerase activity has drastically slowed, since these reactions are always seen to be completed before 5 min using all necessary natural nucleotides and the wild type polymerase. This shows that the chimeric protein can fold sufficiently well to allow for polymerization, but is a lot slower.

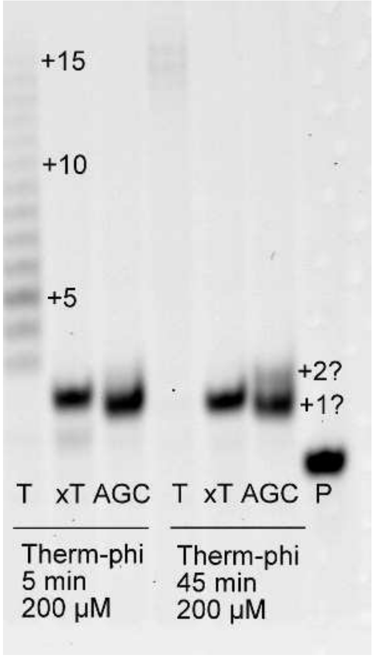


**Figure 4.4: PAGE of extension reaction on a polyA DNA template with the Therminator-ϕ29 fusion after 5 min (left) and 45 min (right) with nucleotide substrate concentrations of 200 µM.**

These are dTTP (“T”), dxTTP (“xT”) or a mix of dATP, dGTP and dCTP (“AGC”). “P” is unextended primer.

## 4. Discussion

Different strategies to recombine selected mutations in Therminator DNA polymerase were explored. The most straightforward of these was site-directed mutagenesis to make different new, clones individually. While this method of generating mutants is highly efficient and offers the most freedom for creating one or more precise substitutions, it is limited in throughput. All sequences also have to be either known or designed. Its major usefulness lies in investigating individual mutations within a well performing clone, combining them to answer questions about specific combinations, and allowing further examination of the nearby sequence space in more detail. This is the case when for example substitutions between positively charged residues are systematically targeted to sample histidine, arginine and lysine.

To obtain a larger throughput, GGS can be used to combine existing populations of mutants, where the newly generated library maximum size is the product of both populations. This means that fragments can be made from libraries using a standard protocol, even without knowing which mutations are present, which can then be combined with other libraries. Since large library sizes can quickly become rate limiting, merely starting with a library as large as possible, and then selecting for improved activity until variation decreases, might not be the most efficient selection strategy. The following shuffling-based strategy is therefore suggested. Several libraries can be made and subjected to a low amount of selection rounds in parallel. The resulting populations of molecules that should be enriched in more active variants can be used as a template to create GGS fragments. Even selected libraries that are reduced in size to for example $10^2$ to $10^3$ different clones can theoretically give rise to combined libraries with sizes of $10^4$ (two libraries with 100 clones) or even $10^9$ (three libraries with 1000 clones each). This way, a standardized recombination step can create a library much larger than the starting library, with much higher local mutation rates already biased to more active substitutions. To generate these combined mutants from a standard library would imply a much larger sequence space to investigate. This means a representative library would have such a large size that the selection experiment becomes the population size limit, causing an immediate loss of genetic variation through a population bottleneck.

Once specific clones are isolated that are sufficiently characterized and found to increase the desired activity without unwanted side effects, these can either be shuffled with library fragments in other regions, or used as templates for the generation of new starting libraries. The latter was used in the specific case of clone E2, on the genetic background of which new mutations in the thumb domain (region T) were generated.

An important limit these shuffling methods have, however, is combinations of mutations within library regions. All mutations within the same region will be retained in the same fragments, and be inherited in complete genetic linkage. This precludes certain combinations to be made, namely those within the same region. And it is precisely those residue positions that are likely to be subject to epistasis (122). It is therefore expected that clones that are the result of recombination of regions that lie in different parts of the protein structure will more likely have additive effects. The major exceptions to this presumed effect are region A, and parts of both region N and M. Some amino acid residues in these locations make direct contacts with residues in these other regions, and are more likely to also show negative or positive epistasis. Selection that follows recombination can remove negative epistatic mutations that in isolation were beneficial, but mutations that are neutral or deleterious in isolation will already have been selected against and might not be combined if they are already efficiently purged from the population. Therefore, recovering mutation combinations that exhibit this type of epistasis likely requires a limited amount of selection rounds with a low selective pressure on the starting libraries, followed by recombination. Mutations in regions that are more likely to have additive phenotypic effects can undergo more stringent selection before recombination. This would be for example between regions F and T or P, since these are functionally and spatially distinct.

The design of all Golden Gate fragments has been restricted to shuffling fragments that are identical at the amino acid residues coded in part by the position markers. This is trivial for shuffling fragments of a single protein species, but not in the case of shuffling two different protein species. Still, it was possible to design fragements so that exchangeable fragments can be produced for both DNA polymerases Therminator and Vent. This was done in such a way that previously designed Vent libraries (made by Dr. Yves Peeters) all have an equivalent Therminator library, even though these libraries do not perfectly overlap. Both polymerases also only share 78% sequence identity at the protein level, and their synthetic genes were independently codon-optimized, leaving little degrees of freedom for fragment and primer design. Combinations of Therminator and Vent libraries would then not just import selected mutations, but also all amino acid differences between both polymerases. This means that from the 7 wild type fragments from both proteins 126 Vent-Therminator hybrids can be made, most of which are likely to fold correctly given the strong structural and protein sequence similarity.

The other recombination-based technique of VersaTile shuffling is not limited by homology of the source genes and therefore has more freedom to combine sequences. It does this by introducing short linkers encoding two small amino acids between the fragments. However, recombining and shuffling fragments in any order from any source is not recommended. If the final protein product has to have a stable fold, the best fragments are those that can independently fold and function correctly as protein domains without much interference from the rest of the protein. Such domains could then even be shuffled in a different relative order, such as in artificial endolysins (123, 124). In the case of thermophilic DNA polymerases, however, most domains have many structural contacts with other domains, and removing or changing a domain will likely destroy the entire function and topology of the protein as a DNA polymerase. It is for this reason that in this study a small “isolated” surface loop was chosen to be exchanged with the corresponding larger TPR2 loop from ϕ29 DNA polymerase. The assumption behind this strategy was that such an insertion would not interfere with the fold of the rest of the protein. Since ϕ29 is mesophilic, it was still an open question whether the grafted TPR2 loop would interfere with the high temperature stability of Therminator (70 °C to nearly 100 °C).

Given that the small equivalent loop of Therminator is partly occluded by a mostly unstructured part of the peptide backbone of the palm domain, whereas the TPR2 loop in ϕ29 is nearly completely solvent exposed, it is possible that this unusually large insertion interacts with other domains, most likely the palm. This is not that surprising since there is no *a priori* reason for the grafted TPR2 loop to have specific interactions with the Therminator protein surface, and no guarantee that it will assume the same narrow, extended fold that it does in DNA polymerase ϕ29, especially at the elevated temperatures that thermophilic family B polymerases are active at. An example of stable loop grafting has been the introduction of a short peptide sequence with an antibody binding site in a relatively unstructured, short hairpin loop of a target CD81 protein (125), which was stabilized by disulphide bridges introduced by targeted mutagenesis. The localization, motion and degrees of freedom of the TPR2 loop on Therminator could be similarly reduced, although a disulphide bridge can make the connection irreversible. While this would increase thermodynamic stability and potentially force the loop to produce the tunnel-like hole through which the DNA template can thread during polymerization, a strong but reversible interaction might be advisable so that the loop can move between a closed and open state that is biased towards the closed state during polymerization.

Still, this experiment is considered as a proof-of-concept that the VTS technique can be used to insert even a very small tile into a DNA polymerase while still remaining functional (albeit at a much slower rate). This information opens the door to another type of mutagenesis. There is after all no reason why such an insertion has to have a specific sequence. If during tile production one or more oligonucleotides are partially or fully randomized, a library of different insertions can be generated. Since the crystal structure of this DNA polymerase is known, regions that are relatively independent from the rest of the protein could be targeted. For example several loops in or near regions P and T could be fully randomized and varied in length. The argument that this strategy might be useful for the evolution of a polymerase dedicated to a xenonucleic acid with drastically changed backbone geometry is that in this case, single substitutions in a relatively constant environment are unlikely to tolerate or promote drastic changes in substrate geometry of the negatively charged backbone of different nucleic acids. Both DNA and dxNA have similar electric charges, so it is likely that the groove with which they are bound and through which they are translocated will require positively charged residues as well as those that are capable of hydrogen bonding. The topology of their backbones is very different though, that of dxNA being much more extended and possibly left handed (20, 22). Rather than changing the biochemical properties of the protein surface, it is the topology of this cleft that needs to change. It is therefore proposed that exchanging specific parts of the thumb domain for libraries of new structures is a useful alternative strategy to evolve a polymerase that needs to recognize such a fundamentally different nucleic acid backbone geometry. These inserted fragments could be completely random, or biased towards codons that code for residues that are similar to those already found in those domains, for example with a relatively high amount of His, Lys and Arg.

# 5

# Discussion and perspectives

## 1. Introduction

In this chapter, the methodology and results are discussed in the context in the state-of-the-art literature. Improvements to the methods are reflected upon, and follow-up experiments and development are outlined.

The development of a polymerase capable of double-stranded dxNA synthesis is the ultimate goal of this project, of which the results in this thesis represent the first steps. The first aim is the evolution of a DNA-dependent mixed DNA/dxNA polymerase that is capable of incorporating and extending multiple dxNT in a growing DNA strand. Such a polymerase, which has now been obtained, should then be further modified, either by design or continued molecular evolution, to produce longer, and eventually fully substituted dxNA single-strands from a DNA template. This goalpost will represent a polymerase that can "transcribe" DNA sequences to dxNA sequences, and would allow for enzymatic production of dxNA strands in an accurate and sequence-specific manner, bypassing the need for chemical synthesis of dxNA oligonucleotides. Eventually, a polymerase that can use dxNA as a template to replicate but can no longer read or write DNA or RNA would be a truly informationally orthogonal XNA polymerase. Such a polymerase is a necessary first requirement for an XNA episome.

## 2. Mutagenesis strategies

### 2.1. Population size and sequence space exploration

The first step in any molecular selection experiment is the generation of variation on which can be selected. Variants with new or improved functions can then be found by screening the survivors afterwards. However, this can only work reliably when those new, functional sequences are sufficiently near the starting sequence within sequence space (126). For this reason, it is important to choose a good starting point to evolve, as well as to introduce mutations at an appropriate rate, so that traversing the fitness landscape is possible (69). If hypothetical proteins with the desired activity are so different from the starting point that a very high number of mutations are required to bridge the gap, the required space to search drastically increases. This in turn increases the necessary population size to plausibly hit those sequences, since sequence space increases exponentially depending on sequence length ($n^4$ for nucleic acids of length n). This quickly makes the necessary population sizes so large that practical limitations become a population size bottleneck (127). In this case it can become realistic that, even if one or more desired mutants potentially exist in the search space, none are present in the created library. A number of mutations that is too large can also cause the evolving protein to cross a stability threshold where negative epistasis between slightly deleterious mutations drastically increases the magnitude of their negative fitness effect (128), making the protein less robust to mutations than initially thought.

On the other hand, creating less variation decreases the search space, and when the search space is of the same order of magnitude as the smallest population bottleneck in the selection steps, all possible mutants can be present in the population. Finding the active ones is then only a matter of time, and with sufficiently low population sizes, screening all of them individually can even remove the need for a separate selection step.

In the case of DNA polymerase evolution for the production of xenonucleic acids, mutants with an increased activity usually have between one and ten substitutions. While it is possible that many of these mutations are neutral, it does show that thermostable polymerases can tolerate large numbers of substitutions (77, 129). This means that mutation rates can be quite high, even concentrated locally in specific regions of the polymerase, and still mutants can be recovered that are not just functional, but more active for these new catalytic activities. This could mean that the adaptive fitness peak where DNA polymerases reside in sequence space has sufficiently flat and high edges that can be explored. The question that still remains is whether these DNA polymerase peaks are connected with XNA polymerase peaks, in which case viable intermediates allow for gradual evolution over a selective gradient, or large fitness valleys exist that need to be crossed with more drastic mutation. The strategy followed in this thesis is based mostly on the former, at least for the first explorations of evolving a dxNA polymerase from a DNA polymerase. However, given the large differences in geometry between DNA and dxNA (20, 22), the binding of a dxNA template, and the formation of a dxNA double helix, it is highly likely that at some point a drastic change is required that cannot simply be bridged by a handful of amino acid substitutions along a largely conserved peptide backbone. In any case, a large genetic difference between a DNAP and XNAP that is difficult to evolve is still desirable, since one of the main goals of an informationally orthogonal system is containment. If a dxNA polymerase was simply a few steps away from naturally occurring proteins along an easily climbable fitness peak, such a nucleic acid would no longer be appropriate from a biosafety standpoint. Any XNA polymerase that produces an orthogonal XNA should therefore be hard to evolve by default.

### 2.2. Targeted randomization

The type of mutation and mutation rate achieved in this study was similar to that in the work of Pinheiro *et al*. (17). It is assumed that mutations that change the interaction with a bound nucleic acid or nucleotide will usually be in close contact with them. A library that is focused on these regions will then be biased towards more functionally changed mutants (130). For this reason, we have chosen six regions that are more or less functionally distinct, in the fingers, palm and thumb domains. While mutations outside these regions can certainly have relevant effects, such as the terminator mutation A485L (93), a compromise was made between library quality and completeness. These target regions were partially randomized by creating a 9% chance of mutation of each base. As shown by sequencing, several clones were discovered with premature stop codons, or frameshift mutations. This type of mutation will, in most cases, render the protein functionless as it amounts to deleting almost half the protein, and most of the core functional domains. The presence of such truncation mutants in a library is thought to be not too problematic since a total loss of function should lead to strong counter selection and a fast removal from the population, and they are relatively rare to begin with.

Most of the mutated clones contained between one and five substitutions, with frequencies that are very similar to the expected ones. Promising selected mutants did not show a consistent trend towards low or high amounts of mutations, and in clones carrying multiple mutations, following mutagenesis experiments are needed to show which of them are necessary and sufficient for the new phenotype. Based on these results, it cannot be determined what an optimal mutation rate in these regions would be.

### 2.3. Recombination through shuffling

It still unclear how general the benefits of recombination in molecular evolution experiments are (131), since most of these experiments are focused on getting improved results rather than investigating the theoretical aspects of the effects of recombination on traversing fitness landscapes. However, several studies have found that using shuffling techniques, deleterious mutations can be more efficiently purged from the population (132), competing high fitness mutants can be less likely to drive each other to extinction (133) and recombined, related proteins are more likely to be functional than random mutants with a comparable genetic distance (134), for example the exchange of a part of the fingers domain from the archaeal Pfu DNAP with the homologous eukaryotic sequence from pol ζ that acquired RT activity (135). We have used recombination techniques to allow these benefits in the evolution of Therminator DNA polymerase.

The most straightforward recombination is that of mutations that are found in different library regions with potentially additive effects. This gives access to parts of sequence space that are unreachable by the region libraries on their own, potentially creating new polymerases that are equal or more than the sum of their parts. Several of these combinations were preliminary tested, mainly by combining substitutions in the minor groove-binding region with the active site substitution H545R from clone E2. However, combining known mutations that were found in screening is limited to those mutations that already have an effect individually. Combinations between mutations in different regions that have epistatic effects that are "invisible" to selection on their own can be sampled by shuffling different libraries altogether. The most pragmatic time point to do this is after few initial selection rounds, so that inactive variants have been largely removed, but (near) neutral, potentially epistatic mutations are not yet outcompeted by variants with high fitness. This can be especially useful if the library size has decreased sufficiently, so a new library is then created on which selection needs to be performed again (since the most active clones are present in much lower relative amounts), but which spans a drastically larger sequence space.

We increased the potential for shuffling to not just combine mutations within Therminator libraries but also to combine fragments of Vent polymerase libraries using Golden Gate Shuffling. If specific mutations were to be found in Vent polymerase, these could then be imported into Therminator polymerase in isolation, or within their normal protein “context” by transferring them with the entire Vent region.

This strategy was taken even further using VersaTile shuffling, where a proof-of-concept experiment was performed by exchanging a peptide loop that was completely different in both sequence and length from the unrelated, mesophilic ϕ29 DNA polymerase. The resulting chimeric polymerase (with the ϕ29 loop grafted on Therminator) retained DNA polymerase activity and was also still able to incorporate one dxNTP, but this activity was much lower compared to the wildtype enzyme. This does show that it is possible to insert a non-conserved amino acid sequence that is not adapted for thermostability, without interfering so severely with folding and activity that it becomes inactive. Identifying and shuffling corresponding loops or regions that could tolerate such structural changes is suggested as a new mutagenesis strategy that is similar to humanization of antibodies (136) but not bound to similarity of the donor and acceptor proteins. It is hypothesized to be most useful in regions that require a more drastic change in protein surface topology, in this case mainly in the thumb domain at the protein and DNA backbone interface. Such insertions could be fully randomized sequences of fixed or variable length, or rationally designed structural elements that are partially randomized but will fold independently and (somewhat) predictably. These elements could be designed based on existing structural elements, such as for example four-helix bundles (137). Structures such as these are preferable when they are already quite thermostable, given the high reaction temperatures of hyperthermophilic archaeal DNA polymerases.

## 3. Library quality

### 3.1. Library size

The molecular population size varies considerably during the workflow of a molecular evolution experiment. Since the maximal genetic variation is limited by the minimal population size, genetic bottlenecks need to be properly controlled to allow for a maximally efficient throughput. Everything begins with the primary library size where variation is generated. The total possible sequence diversity for the six Therminator regions is gigantic, and the synthetic spiked primers can only contain a fraction of the total variation. Only a small fraction of these primers was used in turn to perform the mutagenic iPCR to generate plasmid libraries. Nevertheless, the amount of created unique mutant plasmids (assuming all primers are extended in a 100% yield) reached $6.10^{13}$ molecules. It is the transformation step that limited the primary library size, and this ranged around $10^5$-$10^6$ clones. Most subsequent manipulation steps like emulsifying cells in CST or re-cloning recovered plasmids created molecular population sizes at least one order of magnitude larger. Because most variants are present in multiple redundant copies even before selection has occurred, the effect of genetic drift on rare clones decreases. The longer a population has been under selection, the less the impact of technological limitations such as re-cloning efficiency will become. The main population bottleneck in this work has been the physical plating of transformed cells. Population sizes are determined by the practical density of colonies on solid growth medium and the somewhat variable efficiency of transformation. For this reason, libraries were always directly expressed from the transformation strain, which is an electrocompetent strain with a high transformation efficiency, instead of re-cloning them in a different expression strain.

### 3.2. Genetic parasites

In an ideal evolution experiment, the variants that are enriched and start to dominate the population are those that have the desired traits, in this case improved affinity, specificity and incorporation efficiency for dxNTPs. However, the selective pressure the population is under will usually be merely a proxy for the required function. A variant that, for any reason, is able to outcompete other variants under the experimental environment will be selected for. If the selective pressure for these parasitic variants is larger than the selection for mutants with a desirable activity, such genetic parasites can start to dominate the population and outcompete genuinely useful mutants. For example, hypothetical error prone mutants may be able to bypass templating bases that should require incorporation of a dxNTP, and instead misincorporate a non-Watson-Crick base-pairing nucleotide. An even more direct way of bypassing dxNT-requiring templates is merely adding biotinylated dC nucleotides to the primer 3′ end in a template-independent manner. Mutants with such activities are able to tag plasmids that can be captured by affinity chromatography, and will be selected together with mutants producing biotinylated plasmid-primer complexes by incorporating dxNTPs. While later characterization can easily distinguish such "cheaters" from genuine dxNTP incorporation clones with proper controls (like the use of nucleotide substrate pools that force misincorporation by lacking one or more dNTPs, or mass spectrometry), there is the risk of competition and near complete extinction of mutants with the desired phenotype. Clones with the aforementioned activities were however not discovered, since clones that gave decreased or no signals in ELISA were not further characterized in detail. This screening step thereby effectively selects strongly against hypothetical highly error-prone variants. Some of the clones that showed no improved ELISA activity in dxNT incorporation but still scored high in misincorporation controls could have been mutants that can read through dxNT-requiring templates by misincorporation, but these also were not taken further for characterization and mutagenesis.

Another type of genetic parasite, which was encountered in these CST selection experiments, are deletion mutants. Rather than bypassing the envisioned section criteria during the CST reaction itself, they can take advantage of differences in PCR cloning efficiency in between CST selection rounds. Such deletion mutants usually contained the plasmid backbone, including the origin of replication, the resistance genes that allow them to survive transformation and plating on plates with the appropriate antibiotic, and part of the Therminator gene. These deletions were observed by agarose gel electrophoresis of plasmid libraries and after their re-cloning. Deletions were usually of different lengths (in the order of 400 to 1700 bp), so it was probably not a single, consistent PCR or ligation artefact, but they usually corresponded to a deletion of the mutagenized part of Therminator, amounting to the C-terminal half of the protein that contains all the library regions. Such a deletion completely destroys polymerase activity of the resulting truncated proteins which therefore cannot have tagging activity during CST reactions. However, if these artefacts occur consistently during each round of re-cloning, and if pre-existing and new deletion mutants have a higher transformation efficiency, ligation and/or PCR amplification efficiencies during re-cloning due to their smaller size, they can start dominating the population, as was seen after as few as two CST rounds in some experiments (see Figure 2.5). However, the re-cloning procedure offers several points where DNA fragments can be purified with gel extraction, which typically lowers the yield but allows a selection step for clones of the correct length. This means that when the library size is small enough to be sufficiently redundant, additional gel extraction steps can drastically improve the quality of the library under selection by removing a majority of deletion mutants. These gel extractions would naturally follow the steps where most of these deletion mutants are formed.

## 4. CST as a molecular selection tool for polymerases

### 4.1. Selective pressures during CST

In advance of CST and polymerase characterization, a set of CST primers was designed based on the known promiscuous activity. Knowing that extension of one dxNT and consecutive incorporation of up to two dNTs was possible, these primers were designed as a gradient of hurdles, starting with primers requiring a single dxNT incorporation with very little dNTs before or after, continuing with those that require more dNT elongation, those that require two slightly spaced or consecutive dxNTs, those that require both of these, and finally CST primers with the highest selective pressures: up to three or more spaced or consecutive dxNTs. Based on the results of selection and screening, following selection rounds could then be ramped up very gradually to have a pressure that is high enough but not so high that even the best performing mutants would have difficulty extending the primer. This was done for multiple libraries, with primers requiring the elongation of two consecutive dxNTs. However, these libraries were shown to consist of a very large proportion of deletion mutants and were not screened further. The libraries that yielded the clones that are analyzed in Chapters 2 and 3 therefore underwent one (library M) or two (library A) gradations of low pressure.

### 4.2. "Cheating" polymerases and selection efficiency

CST is the compartmentalization method of choice for evolving polymerase activities where the existing promiscuous activity is very low and when no reverse transcriptase exists. The related selection technique of compartmentalized self-replication (CSR) requires the production of nucleic acids at least a short patch of several dozen bp long (enough to contain the mutagenized region of the polymerase gene), and the resulting XNA fragments need to be transcribed back into DNA for re-cloning and the next round of selection. Obviously, this self-replication presents an impossibly high selective pressure for a polymerase that can hardly incorporate two consecutive dxNTs, and without a polymerase that can read the eventually produced dxNA, the evolutionary cycle hits a dead end. Only when these two conditions are met, namely the presence of an XNA-to-DNA reverse transcriptase and sufficiently long XNA synthesis, can CSR be used. CST is therefore a necessary first step, but should be followed up by short-patch CSR, and eventually CSR of larger fragments, as soon as possible. This is because the selective environment in CSR is a much more accurate environment with a selective pressure that strongly corresponds to the wanted activity. The reaction in CST presents a more indirect selection, which leads to alternative survival strategies that can also be selected upon, but which do not necessarily produce enzymes that are accurate XNA polymerases. As mentioned above, CST selects specifically on tagged plasmids. Every activity that adds biotinylated dC to the primer is selected for and there is no control for fidelity. This is in contrast to CSR, where error-prone variants will encounter an error catastrophe since they are directly responsible for their own replication. The fact that the wild type Therminator enzyme already has misincorporation activities comparable to its dxNT incorporation adds difficulty to the initial setting of the magnitude of the selective pressure. Only once mutants are discovered that can bypass certain templates that the wildtype and most other error-prone variants cannot bypass by “cheating” can this tradeoff be broken. The H545R mutant represents such a step, as it has been shown that there are some templates requiring multiple but spaced dxNT incorporations that the wild type cannot bypass, but the H545R mutant could. Selection with this type of template, instead of templates requiring merely one or two consecutive dxNT incorporations, should increase the fitness of such mutants, since this sets a bar where some variants can perform, and most cannot. This is in contrast to a lower selective pressure where most variants, like neutral variants of the wild type, are able to either bypass the barrier either by genuine dxNT incorporation or misincorporation. The fitness difference between these mutants and those with increased activity is smaller at this latter selection pressure. We presume that setting the bar at the former, higher level will lead to an increased enrichment of clones like E2 and larger fitness differences between them and those that retain wild type activities. Theoretically, error-prone “cheating” polymerases will then also be selected against directly. Since such polymerase mutants were not found, we certainly do not presume them to be so problematic as to compromise CST as a selection tool, especially when a low amount of selection rounds with low selective pressure are performed. Even if they would appear, the ELISA screening would not make them look like promising clones since in essence, the ELISA represents a similar selective pressure for mutants with the intended activity, and a high pressure increase for these theoretical error-prone parasites. They will not misinform inferences on the structure-function relationship of these XNA polymerases when they are eliminated during screening.

### 4.3. Nonspecific binding

Since CST is a technique that eventually separates the different genotypes using affinity chromatography, any DNA that remains attached to the magnetic beads ends up in the re-cloning pool. This includes the specifically bound plasmid-primer-biotin complexes bound to streptavidin on the magnetic beads, but also any plasmid that is bound nonspecifically to the bead surface. Such nonspecific binding is usually countered by increasing the number or stringency of washing steps, and pre-coating the beads with other molecules such as proteins to which the plasmids presumably do not bind, or extra DNA that then competes for these nonspecific binding sites. Multiple of these strategies were tested in our lab (Dr. Y. Peeters, data not shown), and while extra washing steps showed that some plasmids could still be removed from the beads, negative controls routinely showed the presence of some remaining plasmid in the final bead fraction. The PCR reactions that showed the presence of plasmids in washed fractions were not quantitative, so no conclusions can be drawn about the absolute amounts of plasmid in each fraction. Some high-salt washing buffers could also inhibit PCR reactions, possibly resulting in false negatives. These analytical PCR reactions were however tailored to have a low amount of cycles, so that enough product is formed to give the signal of plasmid presence, but that the PCR reaction does not reach saturation. Based on this fact, final eluted fractions with larger bands in positive controls and selected libraries, compared to negative controls, were interpreted as having more bound plasmids, namely the tagged ones. Since any nonspecifically bound plasmids should be of the same size and composition, it is assumed that in any round of CST, a relatively small number of plasmids can hitchhike along with the biotinylated complexes. However, since this nonspecifically bound population is probably a random sample of the library, it is unbiased and will favor beneficial mutants as much as it does deleterious ones and thereby introduce a small amount of genetic drift. Another possibility is the presence of nicked plasmid templates, which could be extended with biotin-tagged nucleotides, regardless of the CST primer extension. These would also bind streptavidin and would thus be extracted, but if there is no correlation between nicked or intact plasmids and the activity of their coded polymerase, these should merely introduce an unbiased background.

# 5. Screening and characterization

## 5.1. ELISA

### 5.1.1. *Template design*

The first method that was used to assess the activity of specific clones after selection was a polymerase activity assay based on ELISA. To make sure that the activity that is screened for closely resembles the selection environment, primer-template duplexes were designed that strongly resemble the sequences in the CST reaction. In the larger throughput screening ELISAs, these were templates that required a short part of DNA synthesis, followed by either one or two consecutive dxNT incorporations, and a longer stretch of DNA. Since this ELISA reaction requires probe binding to give a signal, this last stretch of DNA needed to be sufficiently long, but also lacking A in the sequence if xT was to be incorporated. For this reason, an intermediate linker was designed, so a single probe with all four bases could be used for all templates, without the sequence constraints it would otherwise cause.

### 5.1.2. *Throughput*

Since ELISA screening can be performed in 96-well plates, the activity of nearly 100 uniquely picked clones can be assessed per plate, and because several plates can be screened per experiment, it has a relatively good throughput. Still, this screen is only representative of libraries of a relatively small size and even then the odds of finding the highest performing mutant in the library are small, let alone finding multiple copies of the same mutant. It does offer a view of the overall composition of the library, as clones can be sorted according to activity in comparison to the wild type. The ratio of inactivating, deleterious, neutral and beneficial mutations should change over time as the population gets increasingly enriched in beneficial mutations and purged of deleterious ones. Even without sequencing specific clones, this can be a valuable tool to give insight into the library composition. It is likely that selecting to the point that only a handful of genotypes remain will cause the loss of other potentially useful mutations. The screening of libraries after one or two rounds of selection respectively showed that a little over half of the sampled clones were inactive, around 25% and 5% were less active than the wild type, around 5% to 10% were beneficial and the rest were neutral. Clearly there is still room for further selection on these libraries, but it is estimated that with about three more rounds of enrichment, screening could sample the majority of the best performing clones.

### *5.1.3. Variability and false positives and negatives*

The ELISA signal is measured by absorbance of the chromogenic substrate that is converted by the HRP bound to the probe. This means that the only measure of activity is the magnitude of absorbance, and any experimental error during the entire procedure, from expression to washing away unbound antibody-HRP fusion proteins, can cause false positive or false negative signals. False positives were rarer, as negative controls always showed background level absorbance, but some positive or reference controls sometimes showed lower or even no signal, possibly due to less efficient protein expression. The quality and sampling of the cell lysate can also change the amount of active protein in the reaction from the start. Therefore, experimental replication of the same lysates, and replication of expression of independent clones of the same mutant were investigated. Experimental variation turned out to be relatively small when performed by experienced users, compared to the variability caused by using lysates from independent clones. To control for the variability between experiments, replication should be performed at the level of expressing clones. Replication does influence the amount of clones that can be screened in parallel within the same experiment, keeping the amount of screened clones to around $10^2$ per experiment.

Another potential pitfall of assaying cell lysates, besides poor control on effective protein concentration, is the presence of cellular nucleotides. A reaction with e.g. natural dA, dG and dC and synthetic dxT will not completely lack dT in the reaction if it is present in the lysate. Given the long incubation times and obligate usage of very few d(x)T in the extension reaction, polymerases could eventually use the naturally present dT while stalling on dxT incorporation. For this reason, more detailed activity analysis should be performed with purified protein, or cell cultures or lysates should be pre-treated so that as little natural nucleotides as possible are present in the lysate.

### *5.1.4. Invisibility of stalled intermediates*

Due to the necessity of the synthesis of a probe complement, reaction intermediates that represent extended primers that are so short that the linker and probe cannot bind will not give a signal. The total signal of a reaction is only determined by the integration of all products that are sufficiently long to bind the linker, and these all have a length that required incorporation and complete translocation of the dxNTs (more than one helical turn or ~10 bp). An incomplete, stalled reaction and a negative reaction will look the same, and ELISA will therefore only give information about the activity of polymerases that can produce at least some full-length products. A technique such as PAGE is necessary to reveal this more detailed information.

### 5.2. PAGE

Since short nucleic acids can be visualized at single nucleotide resolution with highly concentrated and denaturing PAGE, this technique is suitable for obtaining much more information about specific checkpoints and stalling of polymerases while they incorporate and extend dxNTs. Because the reactions can be quenched, reaction products from different time points can be compared, adding a temporal dimension to the activity measurement. The main difference with the ELISA screening, besides this higher resolution, is the lower throughput. The time and work required to prepare, load and run slab gels limits the amount of samples that can be run, and so only a limited number of clones can be compared and analyzed. PAGE was thus used only to validate clones performing well in ELISA, including clone E2, and not to detect new clones.

The problems due to the use of cellular lysates as a source of polymerase and potentially cellular nucleotides and the variation in expression efficiencies are circumvented by using purified proteins. Most cellular material that could inhibit or otherwise influence the extension reaction is then removed, and the concentrations of different mutant polymerases can be normalized by estimating their absolute yield with SDS-PAGE, to make sure a higher activity is not just due to a higher effective enzyme concentration.

The ability to see intermediate elongation products at nucleotide resolution, not just those that completed translocation, offers a better view of the polymerase reaction. This way, it was found that the wildtype activity of incorporating up to two dxNTPs stalls specifically because a third consecutive dxNTP cannot be incorporated, not because it cannot be extended at all. Using a range of more complex templates instead of simply polyA or polyT templates showed the major checkpoints where stalling occurs, which could then be correlated with the position of the dxNTs in the polymerase. This structural information then becomes important to understand mutant phenotypes and can help guide future mutagenesis. In this case, the translocation of single dxNT or closely spaced dxNTs is thought to distort the backbone structure, causing suboptimal backbone-protein interactions and steric hindrance. Steric hindrance would introduce such stalling checkpoints, and using PAGE these locations can be targeted with high positional resolution.

## 5.3. Mutation H545R

The best performing clone with dxNT extension activity obtained so far was clone E2 which carries a single amino acid substitution. Since this mutation is the only difference with the wild type, all phenotypic effects of this mutant should be due to this change of His to Arg at position 545. When compared to protein sequences of other archaeal family B DNAPs of varying phylogenetic distance, position 545 is interesting because the amino acids are not identical, even within an otherwise extremely conserved active site. Amino acids occurring at this position are still remarkably conserved: all these polymerases either have His, Tyr or Phe. This implies that there is some variation possible at this position, as long as it is occupied by a bulky aromatic residue. By mutating His to Arg, the guanidinium head could maintain aromatic stacking with a nearby Tyr, while the increase in length could cause the head to displace a bound water molecule, and bring a stronger, positive charge closer to the first phosphate group of the primer backbone. Since the backbone distortion caused by a dxNT incorporation is nearly always accompanied by a stalled intermediate at this position for all tested polymerases, it is hypothesized that this Arg residue stabilizes the phosphate group at this position.

Another effect of this mutation is a decrease in misincorporation and template-independent terminal transferase activity when compared with the wild type. To investigate whether this is the result of changed fidelity, relative comparisons of rates should be made between correct and incorrect incorporations and between the wild type and E2. In absence of this data, a possible alternative explanation is an intrinsically slower incorporation rate of the E2 clone. This could also explain why the E2 clone performed slower on polyA or polyT templates, but was improved at dxNT extension. Since there was selection for dxNT incorporation and extension (also with DNA), it is not surprising that the E2 clone performs better at what it was selected for, even when this comes with an apparent tradeoff, namely decreasing the speed at which the first dxNTs are incorporated. The fact that misincorporation rates decreased, together with lowered activity for arabinose-based nucleotides and the position of the R545 mutation being in direct contact with a backbone phosphate group prone to steric distortion, argues for a mechanism of increased specificity for dxNT substrates. A mutation that increases tolerance for xenonucleotides in a non-specific way, such as the A485L terminator mutation (93), would likely increase error rates and increase promiscuous activities for other XNAs nucleotides.

## 6. Future work

Since the development of a dedicated dxNA polymerase is far from complete, there are several goals that have still to be met. These are given below, from the most immediate work that could be easily implemented, to several suggestions for improving the strategy to evolve a DNA-dependent dxNA polymerase and finally the ultimate aims of the overall research program of developing informationally orthogonal XNAs.

### 6.1. Follow-up experiments

Only a tiny fraction of the variation in all libraries prepared in this study has been sampled to date, simply because the libraries are still large, even after two selection rounds, and several hundred selected clones have been screened for activity. Using the existing protocols, it is quite plausible that several improved mutants can still be found in the current libraries. The found mutations, especially the H545R mutation, could also be introduced in structurally similar DNA or XNA polymerases to investigate if its phenotype is transferrable in a different protein context. Since selection and screening was focused on the active site and minor groove binding regions, and to a lesser extent on the template binding region at the end of the thumb domain, the generated nucleotide binding pocket and finger domain libraries are still largely unexplored, and could be selected and screened for mutants with higher dxNTP affinity (most likely to be found in libraries F and N) or improved translocation (libraries P and T). If the library quality can be monitored and deletion mutants can be controlled using gel extraction purification, selection can enrich beneficial mutants and strongly increase the “hit” rate in following screening.

With standard mutagenesis and the existing protocols for shuffling, such mutants can be combined and shuffled with increased efficiency. To shuffle larger libraries using GGS, or more fragments at once, the GGS protocol would require extra optimization to improve ligation yields.

### 6.2. The role of the exonuclease domain and activity

Since family B polymerases, Therminator polymerase included, have intrinsic exonuclease activity due to the presence of an exonuclease domain, this activity is usually removed by mutation of catalytic Asp/Glu to Ala. This is because in such polymerases, the extended primer can move between a polymerization state and an editing state, each with their own catalytic active site in either the exonuclease or palm domain. This switch to an editing state is presumably influenced by mismatch or distortion-induced stalling, which is expected to happen during XNA synthesis. An active exonuclease domain could thus continuously undo xenonucleotide incorporation by 3’-5’ exonuclease activity. Exo- polymerases will therefore incorporate xenonucleotides irreversibly, but at a cost of lowered fidelity. It is therefore tempting to conclude that once an XNAP is developed that can efficiently polymerize XNA, restoring exonuclease activity could increase fidelity. This is however predicated on two conditions: the ability of the native exonuclease domain to remove XNA from the primer strand (which is not a guarantee given that it is already evolutionarily optimized for DNA) and a preference for XNA with a correct sequence for the productive polymerization state, only preferring the editing state when misincorporations occur. The exonuclease domain itself could also become the subject of mutation and selection, since the evolution of a higher fidelity XNA polymerase will be very useful, and relatively straightforward once it is capable of a self-replication cycle.

### 6.3. Next-generation sequencing to monitor library enrichment

Historically, libraries used for molecular evolution have been subjected to several rounds of selection to enrich it in well performing clones to such a level that the following screening and sequencing will have a high number of positive hits. However, since enrichment due to selection should happen at every round, it is possible in principle to sequence the library before and after selection and compare which genotypes are increasing in frequency. The main limitation to this is the throughput, but by using next-generation sequencing (NGS), it has become possible to sequence large libraries, and gain information about positive clones without the need for a large amount of selection rounds, which are laborious and time consuming. For example, it is possible to sequence a population of over $10^{10}$ human immunoglobulins displayed on phages (138). Biopanning of phages displaying recombinant antibodies that was followed by NGS could recover enriched yet still rare variants after one or two rounds of enrichment (139). If there is a high correlation between enrichment in the first round and further rounds, it may not even be necessary to perform further enrichment (140). Besides this direct measurement of relative fitness of a very large number of clones, NGS can also be used to sequence unselected libraries to control the quality, for example biases or artefacts that are introduced during library production (141).

A potential downside is the relatively short read length of up to a few hundred bp by the more accurate sequencing platforms, such as Illumina (142). While this can be sufficient for spanning single target regions, larger or recombined regions cannot be sequenced in the same clone, and any spontaneous mutations that appeared outside the narrow sequencing window will be difficult to link. If these cannot be recovered but are instead only known by their sequence, recreating the sequenced genotypes will appear as false positives.

### 6.4. Flow cytometry to combine selection and screening

While emulsion-based selection techniques have been the method of choice for molecular evolution of polymerases, there are other selection techniques that could potentially be used. One of these is microfluidic flow cytometry where cells or compartments are sorted based on their properties (143). Compartmentalization, and therefore genotype-phenotype coupling, can be maintained and sorting can be used as the selection mechanism, so in principle it can be used to evolve a population of proteins. Depending on the method of droplet separation and sorting, throughputs can lie in the order of several kilohertz, with electric and acoustic sorting being much faster than thermal, pneumatic and especially magnetically induced sorting (144). A 20 kHz sorting system could sort $10^7$ clones in about 8 minutes, meaning that libraries such as in this work are well within the practical range. A CST or CSR-like approach could be coupled to cytometry, but the usage of water-in-oil emulsions is not compatible with these types of microfluidic devices. However, it is possible to create double water-in-oil-in-water emulsions that can be used to perform fluorescence-activated cell sorting (FACS) (145) for directed evolution of proteins (146).

To be able to select, a selectable signal is needed. The most straightforward signal is a fluorescent signal that can easily be read out. This means the function that is selected for should produce a fluorescent signal. This could be achieved by using a fluorogenic substrate, a FRET pair or quencher/fluorophore pair. In the case of dxNA production, a fluorogenic substrate  such as terminal phosphate-labeled fluorogenic nucleotides (TPLFNs) (147) is less obvious, as the custom-made dxNTP is not simply modified to release a modified leaving group that fluoresces after incorporation. The selected reaction is also very indirect, namely hydrolysis of the leaving group, rather than incorporation and extension directly. A method could be envisioned where a short DNA duplex probe carries a fluorophore and quencher each on a different strand, and a shorter but highly abundant primer, together with nucleotides and a polymerase mutant are added. In the context of the low selection pressures that have to be used at this stage in dxNA polymerase development, the primer should require the incorporation of one or more dxNTs, followed by a stretch of DNA that hybridizes to the part of the template duplex where the fluorophore and quencher are located. This would cause displacement and competition between extended primers and labeled oligonucleotides, resulting in an increase in fluorescence as a positive selection signal.

The resemblance with CST can be decreased even further by avoiding the need for transformation and *in vivo* expression. A mix of precisely concentrated plasmids and an *in vitro* translation system could be emulsified. For this to work, it is even more important that there is a strong control over droplet size and composition, which can be achieved with microfluidic rather than bulk emulsification (148). The added difficulty of this strategy would lie in the necessity of adding the components for the reaction that is being selected on, either from the start (risking interference) or by combining droplets with *in vitro* translated protein with the extension reaction substrates in the microfluidic device, which is technically not straightforward unless specialized equipment and expertise is available for pico-injection.

## 6.5. Towards a fully independent XNA episome

The major hurdle in evolving a dxNA-dependent dxNA polymerase from a DNA-dependent DNA polymerase lies in the viability of the transitional polymerases that are (temporarily) selected for. The strategy of this PhD thesis has been on first evolving a DNA-dependent dxNA polymerase, since this could already have a practical use for dxNA transcription from DNA, for example for enzymatic dxNA aptamer production. However, given the necessity of DNA template binding, the thumb domain will remain selected for the DNA structure. Given that the reverse transitional protein, a dxNA-dependent DNA polymerase, is the one that binds dxNA templates, selection for dxNA template recognition is initially less prudent. This activity only becomes important when dxNA needs to be cloned (after CST/CSR) or sequenced (during aptamer development). At some point however, the thumb domain will need a major overhaul that cannot be achieved by mere point mutations in a structure that is adapted to the DNA helix. It is suggested that double-stranded dxNA-binding polypeptides can be selected for simple binding affinity to dxNA directly by affinity chromatography. Such a large population of peptides can then be prepared as a shuffleable fragment that is then inserted into the thumb domain. A hypothetical example is the bundle of three alpha helices near the C-terminus, one of which contains the T region. This entire structure could be replaced in a dxNA polymerase by a library of peptides that have been previously selected for dxNA binding. The resulting insertion mutants can then be selected on polymerase activity on dxNA templates. This type of mutant libraries can only be generated and combined in the large populations required for molecular evolution with shuffling-based techniques.

The goal of an informationally orthogonal nucleic acid like dxNA that is currently furthest in the future, is that of a self-contained organism with an orthogonal genome. To say that this goal is challenging is an understatement, given that it essentially requires the design of a complete life form that also differs from all existing life. It is needless to say that many hurdles need to be overcome. The synthesis of such a genome alone, which is already an achievement in minimal genome studies for bacteria with the smallest genomes (149), is merely a technical hurdle. The complexity of engineering a complete metabolism, albeit largely based on the natural one, is daunting. A less arduous step is that of an independent XNA episome, analogous to a bacterial plasmid, that is carried by a natural host cell with an otherwise natural DNA/RNA metabolism. If such an episome is made of dxNA or a similar informationally orthogonal polymer, this “genetic firewall” should, in principle, restrain both recombination with the host genome and escape into the environment outside the host cell. Some of the minimal requirements are first of all replication by an XNA-dependent XNA polymerase, of which this work is a first step. Other enzymes that would have to be either evolved or designed are proteins such as for example helicases, ligases, topoisomerases and nucleases. These would be necessary for manipulation of XNA *in vitro*, and to allow *in vivo* functionality that extends from mere linear primed replication to controlled replication and genomic stability.

To encode functional genes on this XNA episome, a transcription polymerase is needed that can transcribe dxNA to another genetic polymer such as xNA, or RNA. These transcripts could either function on their own as aptamers or (xeno)ribozymes, or be translated to proteins. All these choices have opportunities and caveats. Transcription to RNA would instantly break the genetic firewall, since a reverse transcriptase could then easily copy the gene sequence back to DNA. Transcription to anything other than RNA could very well produce a molecule that folds into a functional shape, but if it is to be translated to protein it becomes necessary to adapt or invent an entire translation machinery, complete with tRNAs and ribosomes capable of recognizing and translating non-RNA. A middle ground could consist of a new XNA-to-RNA transcriptase, but with an expanded alphabet, for example an XY base pair as well as AT and GC ones with some tRNAs that have new anticodons with these XY base pairs. If a large enough fraction of the coding sequence consists of XY base pairs, reverse transcription cannot carry sequence information meaningfully back into DNA, and an RNA transcript with unnatural bases should inhibit translation by natural ribosomes and tRNAs. Thus, no protein can be formed from such a transcript by hosts lacking these novel tRNAs and XY base pair system, even if the transcript has the natural RNA backbone.

Although this is a hypothetical example, one thing is clear: an *in vivo* XNA episome that is informationally orthogonal through changes in the backbone and/or nucleobase will also need to be metabolically orthogonal, in the sense that the nucleotides should not inhibit natural polymerases. In the concrete case of dxNTs, they could easily act as chain terminating inhibitors and stall the host polymerases. It is for this reason that even if the dxNA product would be completely orthogonal, the nucleotides themselves could be altered so they are not interfering with the large number of natural proteins that require nucleotides as (regulatory) substrates. A prime target for modification that removes substrate recognition would be an altered leaving group. The combination of multiple differences in the backbone, base and leaving group moiety, should abolish recognition as a nucleotide but retain all the functional parts, in order to have an XNA episome function in a living host. This was the case, for example, where XNA with a HNA backbone and modified bases (5-methyl-isocytosine and isoguanine) resulted in XNAs could not be read through by natural polymerases, *in vitro* and *in vivo* (150, 151).

## 7. Conclusions

To evolve an XNA polymerase from a DNA polymerase that has only a very low promiscuous activity for these xenonucleotides, where the formed XNA does not hybridize to DNA and cannot be reverse transcribed, several issues arise. Selection for activity will have to be set at a very low bar, which allows mutants to be selected for low fidelity, and the reaction that is selected for will have to be indirect. In the case of Therminator polymerase, this involves incorporation of only a couple of dxNTs and their extension has to happen with partly biotinylated DNA, which implies co-selection of DNA replication. The products are also too short and lack an enzyme capable of reverse transcription, so self-replication strategies that safeguard against fidelity loss cannot be used yet. Still it is possible to increase activity and tolerance for dxNTs by a small, local amount of substitution mutations. Once the promiscuous activity is increased to levels where wild type mismatch bypassing is effectively blocked, larger selective pressures can be implemented.

We nevertheless successfully selected a DNA polymerase mutant, H545R, that can incorporate dxNTs after two rounds of CST, albeit slower than the wildtype. However, while the wild type is faster, it stalls after the incorporation of two consecutive dxNTs. The H545R mutant is able to extend and presumably translocate these dxNT-induced distortions in the nucleic acid backbone, to a point where four dxNTs spaced by two dNTs each do not inhibit complete extension. This mutant neither achieved this by becoming more error-prone or by performing template-independent terminal transferase activity, nor did it gain increased general tolerance to any XNA, as it became less efficient for arabinose-based nucleotides.

It is unlikely though that merely substituting residues in the DNA polymerase active site fold can eventually change it into a dedicated dxNA polymerase, given that the contacts between nucleic acid and protein trace the nucleic acid shape and are constrained by the larger scale protein fold. To evolve a thumb domain that binds a much more extended dxNA polymer with a different helical handedness, a more drastic approach of substituting whole regions by different secondary structure elements could prove more efficient, potentially aided by computational methods.

# References

1. Schmidt M. (2010) Xenobiology: a new form of life as the ultimate biosafety tool. *Bioessays*, **32**, 322–31.

2. Ausländer S., Ausländer D. and Fussenegger M. (2017) Synthetic Biology-The Synthesis of Biology. *Angew. Chem. Int. Ed. Engl.*, **56**, 6396–6419.

3. Chaput J.C., Yu H. and Zhang S. (2012) The emerging world of synthetic genetics. *Chem. Biol.*, **19**, 1360–1371.

4. Eremeeva E. and Herdewijn P. (2019) Non canonical genetic material. *Curr. Opin. Biotechnol.*, **57**, 25–33.

5. Herdewijn P. and Marlière P. (2012) Redesigning the leaving group in nucleic acid polymerization. *FEBS Lett.*, **586**, 2049–56.

6. Adelfinskaya O. and Terrazas M. (2007) Polymerase-catalyzed synthesis of DNA from phosphoramidate conjugates of deoxynucleotides and amino acids. *Nucleic Acids Res.*, **35**, 5060–72.

7. Yang S., Froeyen M., Lescrinier E., Marlière P. and Herdewijn P. (2011) 3-Phosphono-L-alanine as pyrophosphate mimic for DNA synthesis using HIV-1 reverse transcriptase. *Org. Biomol. Chem.*, **9**, 111–119.

8. Brudno Y. and Liu D. (2009) Recent progress toward the templated synthesis and directed evolution of sequence-defined synthetic polymers. *Chem. Biol.*, **16**, 265–276.

9. Xu W., Chan K.M. and Kool E.T. (2018) Fluorescent nucleobases as tools for studying DNA and RNA. *Nat. Chem.*, **9**, 1043–1055.

10. Das K. and Arnold E. (2013) HIV-1 reverse transcriptase and antiviral drug resistance. Part 1. *Curr. Opin. Virol.*, **3**, 111–118.

11. Henry A.A., Yu C. and Romesberg F.E. (2003) Determinants of unnatural nucleobase stability and polymerase recognition. *J. Am. Chem. Soc.*, **125**, 9638–9646.

12. Malyshev D.A., Dhami K., Lavergne T., Chen T., Dai N., Foster J.M., Corrêa I.R. and Romesberg F.E. (2014) A semi-synthetic organism with an expanded genetic alphabet. *Nature*, **509**, 385–388.

13. Eremeeva E., Abramov M., Margamuljana L., Rozenski J., Pezo V., Marli P. and Herdewijn P. (2016) Nucleic acids chemical morphing of DNA containing four noncanonical bases. *Angew. Chemie Int. Ed.*, **55**, 7515–7519.

14. Eremeeva E., Abramov M., Margamuljana L. and Herdewijn P. (2017) Base-modified nucleic acids as a powerful tool for synthetic biology and biotechnology. *Chem. Eur. J.*, **23**, 9560–9576.

15. Hoshika S., Leal N.A., Kim M.-J., Kim M.-S., Karalkar N.B., Kim H., Bates A.M., Watkins N.E., SantaLucia H.A., Meyer A.J., *et al.* (2019) Hachimoji DNA and RNA: A genetic system with eight building blocks. *Science*, **363**, 884–887.

16. Joyce C.M. (2010) Techniques used to study the DNA polymerase reaction pathway. *Biochim. Biophys. Acta*, **1804**, 1032–1040.

17. Pinheiro V.B., Taylor A.I., Cozens C., Abramov M., Renders M., Zhang S., Chaput J.C., Wengel J., Peak-Chew S.Y., McLaughlin S.H., *et al.* (2012) Synthetic genetic polymers capable of heredity and evolution. *Science*, **336**, 341–344.

18. Torres L.L. and Pinheiro V.B. (2018) Xenobiotic nucleic acid (XNA) synthesis by Phi29 DNA polymerase. *Curr. Protoc. Chem. Biol.*, **10**, e41.

19. Maiti M., Maiti M., Knies C., Dumbre S., Lescrinier E., Rosemeyer H., Ceulemans A. and Herdewijn P. (2015) Xylonucleic acid: Synthesis, structure, and orthogonal pairing properties. *Nucleic Acids Res.*, **43**, 7189–7200.

20. Maiti M., Siegmund V., Abramov M., Lescrinier E., Rosemeyer H., Froeyen M., Ramaswamy A., Ceulemans A., Marx A. and Herdewijn P. (2012) Solution structure and conformational dynamics of deoxyxylonucleic acids (dXNA): An orthogonal nucleic acid candidate. *Chem. Eur. J.*, **18**, 869–879.

21. Poopeiko N.E., Juhl M., Vester B., Sørensen M.D. and Wengel J. (2003) Xylo-Configured oligonucleotides (XNA, xylo nucleic acid): synthesis of conformationally restricted derivatives and hybridization towards DNA and RNA complements. *Bioorg. Med. Chem. Lett.*, **13**, 2285–2290.

22. Ramaswamy A., Froeyen M., Herdewijn P. and Ceulemans A. (2010) Helical structure of xylose-DNA. *J. Am. Chem. Soc.*, **132**, 587–595.

23. Chin J.W. (2017) Expanding and reprogramming the genetic code. *Nature*, **550**, 53–60.

24. Mayer G. (2009) The chemical biology of aptamers. *Angew. Chemie*, **48**, 2672–2689.

25. Keefe A.D., Pai S. and Ellington A. (2010) Aptamers as therapeutics. *Nat. Rev. Drug Discov.*, **9**, 537–550.

26. Stoltenburg R., Reinemann C. and Strehlitz B. (2007) SELEX—A (r)evolutionary method to generate high-affinity nucleic acid ligands. *Biomol. Eng.*, **24**, 381–403.

27. Gasse C., Zaarour M., Noppen S., Abramov M., Marlière P., Liekens S., De Strooper B. and Herdewijn P. (2018) Modulation of BACE1 Activity by Chemically Modified Aptamers. *ChemBioChem*, **19**, 754–763.

28. Lin Y., Nieuwlandt D., Magallanez A., Feistner B. and Jayasena S.D. (1996) High-affinity and specific recognition of human thyroid stimulating hormone (hTSH) by in vitro-selected 2’-amino-modified RNA. *Nucleic Acids Res.*, **24**, 3407–3414.

29. Biesecker G., Dihel L., Enney K. and Bendele R.A. (1999) Derivation of RNA aptamer inhibitors of human complement C5. *Immunopharmacology*, **42**, 219–230.

30. Ruckman J. (1998) 2’-Fluoropyrimidine RNA-based aptamers to the 165-amino acid form of vascular endothelial growth factor (VEGF165). *J. Biol. Chem.*, **273**, 20556–20567.

31. Burmeister P.E., Lewis S.D., Silva R.F., Preiss J.R., Horwitz L.R., Pendergrast P.S., McCauley T.G., Kurz J.C., Epstein D.M., Wilson C., *et al.* (2005) Direct in vitro selection of a 2’-O-methyl aptamer to VEGF. *Chem. Biol.*, **12**, 25–33.

32. Haruta K., Otaki N., Nagamine M., Kayo T., Sasaki A., Hiramoto S., Takahashi M., Hota K., Sato H. and Yamazaki H. (2017) A novel PEGylation method for improving the pharmacokinetic properties of anti-interleukin-17A RNA aptamers. *Nucleic Acid Ther.*, **27**, 36–44.

33. Johansson E. and Dixon N. (2013) Replicative DNA Polymerases. *Cold Spring Harb. Perspect. Biol.*, **5**, a012799–a012799.

34. Meselson M. and Stahl F.W. (1958) The replication of DNA in Escherichia coli. *Proc. Natl. Acad. Sci. USA*, **44**, 671–682.

35. Rothwell P.J. and Waksman G. (2005) Structure and mechanism of DNA polymerases. *Adv. Protein Chem.*, **71**, 401–440.

36. 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**, 763–73.

37. Raia P., Delarue M. and Sauguet L. (2019) An updated structural classification of replicative DNA polymerases. *Biochem. Soc. Trans.*, **47**, 239–249.

38. Mönttinen H.A.M., Ravantti J.J., Stuart D.I. and Poranen M.M. (2014) Automated structural comparisons clarify the phylogeny of the right-hand-shaped polymerases. *Mol. Biol. Evol.*, **31**, 2741–2752.

39. Greenough L., Menin J.F., Desai N.S., Kelman Z. and Gardner A.F. (2014) Characterization of Family D DNA polymerase from Thermococcus sp. 9°N. *Extremophiles*, **18**, 653–664.

40. Čuboňová L., Richardson T., Burkhart B.W., Kelman Z., Connolly B.A., Reeve J.N. and Santangelo T.J. (2013) Archaeal DNA polymerase D but Not DNA polymerase B is required for genome replication in thermococcus kodakarensis. *J. Bacteriol.*, **195**, 2322–2328.

41. Henneke G., Flament D., Hübscher U., Querellou J. and Raffin J.P. (2005) The hyperthermophilic euryarchaeota Pyrococcus abyssi likely requires the two DNA polymerases D and B for DNA replication. *J. Mol. Biol.*, **350**, 53–64.

42. Lipps G., Ro S., Hart C. and Krauss G. (2003) A novel type of replicative enzyme harbouring ATPase, primase and DNA polymerase activity. *EMBO J.*, **22**, 2516–2525.

43. Bienstock R.J., Beard W.A. and Wilson S.H. (2014) Phylogenetic analysis and evolutionary origins of DNA polymerase X-family members. *DNA Repair (Amst).*, **22**, 77–88.

44. Motea E.A. and Berdis A.J. (2010) Terminal deoxynucleotidyl transferase: The story of a misguided DNA polymerase. *Biochim. Biophys. Acta - Proteins Proteomics*, **1804**, 1151–1166.

45. Goodman M.F. and Woodgate R. (2004) Translesion DNA Polymerases. *Cold Spring Harb. Perspect. Biol.*, **4**, 247–250.

46. Kohlstaedt L.A., Wang J., Friedman J.M., Rice P.A. and Steitz T.A. (1992) Crystal structure at 3.5 A resolution of HIV-1 reverse transcriptase complexed with an inhibitor. *Science*, **256**, 1783–1790.

47. Tsai Y.-C. and Johnson K.A. (2006) A new paradigm for DNA polymerase specificity. *Biochemistry*, **45**, 9675–9687.

48. Yang G., Franklin M., Li J., Lin T.C. and Konigsberg W. (2002) Correlation of the kinetics of finger domain mutants in RB69 DNA polymerase with its structure. *Biochemistry*, **41**, 2526–2534.

49. Schweitzer B.A. and Kool E.T. (1994) Aromatic nonpolar nucleosides as hydrophobic isosteres of pyrimidines and purine nucleosides. *J. Org. Chem.*, **59**, 7238–7242.

50. Fitch W.M. (1967) Evidence suggesting a non-random character to nucleotide replacements in naturally occurring mutations. *J. Mol. Biol.*, **26**, 499–507.

51. Brautigam C.A. and Steitz T.A. (1998) Structural and functional insights provided by crystal structures of DNA polymerases and their substrate complexes. *Curr. Opin. Struct. Biol.*, **8**, 54–63.

52. Kottur J. and Nair D.T. (2018) Pyrophosphate hydrolysis is an intrinsic and critical step of the DNA synthesis reaction. *Nucleic Acids Res.*, **46**, 5875–5885.

53. Johnson K.A. (1995) Rapid quench kinetic analysis of polymerases, adenosinetriphosphatases, and enzyme intermediates. In *Methods in Enzymology*.Vol. 249, pp. 38–61.

54. Stengel G., Gill J.P., Sandin P., Wilhelmsson L.M., Albinsson B., Nordén B. and Millar D. (2007)

Conformational dynamics of DNA polymerase probed with a novel fluorescent DNA base analogue. *Biochemistry*, **46**, 12289–12297.

55. Rothwell P.J., Mitaksov V. and Waksman G. (2005) Motions of the fingers subdomain of Klentaq1 are fast and not rate limiting: implications for the molecular basis of fidelity in DNA Polymerases. *Mol. Cell*, **19**, 345–355.

56. Joyce C.M., Potapova O., DeLucia A.M., Huang X., Basu V.P. and Grindley N.D.F. (2008) Fingers-closing and other rapid conformational changes in DNA Polymerase I (Klenow Fragment) and their role in nucleotide selectivity. *Biochemistry*, **47**, 6103–6116.

57. Kuchta R.D., Mizrahi V., Benkovic P.A., Johnson K.A. and Benkovic S.J. (1987) Kinetic mechanism of DNA polymerase I (Klenow). *Biochemistry*, **26**, 8410–8417.

58. Patel S.S., Wong I. and Johnson K.A. (1991) Pre-steady-state kinetic analysis of processive DNA replication including complete characterization of an exonuclease-deficient mutant. *Biochemistry*, **30**, 511–525.

59. Woolfson D.N., Bartlett G.J., Burton A.J., Heal J.W., Niitsu A., Thomson A.R. and Wood C.W. (2015) De novo protein design: How do we expand into the universe of possible protein structures? *Curr. Opin. Struct. Biol.*, **33**, 16–26.

60. Cuff A.L., Sillitoe I., Lewis T., Clegg A.B., Rentzsch R., Furnham N., Pellegrini-Calace M., Jones D., Thornton J. and Orengo C.A. (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.*, **39**, D420–D426.

61. Fox N.K., Brenner S.E. and Chandonia J.-M. (2014) SCOPe: Structural Classification of Proteins—extended, integrating SCOP and ASTRAL data and classification of new structures. *Nucleic Acids Res.*, **42**, D304–D309.

62. Keefe A.D. and Szostak J.W. (2001) Functional proteins from a random-sequence library. *Nature*, **410**, 715–718.

63. Mansy S.S., Zhang J., Kümmerle R., Nilsson M., Chou J.J., Szostak J.W. and Chaput J.C. (2007) Structure and evolutionary analysis of a non-biological ATP-binding protein. *J. Mol. Biol.*, **371**, 501–513.

64. Hayashi Y., Aita T., Toyota H., Husimi Y., Urabe I. and Yomo T. (2006) Experimental rugged fitness landscape in protein sequence space. *PLoS One*, **1**, e96.

65. Dunitz J.D. and Joyce G.F. (2013) Leslie Eleazer Orgel: 12 January 1927 — 27 October 2007. *Biogr. Mem. Fellows R. Soc.*, **59**, 277–289.

66. Liu C., Cozens C., Jaziri F., Rozenski J., Maréchal A., Dumbre S., Pezo V., Marlière P., Pinheiro V.B., Groaz E., *et al.* (2018) Phosphonomethyl oligonucleotides as backbone-modified artificial genetic polymers. *J. Am. Chem. Soc.*, **140**, 6690–6699.

67. Copley S.D. (2017) Shining a light on enzyme promiscuity. *Curr. Opin. Struct. Biol.*, **47**, 167–175.

68. Peracchi A. (2018) The limits of enzyme specificity and the evolution of metabolism. *Trends Biochem. Sci.*, **43**, 984–996.

69. Laos R., Thomson J.M. and Benner S.A. (2014) DNA polymerases engineered by directed evolution to incorporate non-standard nucleotides. *Front. Microbiol.*, **5**, 1–14.

70. Padiolleau-Lefèvre S., Naya R. Ben, Shahsavarian M.A., Friboulet A. and Avalle B. (2014) Catalytic antibodies and their applications in biotechnology: state of the art. *Biotechnol. Lett.*, **36**, 1369–1379.

71. Chusacultanachai S. and Yuthavong Y. Random mutagenesis strategies for construction of large and diverse clone libraries of mutated DNA fragments. **270**, 319–333.

72. Greener A., Callahan M. and Jerpseth B. (1997) An efficient random mutagenesis technique using an E. coli mutator strain. *Mol. Biotechnol.*, **7**, 189–195.

73. Spee J.H., de Vos W.M. and Kuipers O.P. (1993) Efficient random mutagenesis method with adjustable mutation frequency by use of PCR and dITP. *Nucleic Acids Res.*, **21**, 777–778.

74. Fujii R., Kitaoka M. and Hayashi K. (2004) One-step random mutagenesis by error-prone rolling circle amplification. *Nucleic Acids Res.*, **32**, e145.

75. Reetz M.T., Prasad S., Carballeira J.D., Gumulya Y. and Bocola M. (2010) Iterative saturation mutagenesis accelerates laboratory evolution of enzyme stereoselectivity: rigorous comparison with traditional methods. *J. Am. Chem. Soc.*, **132**, 9144–9152.

76. Cozens C. and Pinheiro V.B. (2018) Darwin Assembly: fast, efficient, multi-site bespoke mutagenesis. *Nucleic Acids Res.*, **46**, e51–e51.

77. Pinheiro V.B. (2019) Engineering-driven biological insights into DNA polymerase mechanism. *Curr. Opin. Biotechnol.*, **60**, 9–16.

78. Trudeau D.L., Smith M.A. and Arnold F.H. (2013) Innovation by homologous recombination. *Curr. Opin. Chem. Biol.*, **17**, 902–909.

79. Ghadessy F.J., Ong J.L. and Holliger P. (2001) Directed evolution of polymerase function by compartmentalized self-replication. *Proc. Natl. Acad. Sci. USA*, **98**, 4552–4557.

80. D'Abbadie M., Hofreiter M., Vaisman A., Loakes D., Gasparutto D., Cadet J., Woodgate R., Pääbo S. and Holliger P. (2007) Molecular breeding of polymerases for amplification of ancient DNA. *Nat. Biotechnol.*, **25**, 939–943.

81. Povilaitis T., Alzbutas G., Sukackaite R., Siurkus J. and Skirgaila R. (2016) In vitro evolution of phi29 DNA polymerase using isothermal compartmentalized self replication technique. *Protein Eng. Des. Sel.*, **29**, 617–628.

82. Ong J.L., Loakes D., Jaroslawski S., Too K. and Holliger P. (2006) Directed evolution of DNA polymerase, RNA polymerase and reverse transcriptase activity in a single polypeptide. *J. Mol. Biol.*, **361**, 537–550.

83. Ghadessy F.J. and Holliger P. (2007) Compartmentalized Self-Replication. In *Methods in Molecular Biology, vol. 352: Protein Engineering Protocols*.pp. 237–248.

84. Pinheiro V.B., Arangundy-Franklin S. and Holliger P. (2014) Compartmentalized self-tagging for in vitro-directed evolution of XNA polymerases. *Curr. Protoc. Nucleic Acid Chem.*, **2014**, 9.9.1-9.9.18.

85. Fernandez-Gacio A., Uguen M. and Fastrez J. (2003) Phage display as a tool for the directed evolution of enzymes. *Trends Biotechnol.*, **21**, 408–414.

86. Smith G. (1985) Filamentous fusion phage: novel expression vectors that display cloned antigens on the virion surface. *Science*, **228**, 1315–1317.

87. Barbas C.F., Kang A.S., Lerner R.A. and Benkovic S.J. (1991) Assembly of combinatorial antibody libraries on phage surfaces: the gene III site. *Proc. Natl. Acad. Sci. USA*, **88**, 7978–7982.

88. Gao C., Lin C.H., Lo C.H., Mao S., Wirsching P., Lerner R.A. and Janda K.D. (1997) Making chemistry selectable by linking it to infectivity. *Proc. Natl. Acad. Sci. USA*, **94**, 11777–11782.

89. Leconte A.M., Patel M.P., Sass L.E., McInerney P., Jarosz M., Kung L., Bowers J.L., Buzby P.R.,

Efcavitch J.W. and Romesberg F.E. (2010) Directed evolution of DNA polymerases for next-generation sequencing. *Angew. Chem. Int. Ed. Engl.*, **49**, 5921–5924.

90. Delespaul W., Peeters Y., Herdewijn P. and Robben J. (2015) A novel helper phage for HaloTag-mediated co-display of enzyme and substrate on phage. *Biochem. Biophys. Res. Commun.*, **460**, 245–249.

91. Holliger P. and Oliynyk Z. (2010) Compartmentalized self tagging, US 7.691576 B2.

92. Steele F.R. and Gold L. (2012) The sweet allure of XNA. *Nat. Biotechnol.*, **30**, 624–625.

93. Gardner A.F. and Jack W.E. (1999) Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res.*, **27**, 2545–2553.

94. Sharp P.M. and Li W. (1987) The codon adaptation index-a measure of directional synonymous codon usage bias, and its potential applications. *Nucleic Acids Res.*, **15**, 1281–1295.

95. Williams R., Peisajovich S.G., Miller O.J., Magdassi S., Tawfik D.S. and Griffiths A.D. (2006) Amplification of complex gene libraries by emulsion PCR. **3**, 545–550.

96. Vashishtha A.K. and Konigsberg W.H. (2016) Effect of different divalent cations on the kinetics and fidelity of RB69 DNA polymerase. *Biochemistry*, **55**, 2661–2670.

97. Makarova A. V., Ignatov A., Miropolskaya N. and Kulbachinskiy A. (2014) Roles of the active site residues and metal cofactors in noncanonical base-pairing during catalysis by human DNA polymerase iota. *DNA Repair (Amst).*, **22**, 67–76.

98. Ichida J.K., Horhota A., Zou K., McLaughlin L.W. and Szostak J.W. (2005) High fidelity TNA synthesis by Therminator polymerase. *Nucleic Acids Res.*, **33**, 5219–5225.

99. Bauwens B., Rozenski J., Herdewijn P. and Robben J. (2018) A single amino acid substitution in Therminator DNA polymerase increases incorporation efficiency of deoxyxylonucleotides. *ChemBioChem*, **19**, 2410–2420.

100. Taylor A.I., Pinheiro V.B., Smola M.J., Morgunov A.S., Peak-Chew S., Cozens C., Weeks K.M., Herdewijn P. and Holliger P. (2015) Catalysts from synthetic genetic polymers. *Nature*, **518**, 427–430.

101. Appella D.H. (2009) Non-natural nucleic acids for synthetic biology. *Curr. Opin. Chem. Biol.*, **13**, 687–696.

102. Herdewijn P. and Marliere P. (2009) Toward safe genetically modified organisms through the chemical diversification of nucleic acids. In *Chemistry & biodiversity*.Vol. 6, pp. 791–808.

103. Anosova I., Kowal E.A., Dunn M.R., Chaput J.C., Horn W.D.V. and Egli M. (2016) The structural diversity of artificial genetic polymers. *Nucleic Acids Res.*, **44**, 1007–1021.

104. Meggers E. and Zhang L. (2010) Synthesis and properties of the simplified nucleic acid glycol nucleic acid. *Acc. Chem. Res.*, **43**, 1092–1102.

105. Nielsen P.E. (2004) PNA technology. *Appl. Biochem. Biotechnol. - Part B Mol. Biotechnol.*, **26**, 233–248.

106. Dunn M.R., Otto C., Fenton K.E. and Chaput J.C. (2016) Improving polymerase activity with unnatural substrates by sampling mutations in homologous protein architectures. *ACS Chem. Biol.*, **11**, 1210–1219.

107. Kempeneers V., Vastmans K., Rozenski J. and Herdewijn P. (2003) Recognition of threosyl nucleotides by DNA and RNA polymerases. *Nucleic Acids Res.*, **31**, 6221–6226.

108. Chen J.J., Tsai C.H., Cai X., Horhota A.T., McLaughlin L.W. and Szostak J.W. (2009) Enzymatic primer-extension with glycerol-nucleoside triphosphates on DNA templates. *PLoS One*, **4**, 1–8.

109. Heuberger B.D. and Switzer C. (2008) A pre-RNA candidate revisited: Both enantiomers of flexible nucleoside triphosphates are DNA polymerase substrates. *J. Am. Chem. Soc.*, **130**, 412–413.

110. Wang P., Amato N.J. and Wang Y. (2017) Cytotoxic and mutagenic properties of C3′-epimeric lesions of 2′-deoxyribonucleosides in Escherichia coli cells. *Biochemistry*, **56**, 3725–3732.

111. Schöppe A., Hinz H.J., Rosemeyer H. and Seela F. (1996) Xylose-DNA: comparison of the thermodynamic stability of oligo(2’-deoxyxylonucleotide) and oligo(2’-deoxyribonucleotide) duplexes. *Eur. J. Biochem.*, **239**, 33–41.

112. Jaziri F., Maiti M., Lescrinier E., Marlière P., Pezo V. and Herdewijn P. (2019) Transliteration of a short genetic message from deoxyxylose (dXyloNA) to deoxyribose (DNA) in Escherichia coli. *J. Syst. Chem.*, in press.

113. Houlihan G., Arangundy-Franklin S. and Holliger P. (2017) Exploring the chemistry of genetic information storage and propagation through polymerase Engineering. *Acc. Chem. Res.*, **50**, 1079–1087.

114. Schlötterer C. and Tautz D. (1992) Slippage synthesis of simple sequence DNA. *Nucleic Acids Res.*, **20**, 211–215.

115. Viguera E., Canceill D. and Ehrlich S.D. (2001) Replication slippage involves DNA polymerase pausing and dissociation. *EMBO J.*, **20**, 2587–2595.

116. Staiger N. and Marx A. (2010) A DNA polymerase with increased reactivity for ribonucleotides and C5-modified deoxyribonucleotides. *ChemBioChem*, **11**, 1963–1966.

117. Tsai C.-H., Chen J. and Szostak J.W. (2007) Enzymatic synthesis of DNA on glycerol nucleic acid templates without stable duplex formation between product and template. *Proc. Natl. Acad. Sci. USA*, **104**, 14598–14603.

118. Williams A.A., Darwanto A., Theruvathu J.A., Burdzy A., Jonathan W. and Sowers L.C. (2009) The impact of sugar pucker on base pair and mispair stability. *Biochemistry*, **48**, 11994–12004.

119. Engler C., Gruetzner R., Kandzia R. and Marillonnet S. (2009) Golden gate shuffling: A one-pot DNA shuffling method based on type iIs restriction enzymes. *PLoS One*, **4**, e5553.

120. Grimon D., Gerstmans H., Briers Y. and Lavigne R. (2018) Polynucleotide Shuffling Method, WO2018114980.

121. Rodriguez I., Lazaro J.M., Blanco L., Kamtekar S., Berman A.J., Wang J., Steitz T.A., Salas M. and de Vega M. (2005) A specific subdomain in phi29 DNA polymerase confers both processivity and strand-displacement capacity. *Proc. Natl. Acad. Sci. USA*, **102**, 6407–6412.

122. Harms M.J. and Thornton J.W. (2013) Evolutionary biochemistry: revealing the historical and physical causes of protein properties. *Nat. Rev. Genet.*, **14**, 559–571.

123. Maervoet V.E.T. and Briers Y. (2017) Synthetic biology of modular proteins. *Bioengineered*, **8**, 196–202.

124. Gerstmans H., Criel B. and Briers Y. (2018) Synthetic biology of modular endolysins. *Biotechnol. Adv.*, **36**, 624–640.

125. Vogt S., Stadlmayr G., Stadlbauer K., Sádio F., Andorfer P., Grillari J., Rüker F. and Wozniak-Knopp G. (2018) Stabilization of the CD81 large extracellular loop with de novo disulfide bonds improves its amenability for peptide grafting. *Pharmaceutics*, **10**, 138.

126. Romero P.A. and Arnold F.H. (2009) Exploring protein fitness landscapes by directed evolution. *Nat. Rev. Mol. Cell Biol.*, **10**, 866–876.

127. Voigt C.A., Kauffman S. and Wang Z.G. (2000) Rational evolutionary design: the theory of in vitro protein evolution. *Adv. Protein Chem.*, **55**, 79–160.

128. Bershtein S., Segal M., Bekerman R., Tokuriki N. and Tawfik D.S. (2006) Robustness–epistasis link shapes the fitness landscape of a randomly drifting protein. *Nature*, **444**, 929–932.

129. Ellefson J.W., Gollihar J., Shroff R., Shivram H., Iyer V.R. and Ellington A.D. (2016) Synthetic evolutionary origin of a proofreading reverse transcriptase. *Science*, **352**, 1590–1593.

130. Wong T.S., Roccatano D. and Schwaneberg U. (2007) Steering directed protein evolution: strategies to manage combinatorial complexity of mutant libraries. *Environ. Microbiol.*, **9**, 2645–2659.

131. Pesce D., Lehman N. and Visser J.A.G.M.D.V. (2016) Sex in a test tube : testing the benefits of in vitro recombination. *Philos. Trans. R. Soc. Publ.*, **371**, 20150529.

132. Yano T. and Kagamiyama H. (2001) Directed evolution of ampicillin-resistant activity from a functionally unrelated DNA fragment: A laboratory model of molecular evolution. *Proc. Natl. Acad. Sci. USA*, **98**, 903–907.

133. Rowe L.A., Geddie M.L., Alexander O.B. and Matsumura I. (2003) A comparison of directed evolution approaches using the β-glucuronidase model system. *J. Mol. Biol.*, **332**, 851–860.

134. Drummond D.A., Silberg J.J., Meyer M.M., Wilke C.O. and Arnold F.H. (2005) On the conservative nature of intragenic recombination. *Proc. Natl. Acad. Sci. USA*, **102**, 5380–5385.

135. Jozwiakowski S.K. and Connolly B. a (2011) A modified family-B archaeal DNA polymerase with reverse transcriptase activity. *ChemBioChem*, **12**, 35–37.

136. Riechmann L., Clark M., Waldmann H. and Winter G. (1988) Reshaping human antibodies for therapy. *Nature*, **332**, 323–327.

137. Murphy G.S., Sathyamoorthy B., Der B.S., Machius M.C., Pulavarti S. V., Szyperski T. and Kuhlman B. (2015) Computational de novo design of a four-helix bundle protein-DND_4HB. *Protein Sci.*, **24**, 434–445.

138. Glanville J., Zhai W., Berka J., Telman D., Huerta G., Mehta G.R., Ni I., Mei L., Sundar P.D., Day G.M.R., *et al.* (2009) Precise determination of the diversity of a combinatorial antibody library gives insight into the human immunoglobulin repertoire. *Proc. Natl. Acad. Sci. USA*, **106**, 20216–20221.

139. Spiliotopoulos A., Owen J.P., Maddison B.C., Dreveny I., Rees H.C. and Gough K.C. (2015) Sensitive recovery of recombinant antibody clones after their in silico identification within NGS datasets. *J. Immunol. Methods*, **420**, 50–55.

140. ’t Hoen P.A.C., Jirka S.M.G., Bradley R., Schultes E.A., Aguilera B., Him K., Heemskerk H., Aartsma-rus A., Ommen G.J. Van and Dunnen J.T. Den (2012) Phage display screening without repetitious selection rounds. *Anal. Biochem.*, **421**, 622–631.

141. Ravn U., Didelot G., Venet S., Ng K., Gueneau F., Rousseau F., Calloud S., Kosco-vilbois M. and Fischer N. (2013) Deep sequencing of phage display libraries to support antibody discovery. *Methods*, **60**, 99–110.

142. van Dijk E.L., Auger H., Jaszczyszyn Y. and Thermes C. (2014) Ten years of next-generation sequencing technology. *Trends Genet.*, **30**, 418–426.

143. Wójcik M., Telzerow A., Quax W. and Boersma Y. (2015) High-throughput screening in protein engineering: recent advances and future perspectives. *Int. J. Mol. Sci.*, **16**, 24918–24945.

144. Say S. and Tan H. (2017) Active droplet sorting in microfluidics : a review. *Lab Chip*, **17**, 751–771.

145. Bernath K., Hai M., Mastrobattista E., Griffiths A.D., Magdassi S. and Tawfik D.S. (2004) In vitro compartmentalization by double emulsions: sorting and gene enrichment by fluorescence activated cell sorting. *Anal. Biochem.*, **325**, 151–157.

146. Zinchenko A., Devenish S.R.A., Kintses B., Colin P., Fischlechner M. and Hollfelder F. (2014) One in a million: flow cytometric sorting of single cell-lysate assays in monodisperse picolitre double emulsion droplets for directed evolution. *Anal. Chem.*, **86**, 2526–2533.

147. Sims P.A., Greenleaf W.J., Duan H. and Xie X.S. (2011) Fluorogenic DNA sequencing in PDMS microreactors. *Nat. Methods*, **8**, 575–580.

148. Paegel B.M. and Joyce G.F. (2010) Microfluidic compartmentalized directed evolution. *Chem. Biol.*, **17**, 717–24.

149. Hutchison C.A., Chuang R.-Y., Noskov V.N., Assad-Garcia N., Deerinck T.J., Ellisman M.H., Gill J., Kannan K., Karas B.J., Ma L., *et al.* (2016) Design and synthesis of a minimal bacterial genome. *Science*, **351**, aad6253.

150. Bande O., Abu El Asrar R., Braddick D., Dumbre S., Pezo V., Schepers G., Pinheiro V.B., Lescrinier E., Holliger P., Marlière P., *et al.* (2015) Isoguanine and 5-methyl-isocytosine bases, in vitro and in vivo. *Chem. - A Eur. J.*, **21**, 5009–5022.

151. Pezo V., Schepers G., Lambertucci C., Marlière P. and Herdewijn P. (2014) Probing ambiguous base-pairs by genetic transformation with XNA templates. *ChemBioChem*, **15**, 2255–2258.

# Bibliography

### Articles in internationally reviewed academic journals

Bauwens B., Rozenski J., Herdewijn P., Robben J., 2018, “A single amino acid substitution in Therminator DNA polymerase increases incorporation efficiency of deoxyxylonucleotides”, ChemBiochem, 19(22), p. 2410-2420. (doi.org/10.1002/cbic.201800411) (99)

### Meeting abstracts, presented at international scientific conferences

Bauwens B., Peeters Y., Malczewska K., Herdewijn P., Robben J. (2017), Poster Presentation: Molecular Evolution of a xylonucleic acid polymerase. Gordon Research Conference on Synthetic Biology. Stowe, Vermont, 30 July – 4 Aug 2017.

### Meeting abstracts, presented at other scientific conferences and symposia

Bauwens B., Peeters Y., Malczewska K., Robben J. (2015), Guided evolution of DNA polymerase for the synthesis of artificial nucleic acids. Meeting of the Belgian Society of Biochemistry and Biotechnology, Biochemical Aspects of Evolution. Louvain-la-Neuve, Belgium, 22 May 2015.

This work was supported by the Flanders Innovation and Entrepreneurship (VLAIO), the Rega Institute and KU Leuven.

# Addendum

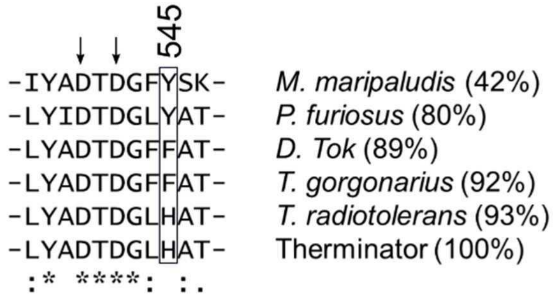

545
-IYADTDGFYSK- M. maripaludis (42%)
-LYIDTDGLYAT- P. furiosus (80%)
-LYADTDGFFAT- D. Tok (89%)
-LYADTDGFFAT- T. gorgonarius (92%)
-LYADTDGLHAT- T. radiotolerans (93%)
-LYADTDGLHAT- Therminator (100%)
:* ****: :.

**Figure S1: Sequence alignment of the active site region of five representative archaeal DNA polymerases compared to the Therminator DNA polymerase.**

A protein BLAST to this polymerase returned 51 archaeal family B DNA polymerases (fragmentary, hypothetical and duplicate protein sequences were excluded), 47 of which belong to the Thermococcaceae family. The catalytic Asp residues D540 and D542, as well as many nearby residues, are either conserved across all polymerases in the dataset (*) allow conserved replacements (:), or are semi-conserved (.). Position 545 is highly variable but is always taken by a large aromatic residue, either His (8/51), Tyr (23/51) or Phe (20/51), with no clear phylogenetic constraints.

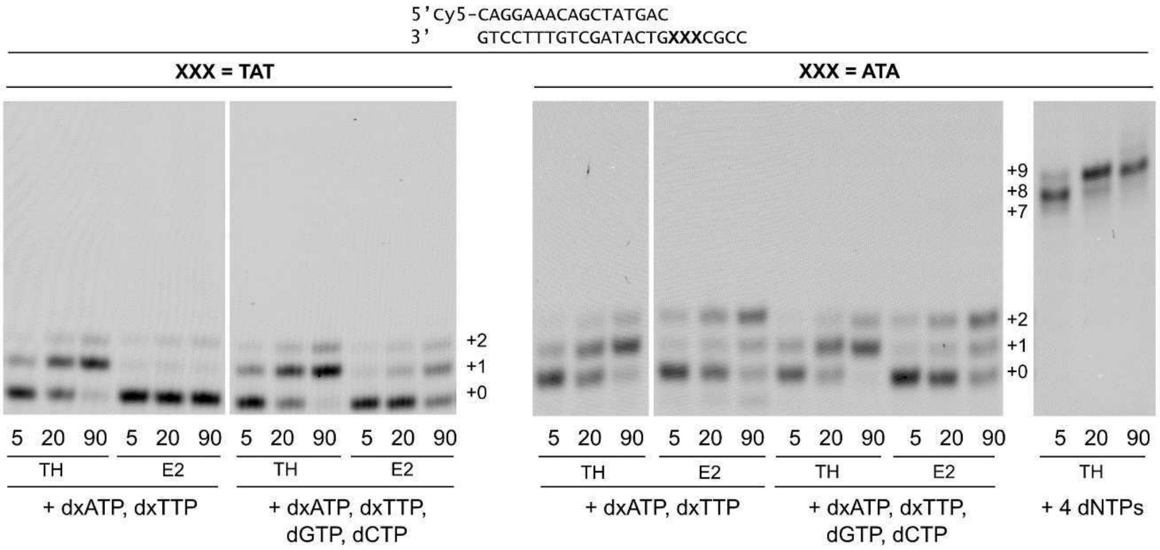
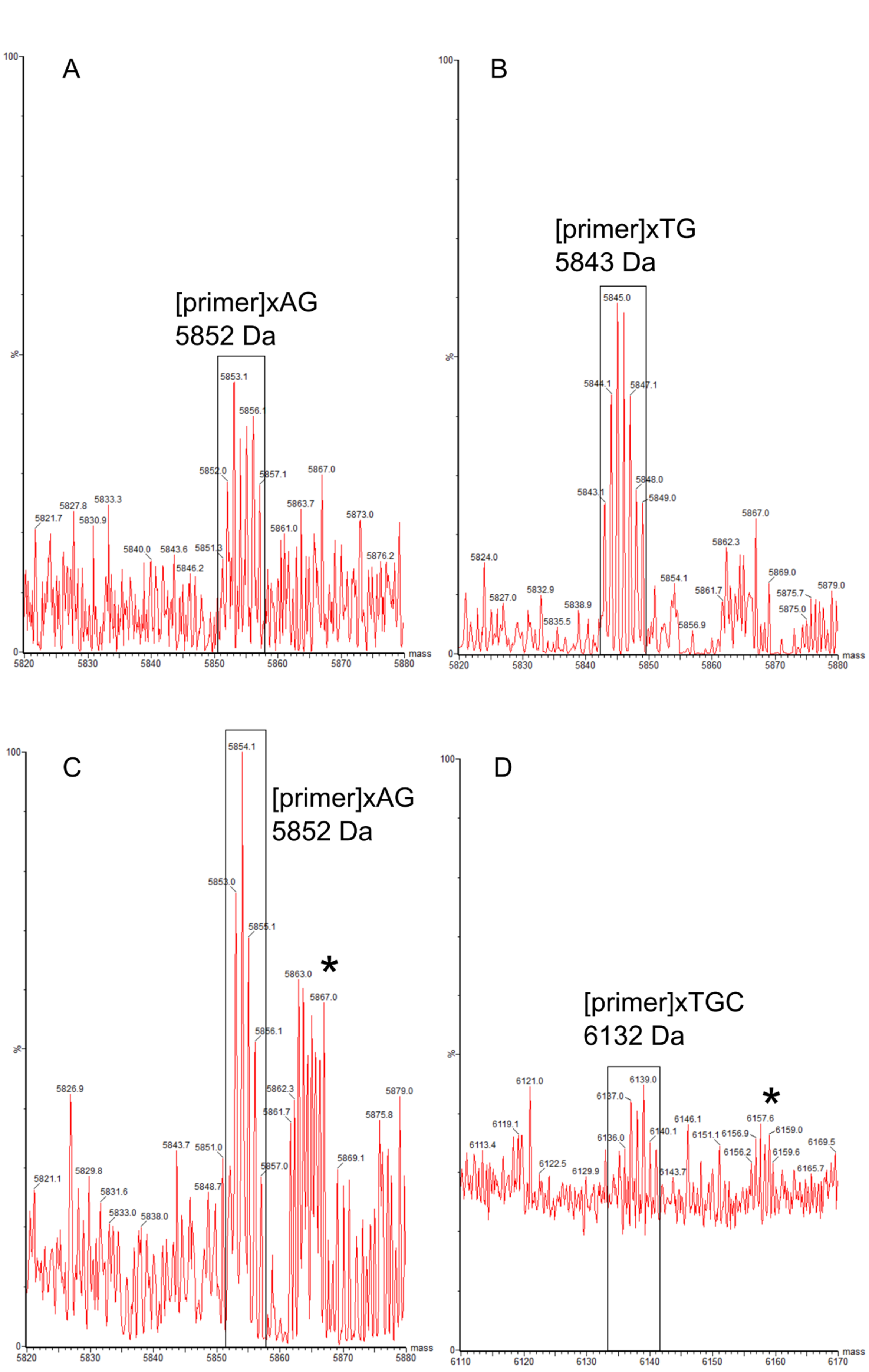


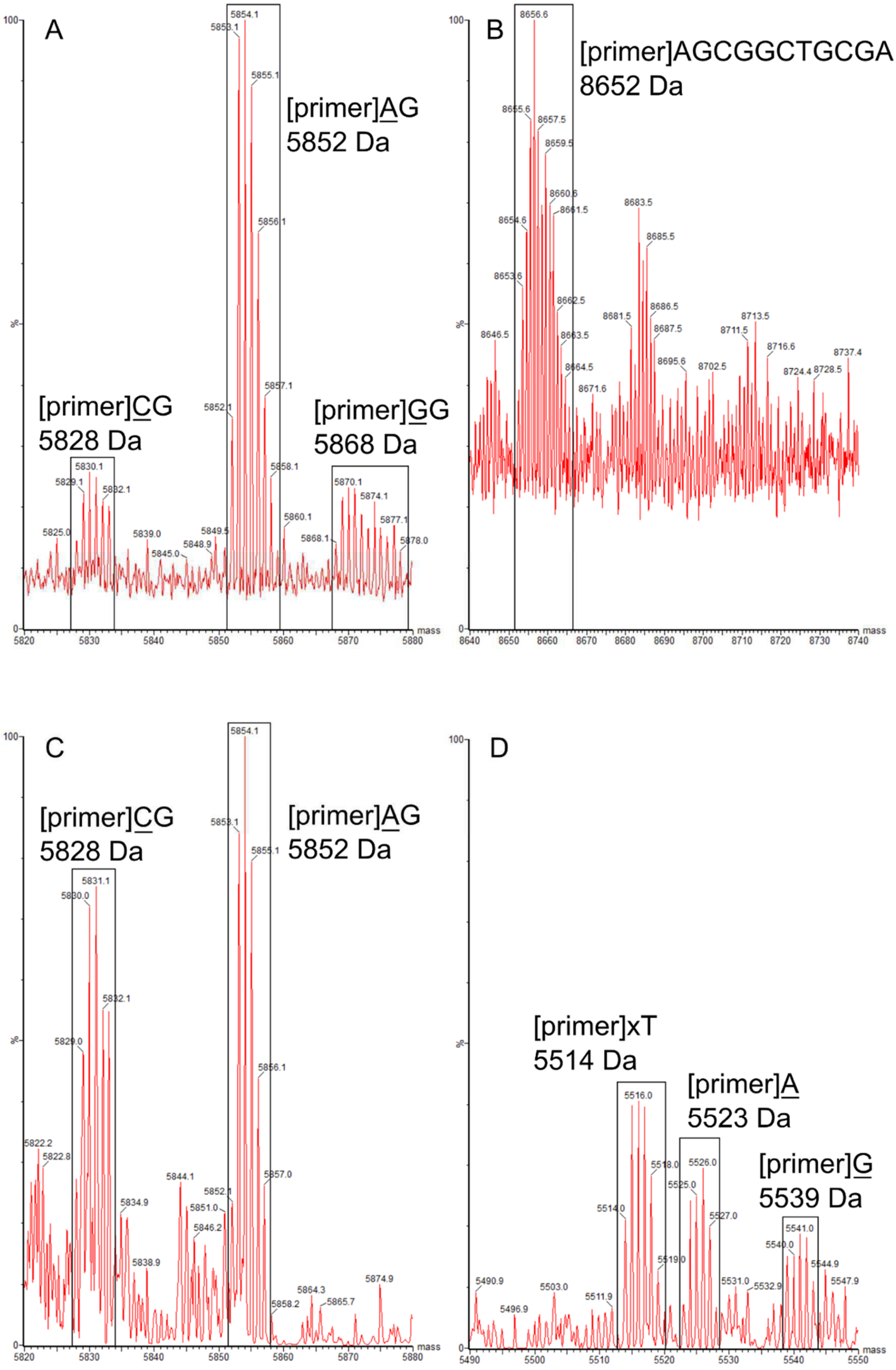
**Figure S2: Primer elongation with mixed dxATP and dxTTP by wildtype Therminator exo$^-$ (TH) and mutant polymerase E2.**

The TAT template (XXX = TAT) was designed to allow incorporation of one dxATP and two dxTTP, the inverse with the ATA template. Reactions ran with dxNTPs in the absence or presence of additional dGTP and dCTP. Samples were taken after 5, 20 and 90 min. P = unextended primer. Right: positive control with natural dNTPs.


A
[primer]xAG
5852 Da
5853.1
5856.1
5852.0
5857.1
5867.0
5827.8
5833.3
5821.7
5830.9
5863.7
5861.0
5873.0
5840.0
5843.6
5851.3
5846.2
5876.2
B
[primer]xTG
5843 Da
5845.0
5844.1
5847.1
5843.1
5848.0
5849.0
5867.0
5862.3
5824.0
5854.1
5869.0
5879.0
5832.9
5861.7
5875.7
5827.0
5838.9
5875.0
5835.5
5856.9
C
[primer]xAG
5852 Da
5854.1
5853.0
5855.1
*
5863.0
5867.0
5856.1
5826.9
5862.3
5879.0
5861.7
5875.8
5843.7
5851.0
5869.1
5829.8
5857.0
5821.1
5848.7
5831.6
5833.0
5838.0
D
[primer]xTGC
6132 Da
6121.0
6139.0
6137.0
*
6146.1
6157.6
6119.1
6159.0
6140.1
6151.1
6156.9
6136.0
6113.4
6169.5
6156.2
6159.6
6122.5
6165.7
6129.9
6143.7
100
%
0
mass
5820
5830
5840
5850
5860
5870
5880
6110
6120
6130
6140
6150
6160
6170

**< Figure S3: Selected parts of electrospray ionization mass spectra of extended oligonucleotide primer CAGGAAACAGCTATGAC.**

Peak heights are normalized for the highest peaks of primer extension products within that sample. Templates 5’-CGCTGCCGCAGTCATAGCTGTTTCCTGCC-3’ and 5’-CGCAGCCGCTGTCATAGCTGTTTCCT GCC-3’ allow extension with TGCGGCAGCG (in presence of d(x)ATP) or AGCGGCTGCG (in presence of d(x)TTP). Peaks labelled with * are m/z to m conversion artefacts that do not correspond to expected elongation products. Expected products are confined by boxes. **A)** MS peaks of +2 extension products using wildtype Therminator in presence of dTTP, dGTP, dCTP and dxATP. Correct extension products have a monoisotopic mass of 5852 Da. **B)** MS peaks of +2 extension products using wildtype Therminator in presence of dATP, dGTP, dCTP and dxTTP. Correct extension products have a monoisotopic mass of 5843 Da. **C)** MS peaks of +2 extension products using Therminator mutant E2 in presence of dTTP, dGTP, dCTP and dxATP. Correct extension products have a monoisotopic mass of 5852 Da. **D)** MS peaks of +3 extension products using Therminator mutant E2 in presence of dATP, dGTP, dCTP and dxTTP. Correct extension products have a monoisotopic mass of 6132 Da.


A
[primer]AG
5852 Da
[primer]CG
5828 Da
[primer]GG
5868 Da
5854.1
5853.1
5855.1
5856.1
5857.1
5852.1
5858.1
5860.1
5868.1
5849.5
5848.9
5845.0
5839.0
5825.0
5829.1
5830.1
5832.1
5870.1
5874.1
5877.1
5878.0
100
%
0
mass
5820
5830
5840
5850
5860
5870
5880
B
[primer]AGCGGCTGCGA
8652 Da
8656.6
8655.6
8657.5
8659.5
8660.6
8661.5
8654.6
8653.6
8662.5
8663.5
8664.5
8646.5
8671.6
8683.5
8685.5
8681.5
8686.5
8687.5
8695.6
8702.5
8713.5
8711.5
8716.6
8724.4
8728.5
8737.4
100
%
0
mass
8640
8650
8660
8670
8680
8690
8700
8710
8720
8730
8740
C
[primer]CG
5828 Da
[primer]AG
5852 Da
5854.1
5853.1
5855.1
5856.1
5857.0
5852.1
5851.0
5846.2
5844.1
5831.1
5830.0
5832.1
5829.0
5822.2
5822.8
5834.9
5838.9
5858.2
5864.3
5865.7
5874.9
100
%
0
mass
5820
5830
5840
5850
5860
5870
5880
D
[primer]xT
5514 Da
[primer]A
5523 Da
[primer]G
5539 Da
5516.0
5518.0
5514.0
5519.0
5526.0
5525.0
5527.0
5531.0
5532.9
5541.0
5540.0
5544.9
5547.9
5490.9
5496.9
5503.0
5511.9
100
%
0
mass
5490
5500
5510
5520
5530
5540
5550

**< Figure S4: Selected parts of electrospray ionization mass spectra.**

Peak heights are normalized for the highest peaks of primer extension products within that sample. Primer sequence: CAGGAAACAGCTATGAC. Templates 5’-CGCTGCCGCAGTCATAGCTGTTTCCTGCC-3’ and 5’-CGCAGCCGCTGTCATAGCTGTTTCCTGCC-3’ allow extension with TGCGGCAGCG (in presence of d(x)ATP) or AGCGGCTGCG (in presence of d(x)TTP) **A)** MS Peaks of +2 extension products using wildtype Therminator in presence of dATP, dGTP and dCTP. Extension products with a misincorporated dC, dA or dG (underlined) have monoisotopic masses of 5828, 5852 and 5868 Da, respectively. **B)** MS Peaks of +11 extension products using wildtype Therminator in presence of dATP, dGTP, dCTP and dTTP. Correct extension products (extended with an untemplated dA) have a monoisotopic mass of 8652 Da. **C)** MS Peaks of +2 extension products using Therminator mutant E2 in presence of dGTP, dCTP and dATP. Extension products with a misincorporated dC or dA (underlined) have monoisotopic masses of 5828 and 5852 Da, respectively. **D)** MS Peaks of +1 extension products using wildtype Therminator in presence of dATP, dGTP, dCTP and dxTTP. Correct extension products (xT) have a monoisotopic mass of 5514 Da, those with misincorporated dA or dG have 5523 and 5539 Da.

```
   1  CGAGGGCAAA CCATGGCTCA CCATCATCAT CATCATTCTT CTGGTGTAGA
  51  TCTGGGTACC GAGAACCTGT ACTTCCAATC CATGATCCTG GACACCGACT
 101  ACATCACCGA AAACGGTAAA CCGGTTATCC GTGTTTTCAA AAAAGAAAAC
 151  GGTGAATTCA AAATCGAATA CGACCGTACC TTCGAACCGT ACTTCTACGC
 201  TCTGCTGAAA GACGACTCTG CAATTGAAGA CGTTAAAAAA GTTACCGCTA
 251  AACGTCACGG TACCGTTGTT AAAGTTAAAC GTGCTGAAAA AGTTCAGAAA
 301  AAATTCCTGG GCCGGCCGAT CGAAGTTTGG AAACTGTACT TCAACCACCC
 351  GCAGGACGTT CCGGCTATCC GTGACCGTAT CCGTGCTCAC CCCGCGGTTG
 401  TTGACATCTA CGAATACGAC ATCCCGTTCG CTAAACGTTA CCTGATCGAC
 451  AAAGGTCTGA TCCCGATGGA AGGTGACGAA GAACTGACCA TGCTGGCTTT
 501  CGCTATCGCT ACCCTGTACC ACGAAGGTGA AGAATTCGGT ACGGGTCCCA
 551  TCCTGATGAT CTCTTACGCT GACGGTTCTG AAGCTCGTGT TATCACCTGG
 601  AAAAAAATCG ACCTGCCGTA CGTCGACGTT GTTTCTACCG AAAAAGAAAT  SalI re-cloning
 651  GATCAAACGT TTCCTGCGTG TTGTTCGTGA AAAAGACCCG GACGTTCTGA  site
 701  TCACCTACAA CGGTGACAAC TTCGACTTCG CTTACCTTAA GAAACGTTGC
 751  GAAGAACTGG GTATCAAATT CACCCTGGGT CGTGACGGTT CTGAACCGAA
 801  AATACAGCGT ATGGGTGACC GTTTCGCTGT TGAAGTTAAA GGTCGTATCC
 851  ACTTCGACCT GTACCCGGTT ATCCGTCGTA CCATCAACCT GCCGACCTAC
 901  ACCCTGGAAG CTGTTTACGA AGCTGTTTTC GGTAAACCTA AGGAAAAAGT
 951  TTACGCTGAA GAAATCGCTC AGGCTTGGGA ATCTGGTGAA GGTCTGGAAC
1001  GTGTTGCTCG TTACTCTATG GAAGACGCTA AAGTTACCTA CGAACTGGGT
1051  CGTGAATTCT TCCCGATGGA AGCTCAGCTG TCTCGTCTGA TCGGTCAATC
1101  TCTGTGGGAC GTTTCTCGTT CTTCTACCGG TAACCTGGTT GAATGGTTCC
1151  TGCTGCGTAA AGCTTACAAA CGTAACGAAC TGGCTCCGAA CAAACCGGAC
1201  GAACGTGAAC TGGCGCGCCG TCGTGGTGGT TACGCCGGCG GTTACGTTAA
1251  AGAACCGGAA CGTGGTCTGT GGGACAACAT CGTTTACCTG GACTTCCGTT
1301  CTCTGTACCC GTCTATCATC ATCACCCACA ACGTTTCTCC GGACACCCTG
1351  AACCGTGAAG GTTGCAAAGA ATACGACGTC GCTCCGGAAG TTGGTCACAA
1401  ATTCTGCAAA GACTTCCCGG GTTTCATCCC GTCTCTGCTG GGTGACCTGC
1451  TGGAAGAACG TCAGAAAATC AAACGTAAAA TGAAAGCTAC CGTTGACCCG
1501  CTGGAAAAAA AACTGCTGGA CTACCGTCAG CGTTTAATTA AAATCCTGGC
1551  TAACTCTTTC TACGGTTACT ACGGTTACGC TAAAGCTCGT TGGTACTGCA
1601  AAGAATGCGC TGAATCTGTT ACCGCTTGGG GTCGTGAATA CATCGAAATG
1651  GTTATCCGTG AACTGGAGGA GAAATTCGGT TTCAAAGTTC TGTACGCTGA  region A
1701  CACCGACGGT CTGCACGCTA CCATCCCGGG TGCTGACGCT GAAACCGTTA  position H545
1751  AAAAAAAAGC TAAAGAATTC CTGAAATACA TCAACCCGAA ACTGCCCGGG
1801  CTGCTGGAAC TGGAATACGA AGGTTTCTAC GTTCGTGGTT TCTTCGTTAC
1851  CAAAAAAAAA TACGCTGTTA TCGACGAAGA AGGTAAAATC ACCACCCGTG  region M
1901  GTCTGGAAAT CGTTCGTCGT GACTGGTCTG AAATCGCTAA AGAAACCCAG
1951  GCTCGTGTCC TCGAGGCTAT CCTGAAACAC GGTGACGTTG AAGAAGCTGT
2001  TCGTATCGTT AAAGAAGTTA CCGAAAAACT GTCTAAATAC GAAGTTCCGC
2051  CGGAAAAACT AGTTATCCAC GAACAGATCA CCCGTGACCT GCGTGACTAC
2101  AAAGCTACCG GTCCGCACGT CGCTGTTGCT AAACGTCTGG CTGCTCGTGG
2151  TGTTAAAATC CGTCCGGGTA CCGTTATCTC TTACATCGTT CTGAAAGGTT
2201  CTGGTCGTAT CGGTGACCGG GCCATCCCGG CCGACGAATT CGACCCGACC
2251  AAACACCGTT ACGACGCTGA ATACTACATC GAAAACCAGG TTCTGCCTGC  PstI re-cloning
2301  AGTTGAACGT ATCCTGAAAG CTTTCGGTTA CCGTAAAGAA GACCTTCGTT  site
2351  ACCAGAAAAC CAAACAGGTT GGTCTGGGTG CTTGGCTGAA AGTTAAAGGT
2401  AAAAAATAAT AAGGATCCGA CTGTGAAGTG AAAAATGGCG CACATTGTGC
2451  GACATTTTTT TTGTCTGCCG TTTACCGCTA CTGCGTCACG GATCTCCACG
2501  CGCCCTGTAG CGGCGCATTA AGCGCGGCGG GTGTGGTGGT TACGCGCAGC
2551  GTGACCGCTA CACTTGCCAG CGCCCTAGCG CCCGCTCCTT TCGCTTTCTT
2601  CCCTTCCTTT CTCGCCACGT TCGCCGGCTT TCCCCGTCAA GCTCTAAATC
2651  GGGGGCTCCC TTTAGGGTTC CGATTTAGTG CTTTACGGCA CCTCGACCCC
2701  AAAAAACTTG ATTAGGGTGA TGGTTCACGT AGTGGGCCAT CGCCCTGATA
2751  GACGGTTTTT CGCCCTTTGA CGTTGGAGTC CACGTTCTTT AATAGTGGAC
2801  TCTTGTTCCA AACTGGAACA ACACTCAACC CTATCTCGGT CTATTCTTTT
2851  GATTTATAAG GGATTTTGCC GATTTCGGCC TATTGGTTAA AAAATGAGCT
2901  GATTTAACAA AAATTTAACG CGAATTTTAA CAAAATATTA ACGCTTACAA
2951  TTTCAGGTGG CACTTTTCGG GGAAATGTGC GCGGAACCCC TATTTGTTTA  CST primer site 1
3001  TTTTTCTAAA TACATTCAAA TATGTATCCG CTCATGAGAC AATAACCCTG
3051  ATAAATGCTT CAATAATATT GAAAAAGGAA GAGTATGAGT ATTCAACATT
3101  TCCGTGTCGC CCTTATTCCC TTTTTTGCGG CATTTTGCCT TCCTGTTTTT  CST primer site 2
3151  GCTCACCCAG AAACGCTGGT GAAAGTAAAA GATGCTGAAG ATCAGTTGGG
3201  TGCACGAGTG GGTTACATCG AACTGGATCT CAACAGCGGT AAGATCCTTG
3251  AGAGTTTTCG CCCCGAAGAA CGTTTTCCAA TGATGAGCAC TTTTAAAGTT
3301  CTGCTATGTG GCGCGGTATT ATCCCGTATT GACGCCGGGC AAGAGCAACT
3351  CGGTCGCCGC ATACACTATT CTCAGAATGA CTTGGTTGAG TACTCACCAG
```

```
3401  TCACAGAAAA GCATCTTACG GATGGCATGA CAGTAAGAGA ATTATGCAGT
3451  GCTGCCATAA CCATGAGTGA TAACACTGCG GCCAACTTAC TTCTGACAAC
3501  GATCGGAGGA CCGAAGGAGC TAACCGCTTT TTTGCACAAC ATGGGGGATC
3551  ATGTAACTCG CCTTGATCGT TGGGAACCGG AGCTGAATGA AGCCATACCA
3601  AACGACGAGC GTGACACCAC GATGCCTGTA GCAATGGCAA CAACGTTGCG
3651  CAAACTATTA ACTGGCGAAC TACTTACTCT AGCTTCCCGG CAACAATTGA
3701  TAGACTGGAT GGAGGCGGAT AAAGTTGCAG GACCACTTCT GCGCTCGGCC
3751  CTTCCGGCTG GCTGGTTTAT TGCTGATAAA TCTGGAGCCG GTGAGCGTGG
3801  CTCTCGCGGT ATCATTGCAG CACTGGGGCC AGATGGTAAG CCCTCCCGTA
3851  TCGTAGTTAT CTACACGACG GGGAGTCAGG CAACTATGGA TGAACGAAAT
3901  AGACAGATCG CTGAGATAGG TGCCTCACTG ATTAAGCATT GGTAGGAATT
3951  AATGATGTCT CGTTTAGATA AAAGTAAAGT GATTAACAGC GCATTAGAGC
4001  TGCTTAATGA GGTCGGAATC GAAGGTTTAA CAACCCGTAA ACTCGCCCAG
4051  AAGCTAGGTG TAGAGCAGCC TACATTGTAT TGGCATGTAA AAAATAAGCG
4101  GGCTTTGCTC GACGCCTTAG CCATTGAGAT GTTAGATAGG CACCATACTC
4151  ACTTTTGCCC TTTAGAAGGG GAAAGCTGGC AAGATTTTTT ACGTAATAAC
4201  GCTAAAAGTT TTAGATGTGC TTTACTAAGT CATCGCGATG GAGCAAAAGT
4251  ACATTTAGGT ACACGGCCTA CAGAAAAACA GTATGAAACT CTCGAAAATC
4301  AATTAGCCTT TTTATGCCAA CAAGGTTTTT CACTAGAGAA TGCATTATAT
4351  GCACTCAGCG CAGTGGGGCA TTTTACTTTA GGTTGCGTAT TGGAAGATCA
4401  AGAGCATCAA GTCGCTAAAG AAGAAAGGGA AACACCTACT ACTGATAGTA
4451  TGCCGCCATT ATTACGACAA GCTATCGAAT TATTTGATCA CCAAGGTGCA
4501  GAGCCAGCCT TCTTATTCGG CCTTGAATTG ATCATATGCG GATTAGAAAA
4551  ACAACTTAAA TGTGAAAGTG GGTCTTAAAA GCAGCATAAC CTTTTTCCGT
4601  GATGGTAACT TCACTAGTTT AAAAGGATCT AGGTGAAGAT CCTTTTTGAT
4651  AATCTCATGA CCAAAATCCC TTAACGTGAG TTTTCGTTCC ACTGAGCGTC
4701  AGACCCCGTA GAAAAGATCA AAGGATCTTC TTGAGATCCT TTTTTTCTGC
4751  GCGTAATCTG CTGCTTGCAA ACAAAAAAAC CACCGCTACC AGCGGTGGTT
4801  TGTTTGCCGG ATCAAGAGCT ACCAACTCTT TTTCCGAAGG TAACTGGCTT
4851  CAGCAGAGCG CAGATACCAA ATACTGTTCT TCTAGTGTAG CCGTAGTTAG
4901  GCCACCACTT CAAGAACTCT GTAGCACCGC CTACATACCT CGCTCTGCTA
4951  ATCCTGTTAC CAGTGGCTGC TGCCAGTGGC GATAAGTCGT GTCTTACCGG
5001  GTTGGACTCA AGACGATAGT TACCGGATAA GGCGCAGCGG TCGGGCTGAA
5051  CGGGGGGTTC GTGCACACAG CCCAGCTTGG AGCGAACGAC CTACACCGAA
5101  CTGAGATACC TACAG
```

**Figure S5: Annotated plasmid sequence.**

Complete sequence of the Therminator protein with His-tag fusion in the pASK plasmid. Targeted residues in the library regions A (active site) and M (minor groove binding region) have been underlined. The codon for residue 545 is boxed. Re-cloning primers for mutagenized part of Therminator were 5’CCGTACGTCGACGTTGTTTC3’ and 5’ACGTTCAACTGCAGGCAG3’, containing a SalI and PstI restriction site, respectively. Re-cloning primers for non-mutagenized part of Therminator and vector backbone were 5’GTTCTGCCTGCAGTTGAAC3’ and 5’GAAACAACGTCGACGTACGG3’, containing a PstI and SalI restriction site, respectively. These primers were used to produce PCR fragments, which are digested by SalI and PstI restriction enzymes and ligated using T4 ligase (Thermo Scientific). Restriction binding sites are underlined. CST primer-binding sites are underlined, the minimal region for incorporation is boxed. Primer for CST round 1 was 5’CAAATAGGGGTTCCGCGCACATTTCCCC3’. Primer for CST round 2 was 5’GAGCAAAAACAGGAAGGCAAAATGCCGC3’.