Sampling the Functional Sequence Neighbourhood of Phi29 DNA Polymerase for XNA Synthesis

Author:

Paola Handal Márquez

A dissertation submitted in partial fulfilment

of the requirements for the degree of

Master of Philosophy in Structural and Molecular Biology

University College London

April 2019

Declaration of Authorship

I, Paola Handal Marquez confirm that the work presented in this thesis titled, 'Sampling the Functional Sequence Neighbourhood of Phi29 DNA Polymerase for XNA Synthesis' is my own. Where information has been derived from other sources, I confirm that this has been indicated in the thesis.

Signed: _____

Date: _____

Abstract

MPhil in Structural and Molecular Biology

Sampling the Functional Sequence Neighbourhood of Phi29 DNA Polymerase for XNA Synthesis

by Paola Handal Marquez

Xenobiotic nucleic acids (XNAs) are unnatural analogues of DNA (and RNA) of great biotechnological and pharmaceutical interest. Their chemical synthesis is highly challenging and expensive, rendering the discovery of functional XNAs impractical. Enzymatic synthesis of XNA is a viable alternative but is limited by the low efficiency and substrate selectivity of the currently available engineered XNA polymerases. A better understanding of the functional space of polymerases is needed to improve their engineering.

Directed evolution techniques were implemented to map the sequence-function relationships of phi29 DNA polymerase (phi29 DNAP), a small and well-characterised polymerase with remarkable processivity and limited HNA (1,5-anhydrohexitol nucleic acid) synthesis activity, in the process of evolving a more efficient HNA polymerase. Diversity in phi29 DNAP was introduced through insertions and deletions, multiple-site saturation mutagenesis and random mutagenesis, targeting different subdomains of the enzyme. Libraries were partitioned through compartmentalised self-tagging (CST), a functional selection platform for XNA synthetases. A single round of selection did not substantially alter the overall activity of libraries for HNA synthesis, and variants isolated in screening were of comparable activity to the wild-type enzyme.

Nevertheless, the results give insight into the functional landscape of phi29 DNAP. Deep sequencing of the libraries used in selection was used to quantify the functional consequence of sequence variation and investigate signs of epistasis for HNA synthesis. Variants that were enriched more than the wild-type, with potentially enhanced HNA synthesis activity, as well as variants that were significantly depleted were identified, both of which contribute to our understanding of the functional sequence space of phi29 DNAP.

Altogether, the directed evolution and deep mutational scanning approaches implemented could be used to develop a new generation of more efficient XNA polymerases.

Impact Statement

Xeno-nucleic acids (XNAs), chemical analogues of nature's only genetic polymers (DNA and RNA), are not only able to store and propagate genetic information, but also broaden the chemical diversity of genetic polymers and their biotechnological and pharmaceutical applications. Still, the chemical synthesis of modified nucleic acids remains highly challenging and expensive. The enzymatic synthesis of XNA has become a more promising alternative but is still limited by the low efficiency and lack of selectivity of the current generation of engineered XNA polymerases.

Directed evolution helps bypass our lack of understanding of the sequence-function relationship of proteins and facilitate their engineering. This approach is particularly relevant to XNA polymerase engineering as there is still a lack of structural and mechanistic information required for engineering approaches involving a more rational design. A lot of progress in the field has been achieved, but there is still a huge gap to bridge to begin exploiting the vast applications of xenobiotic nucleic acids.

In the process of sampling functional variants through directed evolution a lot of useful information that can help us further understand the functional landscape of proteins can be obtained, but it is typically ignored. Instead of focusing solely on isolating the most active variants, deep mutational scanning can help get a better insight into the functional space of polymerases to further improve their engineering and potentially translate modifications to other polymerases or other unnatural substrates.

By combining directed evolution and mutational scanning, this project hopes to generate a platform that can be used to not only generate more efficient XNA polymerases but also to contribute to our understanding of the functional sequence space of polymerases to guide further engineering.

Acknowledgements

Contents

List of Figures

List of Tables

List of Abbreviations

CST	Compartmentalised Self-Tagging
dH_2O	Distilled H_2O
DNA	Deoxyribonucleic Acid
dNTP	Deoxyribonucleotide Triphosphate
D12A	D12A Phi29 DNA Polymerase
D12A THR	D12A Thermostabilised Phi29 DNA Polymerase
EDTA	Ethylenediaminetetraacetic acid
epPCR	Error-Prone Polymerase Chain Reaction
HEPES	4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid
HNA	1,5-Anhydrohexitol Nucleic Acid
hNTP	1,5-Anhydrohexitol Nucleotide Triphosphate
InDel	Insertion or Deletion
MI	Mutual Information
MSA	Multiple Sequence Alignment
NEB 10B	*E. coli* Strain Neb 10-*B*
NGS	Next Generation Sequencing
OD600	Optical Density of a Cell Suspension at a Wavelength of 600 nm
PCR	Polymerase Chain Reactions
PNK	Polynucleotide Kinase
RNA	Ribonucleic Acid
ssDNA	Single-Stranded DNA
TPR2	Terminal Protein Region 2
T7 Express	*E. coli* Strain T7 Express
XNA	Xeno-Nucleic Acids

1. Introduction

The chemical constitution and molecular structure of deoxyribonucleic acid (DNA) allow the storage and propagation of genetic information, essential biological processes for life and evolution [1]. The diversification of the building blocks of natural genetic polymers have enabled the generation of a wide variety of xeno-nucleic acids (XNAs) [2], some of which are also capable of information storage and transfer [3]. Broadening the chemical diversity of genetic polymers has in turn expanded their biotechnological and pharmaceutical applications as well as their use in nanotechnology and material sciences [4, 5, 6]. Continuing to surpass nature's 'limitations' by exploring the XNA world, will not only broaden and refine the applications of XNA but will also help improve our understanding of the origin of life and events that gave rise to nature's genetic information system.

1.1 XNA Diversity and Structure

Nucleic acids are polymers of nucleotides, whose basic structure is composed of an internucleoside phosphodiester backbone, ribofuranose ring and a nucleobase. XNAs refer to nucleic acids with chemical modifications to any of these three chemical moieties [4] and typically display altered physico-chemical properties [3].

For instance, chemical modifications to the nucleobase at the N7 or N9 in purines [7], fluorination of pyrimidines [8] and benzene ring expansions [9] are possible. Some of these nucleobase modifications can result in altered physico-chemical properties that not only expand their functionality but also give rise to novel base-pairings that expand the genetic alphabet and potential information storage density [3]. Modifications at the 2' position of the ribofuranose sugar, such as the incorporation of 2'O-methyl [10l], 2′-F-ANA [11], and 2'-fluoro [12] moieties, have also been investigated. Other sugar modifications include the 'locking' of the ribofuranose ring with a methylene bridge between the 2'O and C4 in LNA (2'-O,4'-C-methylene-b-D-ribonucleic acid) [13, 14] and the replacement of the ribofuranose sugar moiety with six-membered rings such as in HNA (1,5-anhydrohexitol nucleic acid) [15] and CeNA (cyclohexenyl nucleic acid) [16] or a threose sugar in TNA (a-L-threofuranosyl nucleic acids)

[17]. Sugar-modified nucleic acids such as HNA, CeNA and TNA are quite interesting as they retain Watson-Crick base-pairing potential with DNA and RNA and are therefore available for Darwinian evolution in cross-chemistry platforms [18, 3].

XNAs with phosphodiester backbone modifications, such as in GNA (glycerol nucleic acids) [19] and FNA (flexible nucleic acids) [20], where the phosphate is linked to the base, are also possible. Other backbone modifications include the replacement of the alpha oxygen with a sulphur or borano group, giving rise to phosphorothioate [21] and boranophosphate [22] nucleotides, respectively. The complete replacement of the backbone with N-(2-aminoethyl)-glycine in PNA (peptide nucleic acid) have also been generated [23].

1.2 XNA Applications

The physico-chemical properties of natural nucleic acids have allowed the generation of DNA- and RNA-tools with numerous clinical and biotechnological applications. Recent advances in XNA characterisation and synthesis have expanded these applications given that XNA-based technologies not only can display properties of their natural counterpart but are also much more chemically and biologically stable, can have higher specificity and possess a far greater chemical diversity, making them much better candidates for their use in research and industry [5].

Hybridization probes, for instance, are one of the numerous applications of nucleic acids in which XNA outperforms their natural counterpart. These probes consist of single stranded oligonucleotide sequences labelled with radioactive (^{32}P), non-radioactive (biotin or digoxigenin) or chemiluminescent labels that hybridize to complementary nucleic acid sequences with high selectivity and sensitivity [24]. DNA probes have proven useful for *ex-vivo* applications such as filter hybridization reactions for the detection of infectious diseases [25] or *in vitro* applications such as real-time PCR genotyping [26]. Nonetheless, natural nucleic acid-based probes have numerous limitations when used *in vivo* [5, 27]. Molecular beacons (MBs), for instance, are probes consisting of a hairpin structure with the target sequence in the loop and a quencher and fluorophore in close proximity on each end of the hairpin stem.

Once the probe hybridizes to the target sequence, the quencher separates from the fluorophore giving off a fluorescent signal [28]. DNA MBs *in vivo* can give off false positives when the stem portion binds off-target genomic DNA or when bound by DNA/RNA binding proteins inside cells [5]. Molecular beacons of synthetic nucleic acid analogues allow circumventing some of these issues. For instance, locked nucleic acid (LNA)-based MBs have a similar charge to DNA MBs, but different ribose structure and steric properties, which allows to preserve the solubility and specificity of natural MBs while avoiding the binding of endogenous DNA/RNA-binding proteins [27]. LNA MBs also show increased thermostability, higher affinity and greater selectivity, which improve their performance during *in vivo* gene expression and SNP analysis, compared to DNA MBs [27].

Another good example of improving the performance of nucleic acid-based technologies with modified nucleic acids is aptamers. Aptamers are specific single-stranded nucleic acid sequences that selectively bind, through 3D conformational complementarity, a variety of molecules such as sugars, proteins, whole cells, small metal ions and small organic molecules [29, 30]. These molecules can be isolated through *in vitro* selection approaches or SELEX (Systematic Evolution of Ligands by Exponential Enrichment), from a larger pool of sequences, enriching those that selectively bind targets and depleting those that do not. [31, 32]. DNA and RNA aptamers are of particular clinical and biotechnological interest as they bind their targets with high specificity and affinity, have high chemical synthesis and stability, are non-immunogenic, are cheap and quick to produce and can be conjugated to numerous molecules such as nanoparticles, imaging agents or therapeutics, providing them with numerous applications [30, 33]. Still, their *in vivo* applications are limited by their high susceptibility to nuclease degradation, easy filtration and excretion from the body, and insufficient binding affinity and specificity in the *in vivo* complex environment [29, 30]. Aptamer stability can be enhanced through structural modifications that reduce solvent exposure, however modifying aptamer structure tends to interfere with their functionality as their binding affinity often depends on the 3D structure complementarity to their target [29]. A more suitable alternative is to chemically modify aptamers post-selection [34] or use modified nucleoside triphosphates during selection [35,

36]. Modified phosphate backbones, such as the altered backbone orientation (3′ to 2′ phosphoramidate linkage) of threose nucleic acid (TNA) [37] or 2′- modifications to RNA aptamers such as in 2′OMe or 2'F-RNA, provide XNAs with remarkable nuclease resistance and biostability in *in vivo* conditions such as blood serum, while retaining equal or higher selectivity and affinity for their targets compared to DNA/RNA aptamers [38, 39]. Additionally, the broad structural diversity and properties of XNAs, such as the increased hydrophobicity of fGmH (2′-F-dG, 2′-OMe-dA/dC/dU) [39], or the preference of FNA (2'-deoxy-2'-fluoro-ribonucleic acid) for C3'-endo sugar puckering (A-conformation) [40] and its higher hydrogen bonding strength [41], allow them to fold into diverse motifs and efficiently bind a wider variety of targets, surpassing the limited binding abilities and stability of DNA/RNA aptamers.

XNA has also expanded the spectrum of catalytic activities of DNA and RNA. In nature, RNA-based enzymes (ribozymes) found in numerous pathogens are able to catalyse reactions such as phosphodiester bond cleavage [42, 43] and aminoacyl transfer reactions [44]. Through *in vitro* selection approaches it has been possible to isolate multiple nucleic acid catalysts such as ligase ribozymes [45], RNA Polymerase ribozymes [46, 47] and DNAzymes that cleave RNA [48] and DNA [49]. Nonetheless, nucleic acid-based catalysts still possess numerous limitations and are constrained by the reduced structural and physicochemical properties of natural nucleic acids. With the help of techniques such as X-Selex, XNA endonucleases and ligases have been discovered, expanding dramatically the structural space and potential applications of nucleic acid-based technologies [36].

A final application of XNA-based technologies is its potential usage as a bio-containment measure. The engineering of organisms towards their dependence of synthetic building blocks could be implemented as a biosafety measure, due to the absence of XNAs in nature and the lack of compatibility of replication machineries in other organisms [50].

1.3 Limitations of XNA synthesis

Although XNA has numerous biotechnological and pharmaceutical applications, the solid-phase chemical synthesis of oligonucleotides with modified nucleotides is still highly challenging, laborious and expensive and imposes a

limitation on the polymer length [85]. Introduction of chemical modifications post-chemical synthesis of DNA/RNA polymers is also possible but is limited in density and labelling efficiency [85]. The enzymatic synthesis of XNA is slowly becoming a more promising alternative, but it is still limited by the high substrate specificity of natural polymerases, rendering the discovery of functional XNAs very inefficient and costly [36, 52].

Protein engineering techniques are becoming increasingly more important with applications that extend to numerous fields of biotechnology [53]. Consequently, recent advances in protein engineering have shown a lot of potential in surpassing the limitations of DNA polymerases when subjected to XNA processing [54, 51]. The classical method of protein engineering, known as “rational” design, using site-directed mutagenesis [55], has allowed the identification of residues involved in polymerase substrate specificity as well as variants with altered substrate recognition such as the archeal 9°N A485L variant, also known as Therminator DNA polymerase, with abolished exonuclease activity that show enhanced synthesis of one of the most structurally diverse XNAs, TNA [56, 57].

Nonetheless, this approach requires a high level of understanding of the structure, mechanisms and dynamics of polymerases and mutations cannot always be translated across polymerase families due to their structural and sequence diversity [58], making the sampling of multiple mutations or combination of mutations time consuming and inefficient. A more efficient approach for sampling thousands of protein variants simultaneously is through directed evolution, an approach that mimics nature’s way of evolution under desired selection pressures to tailor the activities, properties and substrate specificities of proteins [59].

1.4 Directed Evolution for protein engineering

Directed evolution is an engineering approach that bypasses our lack of understanding of the sequence-function relationship of proteins by subjecting proteins to rounds of random or semi-rational mutagenesis and artificial selection for a desired function [60]. Through directed evolution, the selection pressure should enrich functionally active variants; and naturally reduce the fraction of the population composed by inactive or poorly active variants.

enriched populations can then be subjected to subsequent rounds of selection and screening or can be further diversified and enriched until variants with improved activity can be identified [59]. Selection methods that couple genotype to phenotype along with improvements in sequencing have enabled the high-throughput screening of libraries with sampling capacities reaching into the millions [61]. This method of screening, also known as deep mutational scanning, allows the direct observation and quantification of the input and post-selection populations, where their respective frequencies can be used as a measure of function to map functional landscapes that can further aid protein engineering [61]. Deep mutational scanning for function such as novel ligand binding, catalysis or cellular fitness under selection pressures have uncovered unexpected mutations or combinations of mutations near or distant active sites and mutations with significant impacts on thermostability and enzymatic activity [62].

1.4.1 Gene diversification methods

There are generally two gene diversification approaches used in directed evolution: random mutagenesis and semi-rational design. Random mutagenesis techniques are typically used when there is a lack of structure-function relationship information and are thus implemented to sample a breadth of the sequence space. Semi-rational design techniques are used when functionally relevant residues are known and one wants to sample them in depth to obtain variants with the desired phenotype [63]. Techniques for random mutagenesis can be subdivided into *in vivo* and *in vitro* approaches, where the former typically results in deleterious mutations to the host genome and low mutation rates [63], making the latter a more preferable alternative. *In vitro* approaches commonly used involve error-prone PCR (epPCR), where random mutations are introduced during PCR by DNA polymerases with low fidelity, such as Taq DNA polymerase, under conditions that decrease their substrate specificity further, such as increased $MgCl_2$ concentrations, addition of $MnCl_2$, increased concentrations of Taq DNA polymerase, increased extension times, and non-equimolar or biased ratios of nucleotide concentrations as well as mutagenic nucleotides [64]. This approach, however, tends to result in a bias towards A to G and T to C transitions [65]. Numerous adaptations to the original epPCR approach described by Leung *et al.,* have been developed to reduce the

imbalanced mutational spectrum including using unbalanced concentrations of the four dNTPs [65, 66] or to increase the mutational rate using nucleoside analogues [67]. EpPCR can be easily implemented into directed evolution workflows. For instance, random mutagenesis through epPCR was successfully used to generate the starting diversity in xylanase from *Thermomyces lanuginosus,* which through directed evolution, variants with enhanced alkaline and thermal stability were identified [68]. Incorporating epPCR in the directed evolution of metalloenzymes has also shown to be an efficient approach to identify variants with enhanced selectivity and activity [69].

Nonetheless, approaches involving random mutagenesis tend to result in large libraries that still only sample a small proportion of the sequence space possible and their effectiveness come with the cost of amplification biases [70, 71]. Semi-rational design is often a preferred strategy as it allows the generation of smaller but higher quality libraries [70]. Semi-rational approaches require the identification or pre-selection of potential functionally relevant targets that can be sampled simultaneously. A wide range of tools exists to identify mutagenesis candidates and predict the potential functional impact of mutations, including sequence or structure-based and computational approaches [70]. Sequence-based approaches typically involve the identification of functional hot spots and conserved amino acids by comparing the target sequence with homologous protein sequences. When the target protein and homologous proteins have been structurally characterised, these can also be compared to identify potential functional residues. This approach tends to result in more accurate predictions of functionally relevant residues due to the direct identification of residues located in active sites or domain interfaces [70]. Computational techniques can also help identify residues that do not appear to be directly involved in the catalytic activity of a protein but contribute to its efficiency through overall protein stabilization. Examples of this include programs such as SCide, which allow the identification of stabilisation centre elements (SCEs), residues involved in noncovalent cooperative long-range contacts, which have shown to be involved in maintaining the 3D structure and thermal stability of proteins [72, 73]. Another example is the WHAT IF server (swift.cmbi.umcn.nl) that allows the identification of interdomain contacts, which have shown to be involved in stabilisation of catalytically active protein conformations [74]. Once targets have

been identified, they can then be targeted through a variety of mutagenesis approaches. Numerous site-saturation mutagenesis strategies have been developed and typically involve the use of DNA oligonucleotides harbouring codons in place of the target residue encoding all potential 19 amino acids [75]. ‘Small intelligent libraries’ can be generated by using a combination of complementary oligonucleotides containing either the codon NDT, VMA, ATG or TGG and mixed in a specific ratio that ensures an even distribution of degenerate amino acids and avoids rare codons [75].

For single-site saturation mutagenesis, techniques for site-directed mutagenesis can be generally implemented [76]. Some of these techniques include OE-PCR (overlapping extension PCR) [77]. This approach involves the amplification of the template in two subsequent steps. Initially, the template is amplified in two independent reactions, one with a 5’-end forward primer and reverse degenerate mutagenic primers that should bind the target site and another reaction with forward degenerate primers complementary to the reverse degenerate primers and a 3’-end reverse primer. The two resulting amplicons thus contain overlapping regions that can then be combined into the full-length product carrying the desired mutations through a final amplification step. Another widely used approach for site-saturation mutagenesis is whole plasmid amplification, where two mutagenic outward-facing overlapping or partially overlapping primers amplify the whole plasmid, which can then self-ligate into the final product [78].

Numerous techniques have also been developed for multiple-site saturation mutagenesis. Multiple-site saturation mutagenesis is of great interest as it can uncover positive and negative epistatic interactions in a single round of mutagenesis as opposed to iterative rounds of single-site saturation mutagenesis. For multiple-site saturation mutagenesis, techniques such as MOE-PCR (Multiple overlap extension PCR) [79] or POEP (PAGE-mediated overlap extension PCR) [80] can be implemented to target up to 6 sites. MOE-PCR involves the generation of multiple amplicons using forward and reverse mutagenic primers containing complementary regions. The resulting amplicons are isolated and purified through agarose gel electrophoresis. Pairs of adjacent amplicons are individually mixed and amplified for 15 PCR cycles and then combined for 20 more amplification cycles, to which 3’ and 5’ flanking primers

are added for an additional 10 cycles. In POEP, the same principle of MOE-PCR applies, but amplicons are purified through polyacrylamide gel electrophoresis, which has higher resolving ability that facilitates the separation from the original template and minimises wild-type contamination. In POEP, all amplicons are assembled in a single PCR amplification step with a high-fidelity polymerase. Another approach that has also shown to target up to 6 sites simultaneously is ISOR (Incorporating Synthetic Nucleotides via Gene Reassembly) [81]. This technique involves the PCR amplification of the gene of interest with a 5'-biotinylated forward primer. The PCR product is treated with DNAse I and resulting fragments are mixed with mutagenic oligonucleotides. Fragments are then re-assembled through self-primer extension and can then be captured with streptavidin-coated magnetic beads.

Less time consuming and more effective approaches that can target up to 10-distal sites involve the annealing of the mutagenic primers and flanking primers with overhangs to the denatured DNA template [82]. Subsequent strand-specific PCR amplification is carried out and T4-ligation then seals the nicks. The mutagenised strand is then further amplified with primers that bind to the introduced overhangs in the previous step. Improvements to this approach, that bypasses the limitations imposed by T4 DNA ligase-temperature requirements and its ability to ligate across gaps and mispairs has been developed with an approach called Darwin Assembly [71]. This approach also removes the second strand of the original plasmid and thus further minimises wild-type contamination. Darwin Assembly is a robust and efficient library assembly method that allows the generation of large, high quality and complex libraries with over 10^8 transformants targeting more than 10 distal sites. With this approach, the single-stranded plasmid is generated by nicking the template with a nicking endonuclease and subsequent digestion by exonuclease III. Boundary oligonucleotides and inner mutagenic oligonucleotides are annealed to the template by freezing and thawing. The boundary oligonucleotides contain non-complementary overhangs harbouring Type IIS restriction sites for downstream cloning. The 5' boundary oligonucleotide is also biotinylated for recovery post-assembly. After annealing, the primers are extended through isothermal amplification and the assembled product is recovered with

streptavidin coated paramagnetic beads. The assembly is then PCR amplified and product subcloned into the expression vector through Type IIS cloning.

All the above-mentioned techniques perform with different degrees of success, in terms of efficacy and efficiency, depending on the desired outcome. It is important to consider when selecting the mutagenesis technique that it is not feasible to sample the complete mutational space of proteins without exceeding the limitations of current selection and screening techniques [63]. From plate-based to high-throughput screening only 10^4 - 10^{10} variants can be sampled, implicating that only 2 to 6 sites can be fully randomised respectively [81]. Screening a portion of the library is possible but as diversification increases, so does the proportion of inactive variants, which reduces the probability of identifying active ones [81]. Thus, it is important to select an appropriate mutagenesis technique that suits the selection and screening technique.

1.4.2 Artificial Selection Techniques

Directed evolution requires the partitioning of libraries on the basis of a desired phenotype that can be traced back to its genotype. Numerous artificial selection strategies for directed evolution have been and continue to be developed with the aim of expanding the sampling capacity and improving the robustness of selection.

Phage-assisted continuous evolution (PACE) is a platform for *in vivo* directed evolution that incorporates random continuous mutagenesis within the selection workflow [83]. This technique involves linking the desired protein activity to the production of infectious progeny phage. This is achieved by deleting gene III (encoding protein III (pIII)) required for phage infection and introducing it into *E. coli* host cells in a plasmid, which in turn is under the control of the activity of the desired evolving gene cloned intro the selection phage. Active phage vectors inducing sufficient pIII production in the host will propagate and remain in the lagoon as it is continuously diluted in a fixed-volume vessel. This approach was successfully implemented for the evolution of T7 RNA polymerase (RNAP) variants with activities that exceeded the wild-type T7 RNAP on the wild-type promoter as well as variants able to recognise other promoters [83]. PACE has also been used for the directed evolution of a TEV protease with altered substrate specificity by linking proteolysis to gene III

expression using a protease-activated RNAP (PA-RNAP) [84]. PA-RNAP is composed T7 RNAP fused to T7 lysozyme through a protease-cleavable linker that differs to the target cleavage sequence of the evolving protease. Other activities such as protein–protein binding and recombinase activity have been successfully linked to gene III expression [83], which expands the applicability of PACE to numerous protein targets. PACE is an attractive approach, as it does not require (but would benefit from) the generation of DNA libraries, cloning and transforming cells. Still, it is limited to functions that can be linked to gene expression.

In the case of XNA polymerase engineering, with developments in selection platforms such as Compartmentalised Self-Tagging (CST) [85], Compartmentalised Self-Replication (CSR) [86] and droplet-based optical polymerase sorting (DrOPS) [18], polymerase variants with an expanded substrate spectrum and polymerase functions have been identified [3]. These selection platforms establish strong phenotype-genotype link through the encapsulation of individual variants along with their substrates/products in water-in-oil emulsions, which allows the partition of large libraries on the basis of the enzymatic properties (i.e. substrate recognition, product formation, rate and turnover) of individual variants [85, 86].

In CSR, libraries are induced to begin protein expression and are then emulsified in water-in-oil droplets in order to contain cells within individual emulsion compartments. Emulsions are provided with primers with complementary binding regions that flank the gene encoding the polymerase variant. Once cells within emulsions are lysed through a denaturation step, self-replication of active variants and depletion of inactive ones occurs. The post-replication copy number can therefore be proportionally correlated to the enzymatic turnover. The ‘offspring’ polymerase genes can then be isolated, re-diversified and sub-cloned into the expression vector for another round of selection. The reactions can be subjected to different selection pressures such as high temperatures or high heparin concentrations, to identify variants with desired improved resistance [86]. CSR has been successfully implemented to expand the substrate spectrum of Taq DNA polymerase by including primers containing 3’ mismatches during selection, allowing the selection of variants that can also recognise and incorporate phosphorothioates or fluorescent dye–

labeled nucleotide triphosphates [87]. This approach has also been adapted for the evolution of the mesophilic phi29 DNA polymerase, where the denaturing step is swapped for freezing-thawing cycles [88].

Similar to CSR, CST involves the emulsification of libraries into individual compartments to preserve the phenotype-genotype relationship of individual variants. Libraries of 10^8 variants can be easily sampled with this approach. During selection for XNA synthesis, the emulsions are provided with the desired xNTPs (xenobiotic nucleotide triphosphates or triphosphate analogues) and the biotinylated primer that binds multiple sites on the plasmid harbouring the variant polymerase gene; cells within the confinement of the emulsions are then lysed through heat denaturation so that the polymerases can access the biotinylated primer and extend from it with the provided xNTPs. Emulsions are then disrupted, and the plasmids are captured with streptavidin-coated paramagnetic beads and washed in stringent conditions to enrich stably bound plasmid-harbouring variants that extended the primer more efficiently. Selection parameters such primer concentration, number of primer binding sites, xNTP concentration, buffer composition and incubation time can be optimised in order to adjust the stringency and robustness of the selection. The recovered plasmids can then be subjected to PCR reactions to amplify the gene of interest or a portion of the gene, which can then be sub-cloned back into the expression vector. Assembled vectors can then be transformed and the library can be subjected to additional rounds of selection or screened. CST has been used to identify variants of Tgo DNA polymerase able to processively synthesise HNA, CeNA, LNA, TNA, ANA (arabinonucleic acids) and FANA (2'-fluoro-arabinonucleic acid) [3].

DrOPS also involves the encapsulation of individual variants, but with this approach, monodisperse double-emulsions (water-in-oil-in-water droplets) are generated and libraries are partitioned through optical sorting without the need of a DNA parent plasmid as template. In this approach, cells are encapsulated, lysed and subjected to primer extension assays using a fluorescent reporter. This reporter is comprised of a primer-template complex with a downstream fluorophore that produces a signal only when the template complex is fully extended, displacing an oligonucleotide labelled with a DNA-quencher bound to the unextended region of the reporter. Water-in-oil-in-water

droplets are sorted according to their fluorescence and screened. This approach is cost-effective, allowing the screening of over 10^8 droplets in a single day, facilitating the sampling of large sequence spaces. This approach also bypasses the need of a parent template for the plasmid extension reaction and requirement of DNA-templated synthesis. Through this approach a manganese-independent TNA polymerase with enhanced fidelity was identified [18].

Compartmentalised selection approaches minimise cross-reactivity or cross-catalysis enabling the engineering of proteins towards the desired phenotype and facilitating the identification of their corresponding genotype [126]. Compartmentalisation also allows to modify selection parameters making selections more robust [126]. Microfluidic-based approaches, have the added advantage of sampling from a monodisperse droplet population, which can result in more precise library partitioning, with the caveat of being more costly, time consuming and challenging than other approaches such as CST and CSR.

1.5 Thesis aim and overview

Although recent advances in polymerase engineering and directed evolution have allowed the identification of more efficient variants, the current generation of XNA polymerases still possess inferior selectivity and efficiency compared to their natural counterparts and still retain too much of their DNA polymerase function to efficiently select xNTPs over dNTPs *in vivo* to implement an XNA episome [18]. A better understanding of the functional space of polymerases is needed to generate more efficient XNA polymerases.

This project aims to construct mutational libraries targeting diversity in a polymerase coupled to functional DNA- and XNA-based selection platforms to isolate a more efficient XNA polymerase. In particular, the research described here focuses on exploring three different mutational techniques and the optimisation of an XNA selection technique (CST) in the process of identifying a more efficient 1,5-anhydrohexitol nucleic acid (HNA) synthetase. Phi29 DNAP, a replicative polymerase from the *Bacillus subtilis* bacteriophage Phi29, was the chosen scaffold as it is a small and well-characterised [89] polymerase with extreme processivity and high fidelity [90, 91]. Phi29 DNAP also possesses exceptional strand-displacement activity and high stability, which permits

isothermal amplification and large-scale DNA synthesis [90]. Thus, Phi29 DNAP possesses a lot of untapped potential for the development of XNA-based technologies. Being also a mesophilic polymerase, Phi29 DNAP could facilitate the development of an XNA episome in well-established mesophilic *in vivo* systems.

The exonuclease-deficient mutant, Phi29 DNAP (D12A), can already synthesize 1,5-anhydrohexitol nucleic acid (HNA), 2'-deoxy-2'- fluoro-arabinonucleic acid (FANA) and 2'-fluoro-2'-deoxyribonucleic acid (2'-fluoro-DNA) [92, 93], three XNAs that share similar chemistries and appear to not create significant steric clashes in the catalytic site of Phi29 DNAP [93]. While, the D12A variant is able to synthesise XNAs, it still does so inefficiently, and retains a significant level of DNA polymerase activity.

Numerous mutations from other replicative polymerases, such as Tgo, have shown to expand substrate specificity and processivity with unnatural substrates. These mutations, however, do not map to phi29 DNAP due to its unique structure, size and little helical character [117]. The phylogenetic depth of phi29 DNAP is shallow in open databases (see Figure 1.1), which makes the identification of structurally relevant residues challenging.

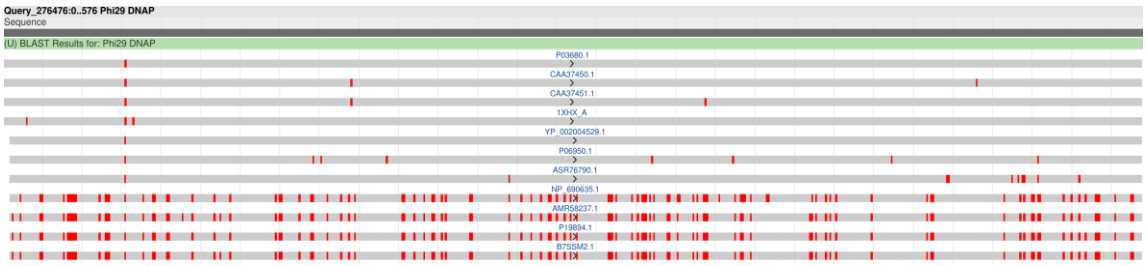


Figure 1.1: Significant alignments of Phi29 DNAP (D12A). Sequences with >80% sequence identity and >80% query cover from a blast search of the D12A protein sequence were selected, aligned using the NCBI Multiple Alignment Tool and viewed using the NCBI MSA viewer. D12A is the first sequence (dark grey) sequence variation is shown in red. Only 11 sequences show significant similarity to phi29 DNAP.

To contribute to our understanding of the sequence-functional landscape of phi29 DNAP while evolving a more efficient HNA polymerase, three engineering approaches were separately implemented targeting different regions of Phi29 DNAP. The first approach, described in further detail in Section 3, involved the generation of InDel libraries of loops belonging to the

exonuclease domain and the terminal protein 2 (TPR2) and thumb subdomains. The second approach involved the multiple-site directed mutagenesis of the finger subdomain through Darwin Assembly, described further in section 4. The last approach involved the random mutagenesis of the thumb subdomain through epPCR, further described in section 5. From the sequence space analysed, mutations of residues that have previously shown to be structurally and functionally relevant were enriched in selection as well as mutations of residues which have not been previously characterised but could be playing a significant role in substrate specificity and processivity. Additionally, a proportion of near wild type sequence variation did not significantly affect fitness and could therefore be used to build the phylogenetic sequence depth of phi29 DNAP.

2. Materials and Methods

2.1 Molecular Biology

2.1.1 PCR

Polymerase Chain Reaction (PCR) was used to generate libraries, amplify fragments or libraries for cloning and recover libraries post-selection for subsequent cloning. Reactions were carried out in Peqstar (Peqlab) and C1000 Touch (BioRad) thermocyclers with the high fidelity Q5 Hot Start DNA Polymerase (New England Biolabs) or KOD Xtreme (EMD Millipore). Reactions with Q5 Hot Start DNA Polymerase typically were carried out in 50 μL reactions with 1X Q5 reaction buffer, 200 μM dNTPs, 0.5 μM forward and reverse primer, 1 ng template DNA or 1 μL re-suspended streptavidin-coated paramagnetic beads and 0.02 U/μL of enzyme. Reaction conditions typically consisted on an initial denaturation at 98°C for 30 seconds; 28 – 35 cycles of 98°C for 10 seconds, 50 - 72°C for 30 seconds and 72°C for 30 seconds/kb of the target DNA product; and a final 72°C extension for 2 minutes. Reactions with KOD Xtreme typically were carried out in 50 μL reactions with 1X KOD reaction buffer, polymerase buffer, 400 μM dNTPs, 0.3 μM forward and reverse primer, 1 ng template DNA or 1 μL re-suspended streptavidin-coated paramagnetic beads and 0.01 U/μL of enzyme. Reaction conditions typically consisted on an initial denaturation at 95°C for 2 minutes; 28 – 35 cycles of 98°C for 15 seconds, 50 - 68°C for 30 seconds and 68°C for 30 seconds/kb of the target DNA product; and a final 68°C extension for 2 minutes. Primers used for each reaction are listed in Appendix B and annealing temperatures were calculated using the Melting Temperature calculator from the NEB website (http://tmcalculator.neb.com/).

2.1.2 Agarose gel electrophoresis

Plasmids, PCR and restriction digested products were analysed in agarose gels prepared in 10 mM Lithium Acetate with 0.8 – 2% (w/v) agarose concentrations depending on the size of the fragment(s) and 0.5X of SYBR Safe stain (Thermo Fisher Scientific). DNA samples were diluted in 6X DNA Gel loading dye (Thermo Fisher Scientific) prior to loading. Gels were run on horizontal electrophoresis systems (AlphaLabs) in 10 mM lithium acetate and visualised

with UV light. For visualising ssDNA, gels were stained post-runing with SYBR Gold (Thermo Fisher Scientific).

2.1.3 Sodium dodecyl sulfate (SDS) polyacrylamide gel electrophoresis

SDS polyacrylamide (SDS-PAGE) gels were carried out to image and quantify protein. Gels were made with two layers, a resolving layer and a stacking layer in vertical systems (Peqlab). The resolving layer was poured first and consisted of 8% (w/v) Acrylamide:Bis-acrylamide, 390 mM Tris-HCl pH 8.8, 0.1% (w/v) SDS, 0.1% (w/v) APS and 0.06% (v/v) TEMED. A layer of 2-butanol was poured immediately after the stacking layer to remove air bubbles. Upon polymerisation of the stacking layer, the 2-butanol was washed with distilled H_2O (dH_2O). The stacking layer poured next was composed of 4% (w/v) Acrylamide:Bis-acrylamide, 140 mM Tris-HCl pH 6.8, 0.07% (w/v) SDS, 0.07% (w/v) APS and 0.07% (v/v) TEMED. Combs were placed to cast the wells for the samples. Samples were diluted in 2X SDS-PAGE sample buffer (100 mM Tris-Cl pH 6.8, 4% (w/v) SDS, 0.02% (w/v) Bromophenol Blue, 20% (v/v) glycerol, 20 mM TCEP) and boiled at 95°C for 5 minutes prior to loading. Gels were run with SDS-PAGE running buffer (25 mM Tris, 192 mM glycine, 0.1% (w/v) SDS pH 8.3) for 2 hours at 200 volts. Gels were then stained with InstantBlue Coomassie Protein Stain (Expedon) for 30 minutes and distained with dH_2O overnight. Gels were imaged using LI-COR Odyssey CLx (LI-COR) and analysed with ImageJ (NIH).

2.1.4 Urea polyacrylamide gel electrophoresis

Denaturing polyacrylamide gels were used to visualize primer extension assays. Gels were composed of 20% polyacrylamide (19:1 acrylamide:bis-acrylamide) with 8 M urea in 1X TBE and casted onto vertical gel systems (Peqlab). Samples were diluted in an equal volume of Urea-PAGE loading solution (98% (v/v) formamide, 10 mM Ethylenediaminetetraacetic acid (EDTA), 0.02% (w/v) Orange G) and boiled at 95°C for 5 min prior to loading. Gels were run with 1X TBE buffer at 25 watts for 2 hours. Gels were imaged using LI-COR Odyssey CLx (LI-COR) and analysed with ImageJ (NIH).

2.1.5 DNA quantification

DNA concentration was quantified using a SpectroStar Nano absorbance plate reader (BMG Labtech). DNA was measured by absorbance at 260 nm and purity by measuring the 260/280 nm for protein contamination.

2.1.6 DNA purification

Libraries were Phenol/Chloroform extracted and ethanol precipitated to reduce the volume of DNA solutions, enhance recovery and remove salts if present. Phenol/Chloroform/Isoamyl Alcohol (25:24:1) was added to the DNA in a 1:1 ratio, vortexed at room temperature and centrifuged at 13,000 *g* for 15 minutes at 4°C. The aqueous phase was recovered and transferred to a new Eppendorf tube and was precipitated by adding 4M ammonium acetate in a 1:10 ratio to the DNA volume, 100% isopropanol in a 3:1 ratio to the DNA volume and 1 μL of 20 mg/mL glycogen azure to facilitate visual detection of the pellet. The mixture was incubated at -80°C for 5 minutes and centrifuged at 13,000 g for 30 minutes at 4°C. The supernatant was discarded, pellets washed with 300 μL 70% ethanol and centrifuged for another 5 minutes. The supernatant was discarded and the pellet re-suspended in 5 – 10 μL of dH_2O. To purify backbones for cloning, the GeneJET PCR Purification Kit (Thermo Fisher Scientific) was used according to manufacturer instructions. DNA extracted and purified from gels was carried out using the Monarch DNA Gel Extraction Kit (New England Biolabs) following the manufacturer instructions.

2.1.7 Oligonucleotide phosphorylation

Oligonucleotides were phosphorylated in 50 µl reactions supplemented with 1x CutSmart buffer (NEB), 1 mM ATP and 0.2 U/µl T4 Polynucleotide kinase (T4 PNK, NEB) for 2h at 37˚C and inactivated for 20 min at 80˚C.

2.1.8 Type IIS cloning

Cloning inserts and backbones were amplified through PCR (See Section 2.1.1) to introduce complementary Type IIS SapI restriction sites. Amplified inserts were then PCR purified with the GeneJET PCR Purification Kit and ~1µg of product was digested with 0.8 U/µl SapI in 1x Cutsmart buffer for 2 hr at 37˚C and inactivated for 20 min at 65˚C. Amplified backbones were treated with 0.4 U/µl DpnI for 1 hr at 37°C and PCR purified with the GeneJET PCR Purification Kit. ~1µg of purified backbone was digested with 8 U/µl SapI in 1x Cutsmart buffer supplemented with 0.25 U/µl Antarctic phosphatase (NEB) and 1x

Antarctic phosphatase buffer. Backbones were then gel purified with the Monarch PCR gel extraction kit. In a 1:3 molar ratio, 1µg of digested backbone was mixed in a 100 µL reaction with the insert and supplemented with 40 U/µl T4 DNA ligase and 1x T4 DNA ligase buffer overnight. The ligation was incubated at room temperature overnight, and then inactivated at 60˚C for 20 min. The ligated product was then ethanol precipitated (See Section 2.1.6).

2.2. Microbiology

2.2.1 *E. coli* strains

NEB 10-B *E. coli* (Genotype: Δ*(ara-leu) 7697 araD139 fhuA* Δ*lacX74 galK16 galE15 e14- ϕ80*d*lacZ*Δ*M15 recA1 relA1 endA1 nupG rpsL* (Str^R)*rph spoT1* Δ*(mrr-hsdRMS-mcrBC,* New England Biolabs) was used for cloning modified backbones and for their amplification. T7 Express lysY/I ^{q} Competent *E. coli* (Genotype: MiniF *lysY lacI*^{q}(Cam^R) / *fhuA2 lacZ::T7 gene1 [lon] ompT gal sulA11 R(mcr-73::miniTn10*--Tet^S*)2 [dcm] R(zgb-210::Tn10*--Tet^S*) endA1 Δ(mcrC-mrr) 114::IS10,* New England Biolabs) was used to clone and express libraries and selections. These strains are referred to as their abbreviated names listed in the Abbreviation table before section 1 throughout the text.

2.2.2 *E. coli* culturing

E. coli cultures were grown in LB medium (1% (w/v) tryptone, 0.5% (w/v) NaCl, 0.5% (w/v) yeast extract). For solid medium, 1.5% (w/v) agar was added. For the expression of libraries and selections, *E. coli* cultures were carried out in 2xTY (1.6% (w/v) tryptone, 1% (w/v) yeast extract, 0.5% (w/v) NaCl). Ampicillin (100 mg/ml stock solution in 50% (v/v) ethanol, 50 µg/ml working concentration) was used as the selection marker. Antibiotic free-media were kept at room temperature, antibiotic stocks at -20°C and antibiotic-containing media at 4°C. Media and glassware were sterilised by autoclaving at 121°C for 20 minutes prior to inoculation. Cell density of liquid cultures was estimated by measuring absorbance at 600 nm (OD600) in a SpectroStar Nano (BMG Labtech) after blanking with fresh media.

2.2.3 Electro-competent cell preparation

An overnight culture of the selected *E. coli* strain was grown from a glycerol stock (stored at -80°C) in 5 mL of LB with no antibiotics at 37°C with shaking. Overnight cultures were then diluted in 200 mL fresh media to an OD600 of 0.1.

When preparing electro-competent cells for library transformations, the overnight culture was diluted 1:100 in 50 mL of LB for each µg of DNA to be transformed. The culture was then allowed to reach an OD600 of 0.4 incubating at 37°C with shaking. The cells were then pelleted through centrifugation at 4000 rpm (3250 *g*) for 30 min at 4°C. The supernatant was discarded and the pellet resuspended in 50 mL filter-sterilised chilled 1 mM HEPES (4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid) pH 7.0. The washes and centrifugation step were repeated 3 times decreasing the volume of buffer in half after each wash. Pellets were finally resuspended in 2 mL 1 mM HEPES containing 10% (v/v) filter-sterilised glycerol and divided into 50 µL or 100 µL aliquots then stored at -80°C. For library transformations, the pellet was resuspended in 400µl 1 mM HEPES and mixed with the µg of DNA to be transformed.

2.2.4 *E. coli* transformation by electroporation

Competent cell aliquots (see Section 2.2.3) were thawed on ice when removed from -80°C storage or used fresh for library transformations. Cells were then mixed with purified DNA (see Section 2.1.6) and transferred to chilled 2mm electroporation cuvettes (BioRad). Cells were transformed using a Gene Pulser II electoporator (BioRad) at 2.5 kV, 200 Ω and 25 µF. 450 µL of LB media was added to each cuvette. For library transformations, the cells were resuspended in 5 mL of fresh LB media. Cells were then incubated at 37°C for 1 hr for recovery. Following incubation, 300 µl – 50 µl cells were plated on solid media with the appropriate antibiotic and incubated overnight at 37°C. For library transformations, cells were pelleted through centrifugation at 4000 rpm (3250 *g*) for 5 min at 4°C, resuspended in 1 mL of LB and plated in 24.5 cm x 24.5 cm LB plates with the appropriate antibiotic. Plates were sealed and incubated overnight at 37°C.

2.2.5 Plasmid purification from *E. coli* and Sanger sequencing

Colonies from transformation plates were isolated and grown in liquid media overnight (See Section 2.2.2). Before cell lysis and plasmid extraction, 500 µl of cell culture was isolated and mixed with filter sterilised glycerol (20% v/v final) for long-term storage and -80°C. The plasmid was then extracted from the remaining cell culture using the GeneJET Plasmid Miniprep Kit (Thermo Fisher Scientific). Plasmids were then quantified as described in Section 2.1.5.

Successful assembly of constructs was corroborated by sending plasmids for sequencing to Eurofins Genomics according to the service provider's instructions.

2.3 Library Construction

Oligonucleotides described in this section are listed in Appendix B. The D12A phi29 DNA polymerase and thermostabilised D12A phi29 DNA polymerase genes sub-cloned into the expression vector pET23 with an ampicillin gene, were kindly provided by Dr. Leticia Torres and are referred to as D12A phi29 DNAP and D12A THR phi29 DNAP throughout this text.

2.3.1 Multiple-site saturation mutagenesis through Darwin Assembly

Darwin Assembly consists of 4 main steps: single-stranded plasmid generation, oligonucleotide annealing, assembly of construct and recovery. For the single-stranded plasmid generation, an Nt.BspQI nicking site was introduced onto the D12A THR phi29 DNAP construct (distant from the polymerase gene) through PCR in a Q5 PCR reaction (See Section 2.1.1) with outward facing primers (iPCR_Nt.BspQI_F and iPCR_Nt.BspQI_R). The forward primer (iPCR_Nt.BspQI_F) was designed to carry the nicking site sequence on its 5' end. Both primers, at 10 µM final primer concentration, were phosphorylated (See Section 2.1.7) prior to the PCR reaction. The PCR product was treated with 0.4 U/µ DpnI for 1 hr at 37°C and PCR purified with GeneJET PCR Purification Kit (See Section 2.1.6). 100 ng of the purified product was blunt-end ligated overnight in 50µl reactions with 8 U/µl T4 ligase and 1x T4 ligase buffer. The ligation was phenol/chloroform extracted and ethanol precipitated (See Section 2.1.6) and transformed into electro-competent 10-B cells through electroporation (See Sections 2.2.3 and 2.2.4). Colonies were screened (See Section 2.2.5) and 1 pmol of plasmid carrying the nicking site was digested with 0.1 U/µl Nt.BpsQI for 1 hr at 50˚C in 1X NEBuffer 3.1 followed by 10 U/µl of Exo III for 2 hr at 37˚C. Successful nicking and single stranded plasmid generation were checked through agarose gel electrophoresis (See Section 2.1.2).

Darwin oligonucleotides are subdivided into inner mutagenic, boundary and outnest oligonucleotides. Three groups of long inner mutagenic oligonucleotides denoted F1, F2 and F3, were designed in order to cover all 16-

target sites involved in interdomain contacts of the finger domain. F1 oligonucleotides targeted the first 5 sites, F2 oligonucleotides targeted the next 5 sites and F3 oligonucleotides targeted the last 6 sites (See Appendix B). The expected library size was of 2.4^4 variants 2(20 residues x 5 sites) x (20 residues x 6 sites)). All inner oligonucleotides were designed to have 11-15bp on each side of the first and last mutation and a C or G on each end to ensure efficient primer binding and ligation during assembly and have annealing temperatures between 65˚C-70˚C (not taking into account mismatches). Four inner mutagenic oligonucleotides containing degeneracies (NDT, VMA, ATG or TGG) were generated per target codon and mixed in a 12:6:1:1 ratio respectively to ensure an even distribution of amino acids and avoid stop or rare codons [75]. Boundary oligonucleotides were designed to anneal on the same strand at the 5’ and 3’ ends of the full phi29 DNAP gene and to contain non-complementary overhangs with SapI recognition sites and outnesting PCR priming sites. Assembly with two different types of boundary oligonucleotides was attempted. The first was the theta oligonucleotide where the priming and termination sequences are linked with a flexible linker, which generates a closed circle post-assembly allowing enzymatic clean up. The second type were a pair of oligonucleotides containing a 5’-biotin-TEG tag for purification of assemblies via biotin-streptavidin pull-down and a 3’inverted-dT repeat for 3’-end protection during assembly. The outnest oligonucleotides were complementary to the overhangs introduced by the boundary oligonucleotide(s) and used for PCR amplification of the assemblies prior to Type IIS cloning. Inner and boundary oligonucleotides were phosphorylated as described in Section 2.1.7 at a final 20 µM and 2 µM concentration respectively. 1 pmol of boundary oligonucleotides and 20 pmol of inner oligonucleotides were annealed to the 0.1 pmol of single stranded template by freezing at -20˚C for 20min and thawing or heating to 95˚C and cooling to 12˚C at 0.1˚C/sec.

One volume of 2x Darwin Assembly mix (0.05 U/µl Q5 High-Fidelity DNA polymerase, 8 U/µl Taq DNA ligase, 2 mM NAD+, 0.4 mM each dNTP, 10% (w/v) PEG 8000, 2 mM DTT and 1x CutSmart buffer) was then added to the assembly mix and the reaction was incubated at 50˚C for 1 hr.

Assemblies performed with the 5’-biotinylated oligonucleotides (P2_DA_F1_SapI and 2_DA_R1_SapI or P2_DA_F2_SapI and

2_DA_R2_SapI, refered to as F1/R1 and F2/R2 in the text) were captured with pre-blocked Dynabeads MyOne Streptavidin C1 beads as described in the Darwin Assembly protocol. Assemblies performed with the theta oligonucleotide (Theta2(F1R2)) were purified through exonuclease digestion as described in the Darwin Assembly protocol using the targeting oligonucleotide pET23P2_NotI_F, designed to bind a NotI restriction site in the assembled product. 1-2 µl of purified assembly were amplified in KOD Xtreme PCR reactions (See Section 2.1.1) with P2_DA_F1_SapI_Out and P2_DA_R1_SapI_Out oligonucleotides or P2_DA_F1_SapI_Out and P2_DA_R2_SapI_Out oligonucleotides. The purified library was then subcloned back into the expression backbone (amplified in a KOD PCR reaction using the DA_iPCR_pET23_SapI_FWD and DA_iPCR_pET23_SapI_RV2) through Type IIS cloning (See Section 2.1.8). The library was then transformed into electrocompetent cells (See Section 2.2.3 and 2.2.4). Colonies were harvested using a cell scraper and resuspended in 5 mL of LB supplemented with 50 µg/ml ampicillin and 25% filter-sterilised glycerol. Resuspended cells were then split into 5 cryovial tubes and stored at -80˚C.

2.3.2 Insertion/deletion (InDel) mutagenesis by iPCR

The exonuclease, TPR2 and thumb loop InDel libraries were generated through inverse PCR (iPCR) with outward facing primers. For the insertions, the forward primers bound immediately next to the 5’ end of the reverse primer and contained one, two or three additional NNS codons on their 3’ end; for the deletions, the forward primer bound one, two, three or four codons upstream from the 5’ end of the reverse primer. A total of 18 different primer pairs, targeting the loops of the three subdomains, were individually phosphorylated at 10µM final (See Section 2.1.7) prior to the iPCR reactions. The iPCR reactions were carried out with Q5 Hot Start High-Fidelity DNA Polymerase as described in Section 2.1.1. The PCR products were treated with 0.4 U/µl DpnI for 1hr at 37°C and PCR purified with GeneJET PCR Purification Kit according to manufacturer instructions. 100 ng of each cleaned up iPCR product were combined into respective exonuclease, TPR2 and thumb subdomain loop InDel libraries and blunt-end ligated overnight in 50 µl T4 ligase reactions with 40 U/µl T4 DNA ligase and 1x T4 DNA ligase buffer. Ligations were inactivated at 60˚C for 20 min, and ethanol precipitated (see Section 2.1.6). The libraries

were then transformed into electrocompetent cells (See Section 2.2.3 and 2.2.4). Colonies were harvested using a cell scraper and resuspended in 5 mL of LB supplemented with 50 μg/ml ampicillin and 25% filter-sterilised glycerol. Resuspended cells were then split into 5 cryovial tubes and stored at -80˚C.

2.3.3 Random mutagenesis through error-prone PCR (epPCR)

The thumb subdomain was amplified through error prone PCR (epPCR) following an adaptation to the protocol by Vanhercke et al. [66], using unbalanced dNTPs (0.04 mM dATP, 0.04 mM dTTP, 0.2 mM dCTP and 0.2 mM dGTP), 1.5 mM $MnCl_2$, 0.05 U/μl Taq DNA polymerase (New England Biolabs), 1x Standard Taq reaction buffer (New England Biolabs), 0.4 ng/μl template (pET23 vector harboring the D12A phi29 DNAP gene or its thermostabilised version) and 0.5 μM forward and reverse primers (P2_HisTAG_SapI_Rv and Lib7_Fw for the D12A phi29 DNAP and P2_HisTAG_SapI_Rv and Lib7_Fw_ThSTP2 for the thermostabilised version). The thermocycling parameters used comprised of an initial denaturation step of 30 sec at 98˚C followed by 30 cycles of 15 sec at 95˚C, 30 sec at 55˚C and 20 sec at 68˚C and a final extension step of 5 min at 68˚C. The library was purified with the GeneJET PCR Purification Kit and subcloned back into the expression backbone (amplified in a Q5 PCR reaction using the iPCR pET23_Lib7_Fw and iPCR_Lib7_rv oligonucleotides for cloning in the library on the D12A background or iPCR pET23_Lib7_Fw and iPCR_Lib7Rv_ThSTP2 for cloning the library on the thermostabilised version) through Type IIS cloning (See Section 2.1.8). The library was then transformed into electrocompetent cells (See Section 2.2.3 and 2.2.4). Colonies were harvested using a cell scraper and resuspended in 5 mL of LB supplemented with 50 μg/ml ampicillin and 25% filter-sterilised glycerol. Resuspended cells were then split into 5 cryovial tubes and stored at -80˚C.

2.4 Selection: Compartmentalised Self-Tagging (CST)

2.4.1 Emulsification of libraries

Libraries were aliquoted from glycerol stocks in 5 mL of LB supplemented with 50 μg/mL Ampicillin. Cultures were grown to OD600 0.8 by incubating at 37°C with shaking and then induced with 1mM IPTG for 3 hr at 30°C with shaking. Up to10^8 cells can be emulsified and selected through CST, thus assuming that

approximately 1 mL of a OD600 = 0.4 has 10^8 cells, 10^8 cells were isolated and pelleted through centrifugation at 4000 rpm (3250 *g*) for 10 minutes. Cells were then re-suspended in 100µl of activity reaction mix, composed of 30 pmol of a short biotinylated oligo (CST_04(7)exoR), 200µM of hNTPs, 1x Phi29 reaction buffer, 1x Bovine Serum Albumin (BSA, Sigma Aldrich), 1 M betaine, 2µL NotI, 1 mg/ml lysozyme and 5µg/ml polymyxin in 100µL molecular grade water. The re-suspended cells were aliquoted in a 2 mL round bottomed Eppendorf tube containing a 5-mm steel bead and then overlaid with 500µL of an oil mix composed of 4.5% SPAN 80, 0.45% TWEEN 80 and 0.05% Trit on X-100 diluted in mineral oil. Stability mixes were prepared on separate 2 mL round-bottomed eppendorfs with a bead, 100µL molecular grade water and 500µL of the oil mix. The eppendorfs were transferred to the Tissuelyser for emulsification at 14 Hz for 30 seconds for the activity mixes and 20Hz for 20 seconds for the stability mixes. 250µL of stability mix were added to each activity mix immediately after and mixed gently.

2.4.2 Library selection

The activity mixes were incubated at -20°C for 1 hr (to promote cell lysis), then at 30°C for 30 min (allow lysozyme/polymyxin to lyse the cells), then -20°C for 1 hr (to denature the plasmid), then 30°C for 10 min to 12 hr+ (for binding and primer extension) and finally at 65°C for 20 min (inactivate p2). The emulsions were then disrupted with 500µL of 90% butanol and vortexed. The beads were removed and the bottom phase of each selection (100µL approx.) recovered by spinning for 5min at 13,000rpm. The selections were incubated with 50 units of trypsin for 2 hr at 37°C (to digest remaining phi29 DNAP bound to templates).

2.4.3 Library recovery

The selections were then captured with 10µL MyOne C1 streptavidin-coated paramagnetic beads (washed three times in BWB1X (20 mM TRIS-HCl pH 7.4, 2 M NaCl, 0.2% v/v Tween-20, 2 mM EDTA) and resuspended in 100µL BWB2X) for 1-3 hours in a rotating mixer at room temperature. The bead suspensions were then transferred to a magnetic stand and beads were washed with 1ml BWB1X, 1ml TBT2 (200 mM NaCl, 100 mM Tris-HCl, pH 7.4, 0.2% (v/v) Tween 20, 1 mg/mL BSA), 100ul TBT2 20% Formamide (to remove loosely bound plasmids) and resuspended in 50µL Tris-HCL 10mM pH7.5. Selections were then PCR amplified with Q5 Hot Start High-Fidelity DNA

Polymerase (See Section 2.1.1) using the corresponding selection oligonucleotides for each selection library and backbone. After amplification, selections were cloned into the expression backbones through Type IIS cloning and transformed into fresh electrocompetent *E. coli* cells (See Sections 2.1.8 and 2.2.4). Colonies were harvested using a cell scraper and resuspended in 5 mL of LB supplemented with 50 µg/ml ampicillin and 25% filter-sterilised glycerol. Resuspended cells were then split into 5 cryovial tubes and stored at -80˚C. From the glycerol stocks cultures were grown to go into selection or screening.

2.5 Activity assays

All activity assays were analysed through denaturing urea PAGE gel electrophoresis (See Section 2.1.4) and visualised using a LI-COR Odyssey CLx (LI-COR) and analysed with ImageJ (NIH).

2.5.1 His-tagged protein purification

To test the activity of the initial libraries and selections, phi29 DNAP was extracted, purified and concentrated prior to activity assays. Libraries were aliquoted from glycerol stocks into 50 mL 2xTY media supplemented with 50 µg/mL ampicillin. Cultures were incubated at 37°C with shaking until reaching an OD600 of 0.6 - 1. Cultures were then induced with 1mM IPTG for 3 hr at 30°C. Cells were then pelleted through centrifugation at 4000 rpm (3250 *g*) for 30 min at 4°C and the supernatant was discarded. Pellets were resuspended in 10 mL of lysis buffer (50mM NaH_2PO_2, 300mM NaCl, 10 mM imidazole, pH 8.0) supplemented with 1 mg/mL of lysozyme from chicken egg white and incubated for 30 min at room temperature in an end-over-end shaker. Cells were pelleted by centrifugation for 30 min and the cleared lysate was isolated. The Super Ni-NTA Agarose Resin was prepared by aliquoting 1 mL into a 15 mL Eppendorf. The supernatant was discarded and the resin resuspended in 2.5mL of the lysis buffer and allowed to settle. 2mL of supernatant were discarded and the cleared lysate was added to the resin. The lysate and resin were incubated at 4°C for 1hr. The lysate was discarded and the resin washed with 5 mL of wash buffer (50mM NaH_2PO_2, 300mM NaCl, 20 mM imidazole, pH 8.0) three times by centrifuging for a couple of seconds and discarding the supernatant after each wash. The resin was then resuspended in 2.5mL of elution buffer (50 mM

NaH_2PO_2, 300mM NaCl, 500mM imidazole, pH 8.0). The supernatant was then recovered and transferred Amicon Ultra-4 Centrifugal 50 kDa Filter Unit and concentrated through centrifugation at 4000 rpm (3250 *g*) for 5 min at 4°C. Buffer exchange was then carried out by adding 4mL of Phi29 DNA storage buffer (10 mM Tris-HCl, 100 mM KCl, 1 mM DTT, 0.1 mM EDTA, pH 7.4 at 25°C). The concentrated protein was quantified through SDS PAGE gel electrophoresis (See Section 2.1.3). 50% Glycerol, 0.5% Tween 20, 0.5% Nonidet P40 were then added for long term storage at -20°C.

2.5.2 HNA and DNA synthesis primer extension assays

Primer extension assays were carried out to test the HNA and DNA synthesis activity of libraries and selections (as well as D12A THR Phi29 DNA polymerase and WT Phi29 DNA polymerase from NEB). Protein was purified (See Section2.5.1) prior to assays to minimise background extensions with dNTPs present in unpurified lysates. The reaction mix comprised of 1µl Phi29 reaction buffer (typically 50 mM Tris-HCl, 10 mM $MgCl_2$, 10 mM $(NH_4)_2S0_4$, 4 mM DTT, pH 7.5 at 25°C), 0.1µl BSA 100X, 2µl Betaine 5M, 3pmol of single stranded DNA template (TempN), 1pmol of a fluorophore-labelled-exonuclease-resistant DNA primer (Tag01F3-exoR), 0.8µl hNTPs 2.5mM, 1- 5.7 µl of purified protein and molecular grade water to 10µl. The reactions were incubated at 30°C for 10 min - 3 hr and inactivated at 65°C for 20 min.

2.5.3 Strand displacement activity assays

The processivity of the libraries and selections (as well as that of WT and D12A THR p2) was measured by including a probe (TempNblock+20-ExoR) in the primer extension assay (See Section 2.5.2) that should bind 20 bases upstream from the primer-binding site. An inverted dT was incorporated to the 3'-end of the oligonucleotide, which inhibits 3' exonuclease degradation and extension by DNA polymerases. Full extensions were expected only from variants able to displace the probe from the template during synthesis. Prior to the extension assays, 1 pmol of primer and 1 pmol of probe were pre-annealed to 3 pmol of template by incubating them at 95°C for 5 min and cooling them down to 4°C at a rate of 0.1°C/sec in a thermo-cycler.

2.5.4 Small-scale expressions and activity assays

Selections were streaked onto 24.5 cm x 24.5 cm square LB plates supplemented with 50µg/ml Ampicillin directly from glycerol stocks and

incubated overnight at 37°C. 96 colonies were picked and inoculated in a 96-well plate (round bottom, 200 µL 2xTY 50µg/ml Ampicillin) and grown for 3hr at 37°C with shaking at 250 rpm. Cultures were then induced with 1mM IPTG for 3 hr at 30°C. Cultures were transferred to a V-shaped 96-well plate and pelleted by centrifugation at 4000 rpm (3250 *g*) for 20 min at 4°C. Supernatant was removed and pellets were dried upside down. 20 µL of filter-sterilised 2x Storage buffer (20 mM Tris-HCl pH7.5, 200 mM KCl, 2 mM DTT, 0.2 mM EDTA) were added. Cells were then lysed by adding 1 mg/mL lysozyme and 5 µg/mL polymyxin and incubating them for 30 min in an end-over-end shaker at room temperature followed by centrifugation (15 min 4000 rpm (3250 *g*) at 4°C). 1.5 µl of lysate was mixed with 9 µl of activity assay mix composed of 1 µl Phi29 reaction buffer (50 mM Tris-HCl, 10 mM MgCl2, 10 mM $(NH_4)_2S0_4$, 4 mM DTT, pH 7.5 at 25°C), 0.1µl BSA 100X, 2µl Betaine 5M, 1 pmol of Tag01F3-exoR pre-annealed to 3 pmol of single stranded DNA template (TempN), 0.8µl hNTPs 2.5mM and molecular grade water to 9 µl. Samples were dispensed with 10 µl mineral oil to prevent evaporation. Primer was pre-annealed to the template by incubating them together at 95°C for 5 min and cooling them down to 4°C at a rate of 0.1°C/sec. The 10 µl reactions were incubated at 30°C for 3 hr and inactivated at 60°C for 20 min.

2.4 Deep sequencing data clean up and analysis

2.4.1 Deep sequencing library preparation

NGS-based amplicon sequencing was carried out for all libraries at Genewiz UK Ltd. Libraries were prepared by generating amplicons of ~400 bp through PCR using KOD Xtreme Hot Start DNA Polymerase (EMD Millipore). Amplicons were purified from agarose gels stained with SYBR safe (Life Technologies) using Monarch DNA Gel Extraction kits (NEB) and ethanol precipitated and resuspended in 25 µL of molecular grade water to a final 20 ng/µL concentration. Amplicons were quantified using a SpectroStar Nano (BMG Labtech).

2.4.2 Deep sequencing data clean up

The sequencing data were cleaned-up and processed using the Galaxy public server (usegalaxy.org). The Fastq-join (Galaxy Version 1.1.2-484) tool was used to join paired-end reads on the overlapping ends allowing 0 mismatches

and a minimum of 10 bp overlap. Fastq reads were converted to Fasta format using FASTQ to FASTA converter (Galaxy Version 1.0.0), discarding reads with unknown bases. Reads were trimmed at the 3' and 5' using the Cutadapt (Galaxy Version 1.16.4) tool inputting adapter sequences immediately upstream or downstream of the mutagenised region of the libraries and with 100% adapter overlap, no minimum or maximum length of trimmed reads and 1 bp mismatch maximum between adapter and read. Reads that did not contain adapter were discarded. Trimmed reads were filtered by length using the Filter sequences by length (Galaxy Version 1.1) tool with minimum and maximum read lengths corresponding to the expected length post-trimming. Reads were converted to protein sequences using transeq (Galaxy Version 5.0.0). Protein sequences were exported as fasta files. In addition, the sequence lengths of reads from the InDel loop libraries/selections were calculated using the Compute sequence length (Galaxy Version 1.0.0) tool and the outputs were exported in tabular format.

2.4.3 Enrichment and Fitness scores

All the insertions and deletions of the loop libraries were counted in RStudio version 3.5.1 for Mac OS X. The frequency of each InDel was calculated by dividing the count number of individual variants over the total number of counts in Excel. Enrichment ratios were calculated by dividing the frequency of each variant after 1 round (R1) by its frequency in the R0. Following the approach of Fowler et al. [118] to compare the fitness of each variant to the wild-type, fitness scores were calculated by dividing the enrichment ratio of each variant over that of the wild-type. To statistically determine if enrichments were significant, proportions from the R1 and R0 were compared with a two-tailed pooled two proportion Z-tests [124] as described in the following equation:

$$Z_t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})(\frac{1}{n_1}+\frac{1}{n_2})}}$$

$$\hat{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1+n_2} \quad (1)$$

where n_1 and n_2 are the sample sizes and $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions using RStudio version 3.5.1 for Mac OS X. For the multiple-site saturation mutagenesis and random mutagenesis libraries, sequences were initially aligned using clustal-omega-1.2.4 locally and visualised using BioEdit

Sequence Alignment Editor [103]. Frequencies of each amino acid at each position pre- and post-selection were calculated and compared in Matlab 2018a for Windows 10 using a script written by Dr. Vitor Pinheiro (See Appendix C for script) that used the unpooled version of the two proportions Z-test [124] as described in the following equation:

$$Z_t = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}} \quad (2)$$

where n_1 and n_2 are the sample sizes and $\hat{p}_1$ and $\hat{p}_2$ are the sample proportions. The Bonferroni correction was incorporated into the analysis to correct for multiple testing.

2.4.4 Entropy and Mutual information

Entropies of the multiple-site saturation mutagenesis library were calculated using BioEdit [103] according to equation 3, where $H(l)$ is the entropy at position l and $f(b,l)$, is the frequency at which residue b is found at position l. Entropies at each position in the MSA were plotted in Microsoft Excel. Mutual information (MI) was calculated using equation 4, where $P(a_i, b_j)$ is the frequency of amino acid a at position i and amino acid b at position j in the same sequence, $P(a_i)$ is the frequency of amino acid a at position i and $P(b_j)$ is the frequency of amino acid b at position j. The obtained MI scores were APC corrected [128], sequences were clustered with Hobohm 1 algorithm [129] and then z-score transformed to remove background, improve accuracy and allow the comparison across different protein families respectively. Networks were constructed by linking nodes (positions) when MI>6.5, a threshold shown high sensitivity and specificity in phylogeny studies [106]. To measure residue conservation, the KL divergence was calculated with equation 5, where $P(i)$ is the frequency of amino acid i in a position in the MSA and $Q(i)$ is the frequency of amino acid i in nature. MI and KL divergence calculations and network plotting were done in the MISTIC2 beta server (mistic2.leloir.org.ar).

$$H(l) = -\sum f(b,l)\ln(f(b,l)) \quad (3)$$

$$MI(i,j) = \sum_{a,b} P(a_i, b_j)\log\left(\frac{P(a_i,b_j)}{P(a_i)P(b_j)}\right) \quad (4)$$

$$KLcons_i = \sum_{i=1}^{N} ln\frac{P(i)}{Q(i)} \quad (5)$$

3. Insertion and deletion (InDel) mutagenesis of phi29 DNAP loops

3.1 Introduction

Phi29 DNAP displays exceptional processivity in DNA synthesis [91] but this is greatly reduced in HNA synthesis. Polymerase processivity refers to the ability of polymerases to synthesise DNA in a single template-binding event before dissociating, which in turn influences their efficiency [94]. Phi29 DNAP, as other members of the family B DNA polymerases, is composed of an N-terminal exonuclease domain and a C-terminal polymerisation domain containing the palm, fingers and thumb subdomains [113]. Belonging to the protein-primed subclass of polymerases, the polymerisation domain of phi29 DNAP also contains the terminal protein regions 1 and 2 (TPR1 and TPR2) [95]. The TPR2 confers strand-displacement capacity to phi29 and contributes to its high processivity [95]. The TPR2 along with the palm and thumb subdomains form an internal clamp that accommodates tightly and stabilises the newly synthesised DNA duplex, which is thought to also enhance processivity in a similar manner as sliding-clamp proteins [95].

The D12A mutation in the exonuclease domain, which abolishes phi29 DNAP proofreading activity [93], enhances HNA synthesis; nonetheless, it still performs with lower processivity and efficiency compared to the wild-type protein synthesising DNA. A potential approach to enhance phi29 DNAP processivity and efficiency could be by stabilising the nascent duplex product during HNA synthesis. Three loops belonging to the exonuclease domain, TPR2 subdomain and thumb subdomain of Phi29 DNAP come in close proximity with the nascent DNA duplex during synthesis (See Figure 3.1a). Protein loops have significant functional, stability and folding roles [119] and are therefore suitable engineering candidates. Loop engineering has been used, for instance, to improve the thermostability of a mesophilic transketolase [120] as well as to expand structure and sequence space of the antigen-binding site in antibodies [121]. Modifying the length and composition of loops belonging TPR2 and thumb, two subdomains associated with phi29 DNAP processivity as well as a neighbouring loop belonging to the exonuclease domain, could help

optimise the space for the nascent duplex and stabilise it during HNA synthesis to improve its efficiency.

The directed evolution of loop-modified variants could result in a polymerase with enhanced loop flexibility that reduces steric clashes with the DNA-HNA product or a polymerase with stronger DNA-HNA binding ability, both of which would improve the processivity of phi29 DNAP.

Therefore, diversity was introduced to the three loops through insertion and deletion (InDel) mutagenesis, where insertions carried NNS codons to introduce further diversity. All libraries were constructed on a D12A thermostabilised phi29 DNAP (D12A THR) background.

3.2 Results and Discussion

3.2.1 Library construction, expression and activity

The length of the exonuclease, TPR2 and thumb loops of phi29 DNAP were modified by introducing up to 3 codon insertions and up to 4 codon deletions (See Figure 3.1b) through iPCR as detailed in Section 2.3.2.

All individual libraries and a combination of all 3 (denoted ‘Mix’) underwent a first round of selection (R1) for HNA synthesis as detailed in Section 2.4, initially with a short primer extension incubation time (30 min) to pool highly efficient HNA processing variants, as well as a long incubation time (12+ hrs) to benefit variants that are potentially less processive but display reduced side reactions (i.e. pyrophosphorolysis) that would also result in a more efficient HNA polymerase. The Mix loop selection was included to directly compare the enrichment/activity across all libraries. As shown in Figure 3.2a, although exonuclease selection appears to have resulted in a larger than expected recovery product, the PCR recovery of all 30 min selections appears to be highly similar to the overnight selections which suggests that 30 min incubations may not be the most stringent incubation time or that the PCR has plateaued and the differences between populations has collapsed. The exonuclease and combined Mix loop library selections of either incubation time do appear to have recovered less material compared to the rest. Primer extension assays of 3 hrs were carried out on purified protein from all R1 selections as well as on the original libraries (R0) as detailed in Section 2.5.2. As shown in Figure 3.2b, 3 hr primer extension assays resulted in the

incorporation of 57 nucleotides, the maximum amount of nucleotides able to be incorporated in the given DNA template.

The fact that all libraries and selections reached full extension does not provide sufficient information to determine if there was enrichment of active variants and depletion of inactive variants during selection. Reducing the primer extension time before the reactions plateau would allow a more accurate comparison. Nonetheless, the proportion of variants halting at the first incorporations is greater in the R0 populations than the in the R1 populations, which indicate that more processive and efficient variants have been enriched after one round of selection.

In order to observe a clearer comparison between libraries and selections, assays with a shorter extension time (30 min) were carried out (see Figure 3.2c). All libraries, once again, appear to reach full extension even with a 30 min extension time, however the stronger signal of full extension of the selections compared to their respective R0, indicate enrichment. Still, it is possible that the signals are an artefact of protein concentration, which was not standardised across reactions.

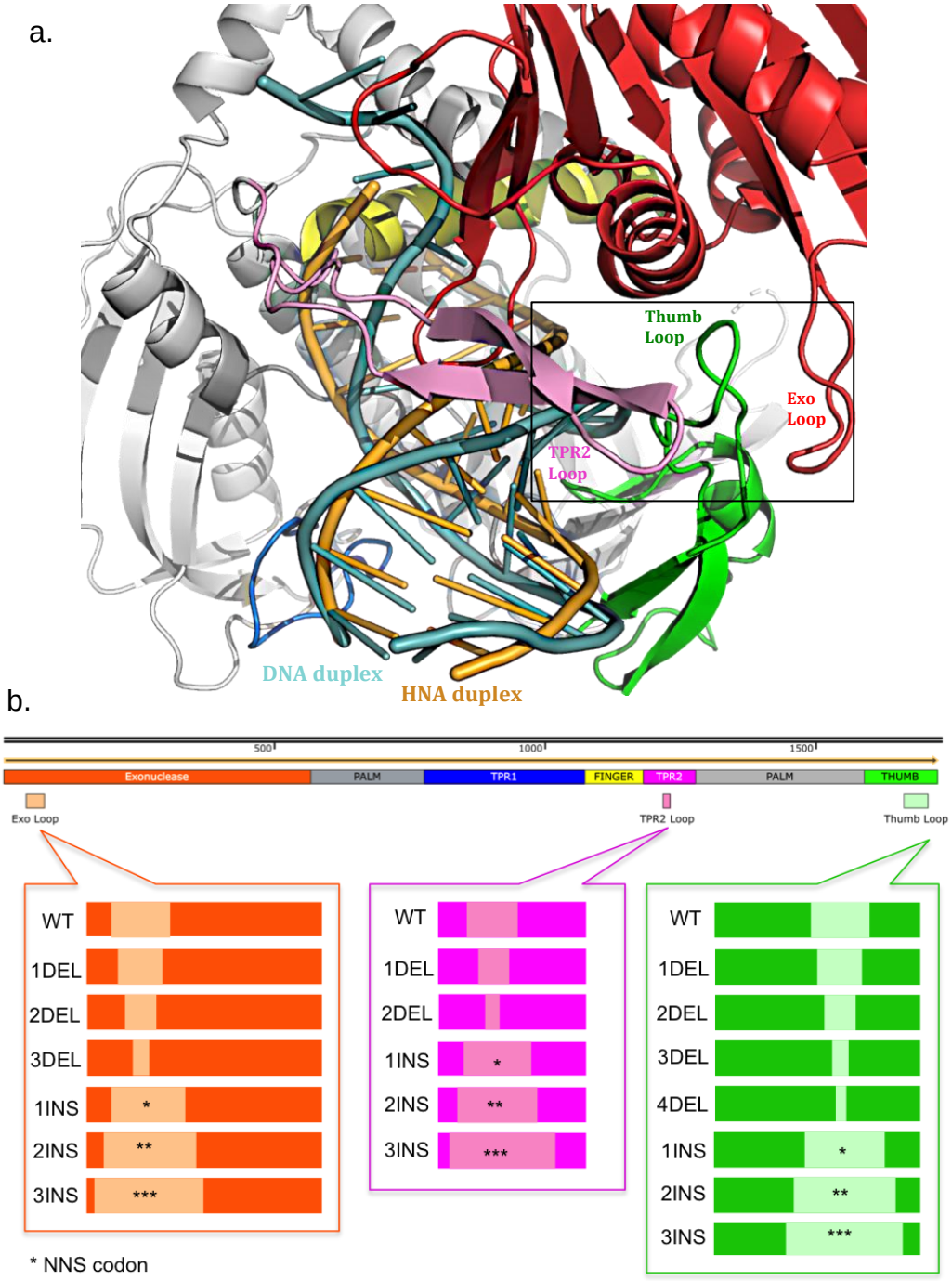


Figure 3.1: Phi29 DNAP InDel library construction. (a) Phi29 DNAP (PDB ID 2PYJ) complexed with DNA and a superimposed HNA duplex (PDB ID 481D); the exonuclease domain loop is shown in red, the TPR2 subdomain loop in pink and the thumb subdomain loop in green. (b) InDel library design on the three phi29 DNAP loops. DEL represents codon deletion and INS codon insertion. The '*' represents each NNS codon introduced.

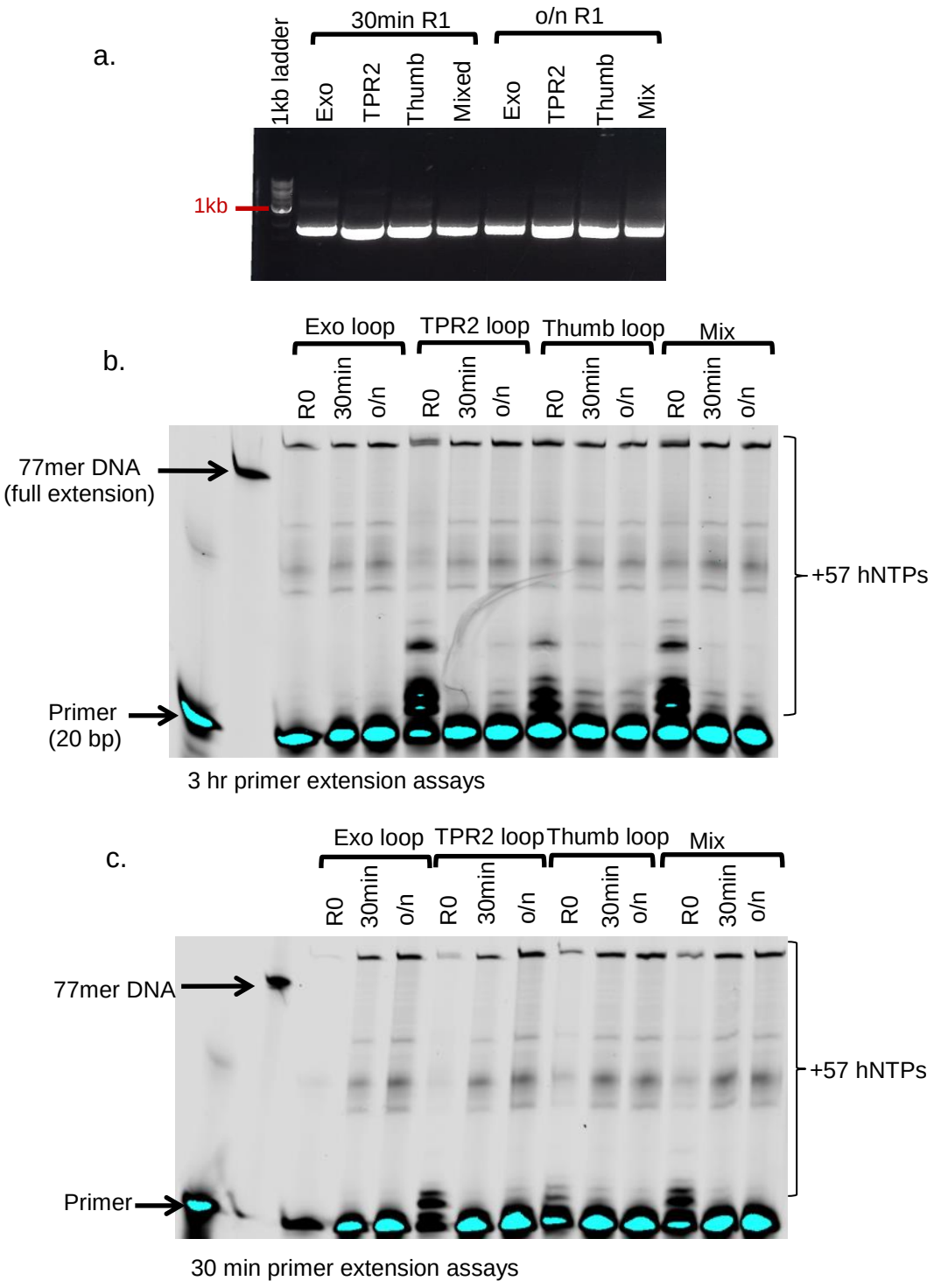


Figure 3.2: InDel library selection for HNA activity. (a) PCR recovery of exo (expected 1kb band size), TPR2 (0.7kp) and thumb (0.7kb) loop libraries after one round of selection for HNA synthesis. (b) 3 hr HNA primer extension assays of selections. (c) 30 min. HNA primer extension assays of selections. HNA migrates slower than DNA in a denaturing PAGE gel [92], explaining why it does not match the

77mer DNA full extension marker. Fluorophore signal saturation is observed at the bottom of the gels due to excess un-extended primer.

3.2.2 Optimising primer extension assay conditions

In order to determine if enrichment is occurring and compare the activity among selections, limiting conditions of primer extension assays where libraries do not reach full extension would have to be determined. By then increasing protein concentration and/or primer extension time or modifying the reaction buffer composition, one can determine which selections are more efficient (i.e. requiring less protein concentration or extension time to reach full extension). To identify these limiting conditions, different conditions were tested through HNA and DNA primer extension assays using D12A THR phi29 DNAP (D12A THR) as well as wild-type phi29 DNAP obtained from New England Biolabs. It is expected that D12A THR will outperform any of the libraries at HNA/DNA synthesis due to the presence of inactive variants in libraries after one round of selection and will outperform the WT in HNA synthesis due to the decreased substrate specificity of the D12A THR variant. Nonetheless, the limiting conditions identified using the D12A THR can then be used in activity assays of libraries and selections for more accurate comparisons.

As shown in Figure 3.3a, for D12A THR, reducing the primer extension time for HNA synthesis to 10 minutes with Phi29 DNAP reaction buffer (New England Biolabs) (Fig 3.3a, lane 3) was sufficient to observe partial HNAs extension. Halving the magnesium concentration to 5 mM of the phi29 DNA reaction buffer had seemingly no impact on the extension efficiency (Fig 3.3a, lane 4), but reducing the concentration to 1 mM (1/10th of the standard Mg2+ concentration) did reduce the efficiency of HNA synthesis slightly (Figure 3.3a, lane 5). Interestingly, for DNA synthesis, full extension was reached at all magnesium concentrations with no significant yield difference. Manganese ions have shown to reduce the substrate specificity of polymerases and facilitate the processing of unnatural substrates [54], thus primer extension assays with $MnCl_2$ instead of or in combination with $MgCl_2$ were also tested.

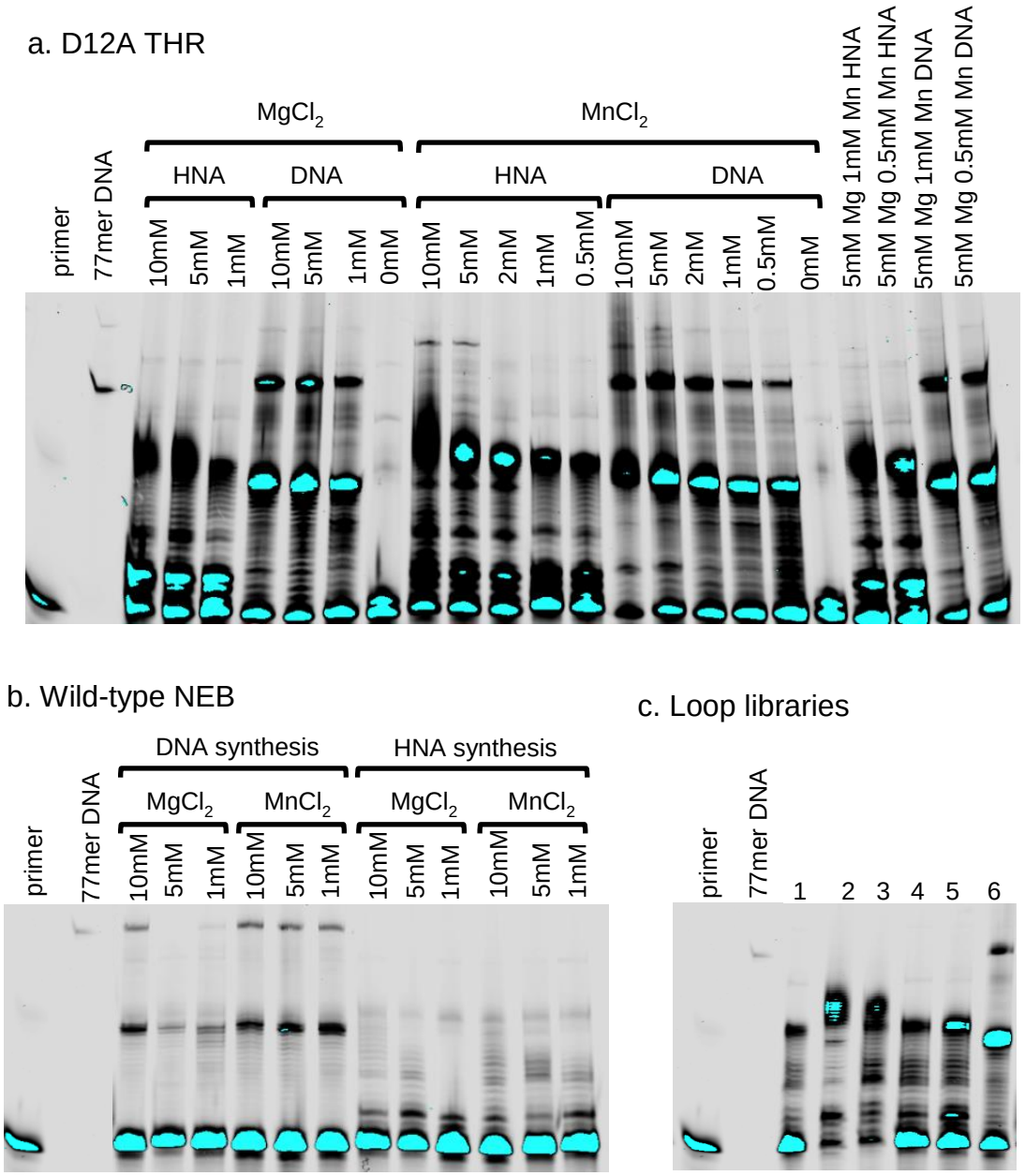


Figure 3.3: Optimising primer extension assays. Testing different magnesium and manganese concentrations during 10 min HNA and DNA primer extension assays with (a) D12A THR phi29 DNAP and (b) phi29 DNA polymerase (NEB). (3c) 10 min HNA synthesis with 10mM $MgCl_2$ of the exonuclease, TPR2, thumb and Mix loop libraries after a 30 min selection for HNA synthesis (lanes 1 – 4) and Mix loop library after a 30 min selection for DNA, synthesising HNA (lane 5) and DNA (lane 6) for 10 min with 10mM $MgCl_2$. Smears or bands above the full extension mark indicate template-independent synthesis.

As shown in Figure 3.3a, for HNA synthesis, D12A THR displays more efficient activity with Mn2+ ions than with Mg2+ ions during 10 min reactions, only reaching full extension with 5 and 10 mM $MnCl_2$. On the other hand, Mn2+ ions appear to reduce the efficiency of DNA synthesis and lead to a higher degree of template-independent synthesis (shown as bands above the full extension mark in Figure 3.3a). Combining $MgCl_2$ with $MnCl_2$ does not appear to have a significant additive effect in HNA or DNA synthesis. DNA synthesis with the wild-type enzyme (NEB, figure 3.3b) under the same conditions resulted in full extensions with significantly less background, template-independent synthesis and partial extensions with either magnesium or manganese compared to the D12A THR variant. Still, as seen previously, DNA and HNA synthesis with manganese appears to be more efficient than with magnesium, but as expected, the wild-type synthesising HNA performs with significantly less efficiency than the D12A THR variant.

10 minute HNA/DNA synthesis with standard reaction buffer composition was thus chosen as a starting point to screen the selections. Since there was no apparent difference between the 30 min and overnight selections, the more stringent selections of 30 min were chosen for further characterisation. Additionally, two mixtures of all three libraries subjected to 30 min HNA and 10 min DNA synthesis selections were also carried out and included in the primer extension assays. As shown in Figure 3.3c, the TPR2 (lane 2) and Thumb (lane 3) InDel selections demonstrate higher HNA synthesis compared to the Exonuclease (lane 1) selection. The mixed library selected for DNA synthesis appears to be slightly more efficient at HNA synthesis (lane 5) than the mixed library selected for HNA synthesis (lane 4), which could be due to a greater enrichment of unmodified D12A THR variants and more efficient depletion of unstable variants during the less stringent selection for DNA synthesis. The HNA synthesis of mixed library selection for HNA synthesis appears to be less efficient than that of the TPR2 or Thumb selections, resembling more the exonuclease selection. The mixed library selected for DNA does reach full extension during DNA synthesis as the D12A THR variant but with less background, which could be due to differences in protein concentration (excess

protein results in more background observed in the D12A THR assays). Overall, the TPR2 and Thumb loop selections appear to synthesise HNA more efficiently than the exonuclease loop selection, as these extended the primer further, nearly reaching full-extension, within the same timeframe (10 min).

3.2.3 Screening isolated variants

To determine if altering the loop length of the thumb and/or TPR2 subdomains does enhance HNA synthesis and to potentially identify an efficient HNA synthetase, variants from each library were individually screened. 3 colonies from the thumb HNA selection and 2 colonies from the TPR2 HNA selection were isolated at random and subjected to primer extension assays of 10 minutes with increasing protein volumes (not standardised) as shown in Figure 3.4.

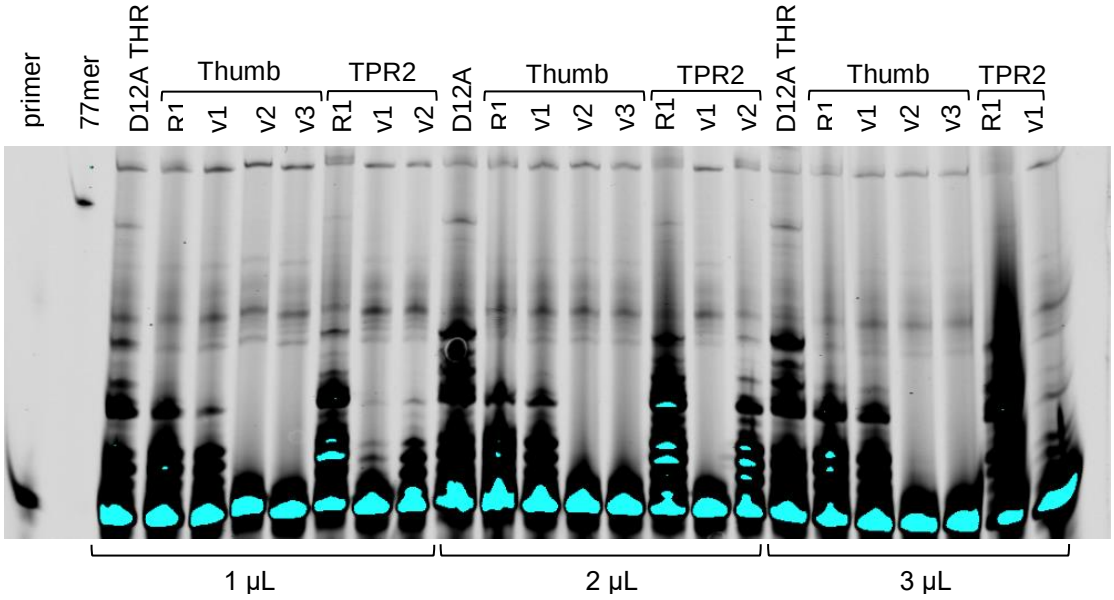
2 out of the 3 variants (Thumb V2 and V3) isolated from the thumb HNA selection contain a frameshift, which, in consequence, should not demonstrate HNA synthesis activity at any concentration. With this in mind, one can deduce that the full-extension bands observed in Figure 3.4, are in fact background signals, especially seeing that saturation of these bands does not increase with the increasing of protein concentration. Primer degradation can originate when reactions are challenging and exonuclease activity is present [92], resulting in background signals as those observed here. This observation is further supported by the fact that the thumb and TPR2 R1 selections reached full extension but did not reach full extension under equivalent conditions previously. Thus, the bands observed from the middle to the top of the gel are not taken into account for the following observations and conclusions.

Thumb V1, containing the P562 and G563 deletions appears to have HNA synthesis activity but is not significantly different to the bulk activity of the R1 population. Due to time constraints, it was not feasible to standardise enzyme concentration prior to the polymerase activity experiments. As a result, it cannot be ruled out that the apparently higher activity of the library is not the result of higher protein concentration. However, even comparing Thumb V1 with 3x more protein to R1 with 1x protein does not reveal a significant enhancement in activity. For the TPR2 selection, all variants isolated contained an N409

a.

Variant No.	Mutation
Thumb V1	P562-G563del
Thumb V2	V561fs (1677_1680delTCCG)
Thumb V3	Q560_V561insLM V561fs (1676_1677delTT)
TPR2 V1	N409del
TPR2 V2	N409del

b.



deletion, which significantly underperforms in HNA synthesis when compared to the R1 population at all concentrations. All variants from either selection and at

Figure 3.4: InDel library variant screening. (a) list of InDel variants isolated at random from the thumb and TPR2 30 min R1 selections (b) 10 min HNA primer extension assays of D12A THR and isolated variants described in (a) with increasing protein volume (not standardised).

any concentration also appear to underperform when compared to the D12A THR. The bulk activity of the TPR2 R1 population, however, does appear to behave similarly to that of the D12A THR, suggesting that the selection contains potentially more efficient variants than the N409del, which can potentially be more efficient than the D12A THR.

Standardization of protein concentration would enable improved comparisons between different mutants or mixes thereof. This was not

implemented in the time frame of the project due to time constraints. The background highlighted in Figure 3.4 is probably the result of incomplete denaturation combined with template independent synthesis. It was hypothesised that removing single-stranded nucleic acids, such as unused templates, could reduce the background. Exonuclease VII was able to reduce the background in gels considerably (not shown) and was added to all subsequent experiments as a post-extension clean up step.

3.2.4 Optimising Selection Stringency

To deplete unstable mutations and further enrich the population with variants of greater HNA synthesis efficiency, subsequent rounds of selection can be performed. After a second round of selection, ~70% of sequences analysed contained frameshift or were truncated and the overall library activity did not improve, suggesting that the proportion of undesired phenotypes due to non-specific recovery or presence of parasites significantly increased. Parasites emerge due to methodological constraints that partition with the desired phenotype, such as variants able to use the low cellular concentrations of dNTPs present in the emulsion [85]. High levels of background and parasites tends to undermine enrichment and parasites exponentially increase with subsequent rounds [85], thus instead of going into a second round of selection, more stringent first rounds of selection were carried out in order to accelerate the directed evolution of a more efficient HNA synthetase.

Among the conditions that can be modified to increase the stringency of selection, the following were considered: extension time, nucleotide concentration, selection primer concentration, and magnesium concentration. Lowering the concentration of hNTPs could push the selection of variants able to recognise and incorporate hNTPs more readily, thus the hNTP concentration was lowered 5-fold, this condition is hereinafter referred to as “condition 1” or c1 (See Figure 3.4a). A potential issue with c1, however, is that a very low hNTP concentration could also push the selection of polymerase variants able to use the low cellular concentration of dNTPs within the emulsion, thus giving rise to parasites. The biotinylated selection primer is what links phenotype to genotype in CST, thus its optimisation can significantly impact selection. Decreasing the selection primer concentration would result in a reduced proportion of primed

templates and, in consequence, less polymerase binding and extending events. This could in turn reduce the probability of recovering poorly extended primers and favour the recovery of highly processive variants able to bind the low abundance of primed templates and begin synthesis efficiently. Thus, reducing primer concentration 4-fold (c2) was tested. As shown in previous primer extension optimisation assays, decreasing the magnesium concentration to 1 mM had a significant impact in synthesis activity, thus this concentration was also tested (c3). From previous activity assays, it also became clear that D12A THR can synthesise HNA within 10 min, the HNA synthesis incubation time during selection was reduced to 10 min in all of the above-mentioned conditions. Finally, c4 comprises a combination of c1, c2 and c3, and should therefore be the most stringent selection.

As shown in Figure 3.5b, the post-selection recovery of the stringent selections appears to be lower than the recovery of 30 min or overnight selections with standard hNTP, primer and magnesium concentration conditions and the same number of PCR cycles seen in Figure 3.2a. This thus suggests lower initial template concentrations recovered from the stringent selections, which would be compatible with a higher stringency selection. Protein was quantified using Phi29 DNAP from NEB as a standard and protein concentrations were standardized for the primer extension assays. Due to phi29 DNAP activity's dependency on concentration and half-life, the wild-type (D12A/D12A THR) was always re-extracted and purified along selections and stored for the same amount of time prior to activity assays. This ensures that the activity from the WT is relatively proportional to the observed activity of selections. Primer extension assays of 30 min were carried out (Figure 3.5c) and it appears that the exonuclease loop library selections, once again, performed with less efficiency compared to the rest.

Exonuclease R1c4, the most stringent selection, appears to have slightly more activity than any of the other selection conditions based on the signal strength of the extended products. Exonuclease R1c1 showed no activity, which due to the reaction volume constraint could not be properly standardised possibly resulting in insufficient protein in the reaction to observe a signal. The TPR2 selections, aside from R1c1, show similar activity levels, which also resemble the bulk activity of the R0 population. For the thumb and TPR2

selections, R1c1 once again had the lowest activity and the rest resemble the activity of the R0, thus it appears to be poor enrichment. Still, early rounds of selection tend to not significantly alter the bulk activity of a population [85], even with strong enrichment the fraction of rare clones will remain a small proportion of the total population and the increased activity may remain below detection limits. All selections also display lower HNA synthesis activity when compared

a.

Selection Conditions	Standard	R1c1	R1c2	R1c3	R1c4
Time	3 hr	**10 min**	**10 min**	**10 min**	**10 min**
hNTPs	200 μM	**50 mM**	200 mM	200 mM	**50 mM**
CST_4(7)	10 μM	10 μM	**2.5 μM**	10 μM	**2.5 μM**
Mg2+	10 mM	10 mM	10 mM	**1 mM**	**1 mM**

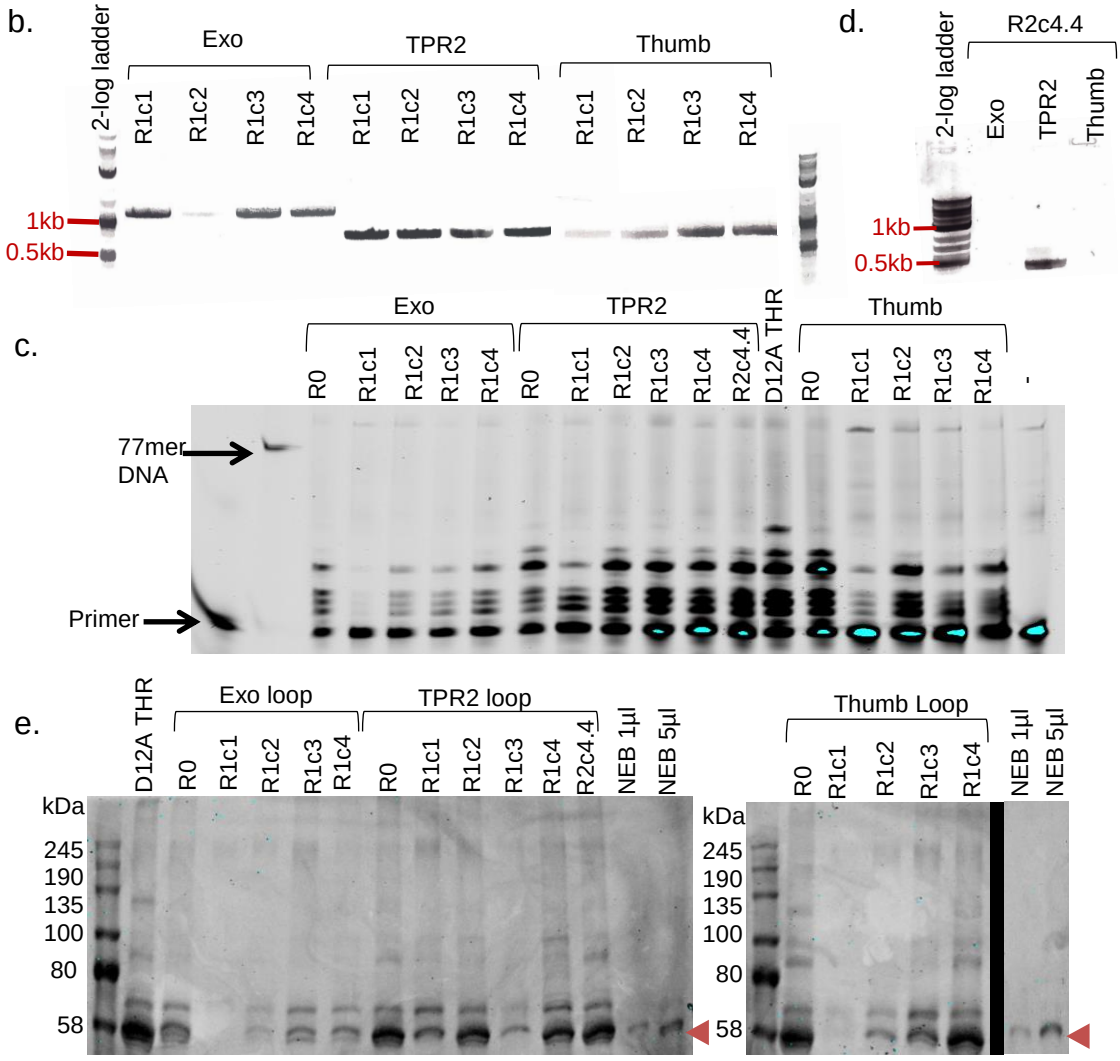


Figure 3.5: Optimising stringency of InDel library selections. (a) Table of standard (black) and stringent selection conditions (red). (b) PCR recovery of exo (expected 1kb band size), TPR2 (0.7kp) and thumb (0.7kb) loop libraries subjected to stringent

selection conditions described in (a). (c) 30 min HNA primer extension assays of exo, TPR2 and thumb loop libraries post-stringent selections with standardised protein concentrations. (d) PCR recovery of exo, TPR2 and thumb R1c4 (most stringent) selection subjected to a second round of selection with the same conditions as R1c4 but with a lower capture time (1 hr instead of 2 hr) and two 20% formamide in TBT washes instead of one. (e) Protein gels of purified libraries used in (c) and phi29 DNAP (NEB) used for protein standardization.

to D12A THR, which is expected due to the presence of inactive/inefficient variants in the population contributing to the overall protein concentration.

Overall, it appears that the combined effect of decreasing NTPs, selection primer and magnesium concentrations, is a more effective selection pressure than low hNTP concentration alone but similar or slightly more effective than lowering the primer or magnesium concentration alone. Limiting primer and NTP concentrations should slow down the reaction rate. Low NTP concentration in particular, should facilitate nucleotide sampling thereby improving fidelity and reducing pyrophosphorolysis. Similarly, lower magnesium concentration has shown to increase substrate residence time and increase fidelity. This is as magnesium ions play a role in coordinating the α- and γ-phosphates of the incoming nucleotide to facilitate the phosphodiester bond formation and stabilising the ternary conformation of phi29 DNAP [122]. The observation that the lower NTP concentration alone pooled a less efficient fraction (in terms of product length during primer extensions) suggests that the tested concentration does not constrain polymerase function as lower magnesium or primer concentrations do, allowing enrichment of an overall less efficient population. It is also possible that low NTP concentration and excess magnesium in fact favours enrichment of efficient variants but with lower fidelity, which, due to the still present exonuclease activity of phi29 DNAP, would result in an enhanced pyrophospholysis rate. This would result in a selection with seemingly less efficient activity. Low magnesium and primer concentrations alone, resulted in more efficient selections (in terms of product length during primer extensions), indicating that these conditions potentially enriched variants with reduced synthesis efficiency but enhanced fidelity and thus reduced pyrophosphorolysis or variants with high efficiency and fidelity.

A second round of selection (R2c4.4) were carried out on the exonuclease, TPR2 and thumb R1c4 selections, with the same selection conditions of the c4 selection but with a shorter plasmid capture time (1 hr instead of 2 hrs) and an additional TBT2 20% formamide bead wash step. The low ionic strength and denaturant formamide washes favours the removal of loosely bound plasmids. Since the probability of plasmid retention is dependent on the primer extension (as proxy for Tm of the complex) this will reduce the proportion of inactive polymerases and background [85]. The recovery of this highly stringent R2 selection was poor for the exonuclease and thumb selections (See Figure 3.5d). Only the TPR2 R2c4 allowed recovery of template and protein expression. The activity of TPR2 R2c4.4, however, was not significantly different from the R1c4 in the conditions tested (See Figure 3.5c).

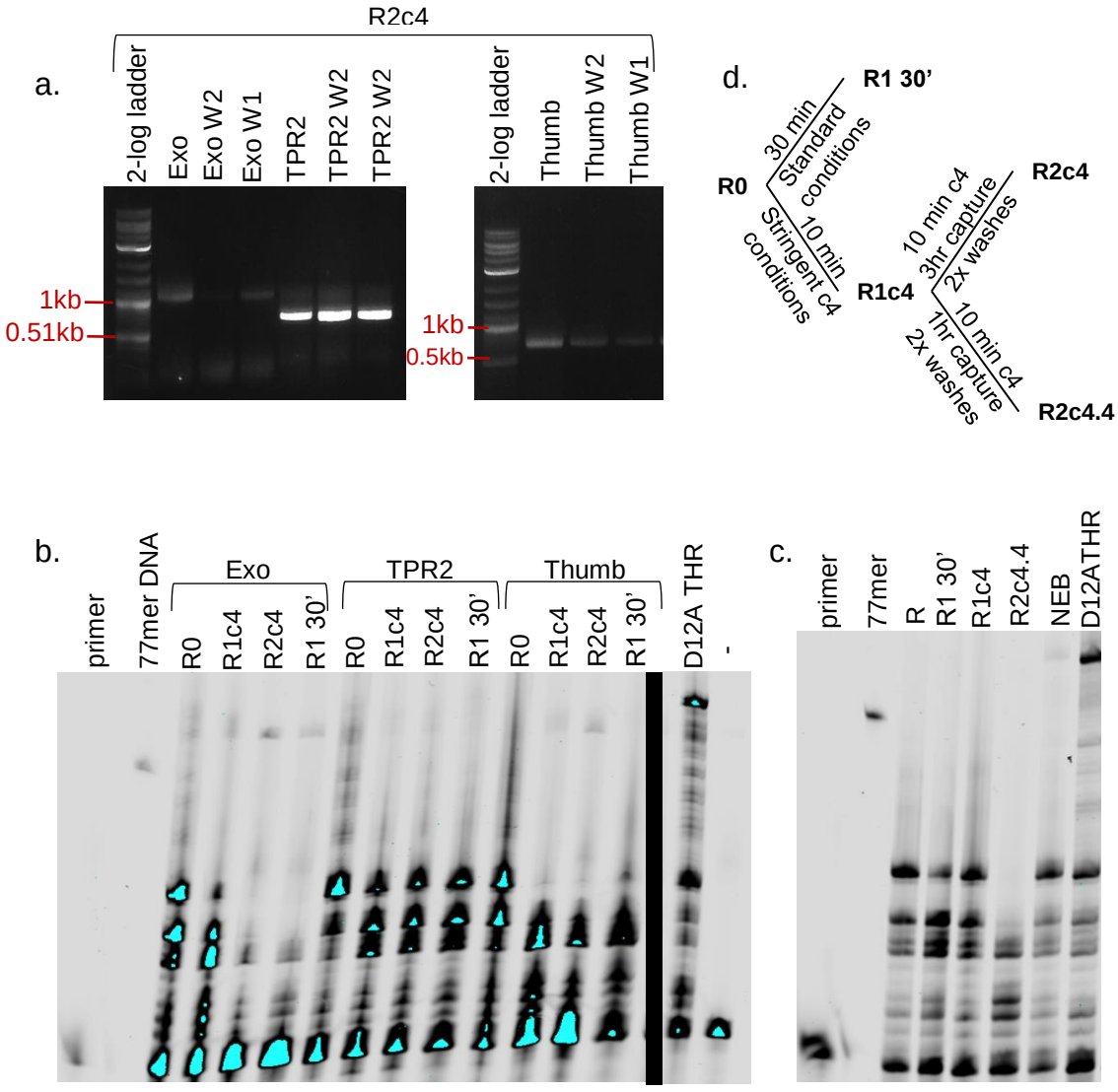


Figure 3.6: Second round of selection of InDel libraries. (a) PCR recovery of exo (expected 1kb band size), TPR2 (0.7kp) and thumb (0.7kb) R1c4 selection subjected to

a second round (R2c4) including recovery after each formamide wash. (b) 3 hr HNA primer extension assays of exo, TPR2 and thumb loop libraries before selection (R0), after a 30 min selection with standard conditions (R1 30'), after a stringent selection (See Figure 3.5a for conditions, R1c4) and after a second round of selection on the R1c1 selections (R2c4) with the same stringent conditions as R1c4 but with a 3 hr capture and 2 formamide washes. (c) Overnight HNA primer extensions of TPR2 loop libraries R0, R1 30', R1c4, R2c4 and R2c4.4 (same as R2c4 but with a 1 hr capture) with standardised protein concentrations. (d) Summary of selections and conditions used for primer extensions in (b and c).

A second round (R2c4) of selection were repeated on the R1c4 selections with the same selection conditions of the c4 selection but with a 3hr capture (instead of 2hr) and 2x TBT2 20% formamide washes (instead of 1hr). After each wash, the eluates were kept to identify the fraction of product lost after each wash. As shown in Figure 3.6a, all selections allowed sufficient recovery even after 2 washes and depict the loss of potentially loosely bound plasmids after each wash. Primer extension assays of 3 hours were carried out on the original R0 libraries, R1c4 selections, R2c4 (on the R1c4 backgrounds) selections, and the 30 min R1 selections with standard selection conditions from Figure 3.2. Protein was not standardised, but overall the same pattern previously observed emerges. The Exo R2c4 and R1 30 min, display equally low activity compared to the R1c4 selection. As previously observed, the Exo R1c4 also displays similar activity to the R0. All thumb selections show similar activity, which is lower than the bulk activity of the R0. The TPR2 R1 and R2 selections and starting R0 library all display a similar HNA synthesis activity, which is superior to any of the Exo or Thumb selections. Here, the Exo R1v4 appears to have similar activity to the TPR2 selections in contrary of what was observed before, nonetheless this is possibly due to an artefact of protein concentration as the protein was not standardized. The TPR2 primer extension assays were repeated but with an overnight incubation (See Figure 3.6c) and revealed that in fact, the R2c4 is less efficient than the R0 and R1 selections.

3.2.5 Small-Scale Screening of the TPR2 R1c4 Selection

The TPR2 selections, displaying more efficient HNA synthesis than the other loop libraries in terms of primer extension length, were selected for further screening. All selections of the TPR2 library showed similar activity during

primer extension assays, TPR2 R1c4 was initially chosen for isolating colonies and test their activity. Small-scale protein expression and activity assays for 3 hr were carried out on 96 colonies from the library (See Appendix D). Variants (3E, 3F, 6F, 9B, 9C, 10D, 10E, 10G) with ranging levels of activity were sequenced and lysates were assayed again in 3 hr primer extension assays (See Figure 3.7).

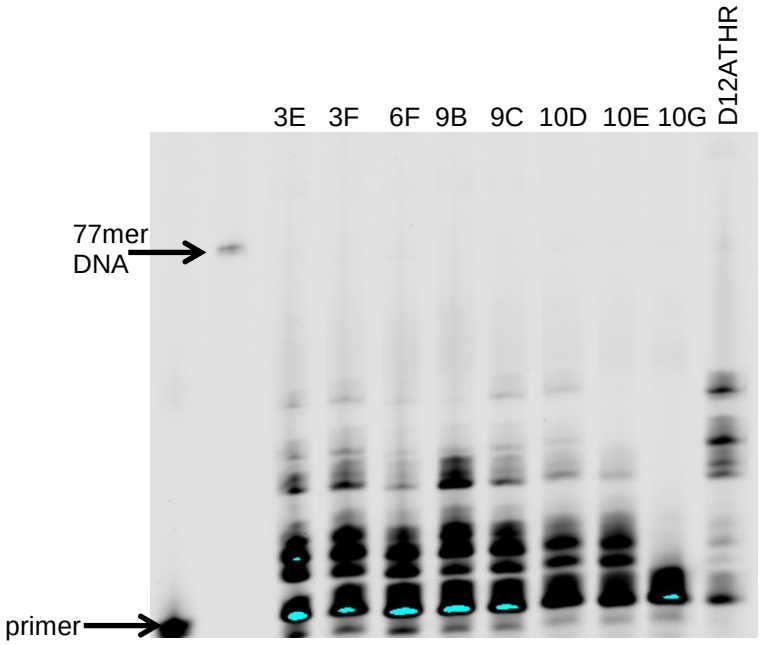


Figure 3.7: Small-scale screening of the TPR2 InDel R1c4 selection. 3 hr HNA primer extension assays with lysates of 8 out of 96 screened colonies from the TPR2 InDel R1c4 selection displaying ranging activity levels. Colonies 3E, 3F, 6F, 9C, 10D contain an N409del, colony 9B an E408_V409insEYW and 10E an E408_V409insGTA. 10G did not yield optimal sequencing results.

From the 8 colonies selected, 3E, 3F, 6F, 9C and 10D contained the same N409del previously identified, 9B contained an E408_V409insEYW and 10E an E408_V409insGTA. 10G did not produce optimal sequencing results. The activity of all N409del variants looked similar as expected, with the exception of 6F that showed slightly less activity. 9B shows slightly more activity than 10E, but both seem to fall behind the activity of the N409del variants. The 10G variant showed no activity, which corroborates the poor sequencing results. Although these assays used unpurified protein directly from the lysates

and protein was not standardised, the consistency between N409del suggests that the protein expression levels were not widely variable in the conditions tested, thus making the lysate platform a reasonable screening approach.

Variants 3F, 9B and 10E were selected for further screening and were thus expressed in large scale and purified. Variation in protein concentration was observed (not shown), thus, without standardising protein concentration, a preliminary assay with increasing amounts of protein supplemented to the DNA and HNA synthesis reactions were carried out. As shown in Figure 3.8a, for both DNA and HNA synthesis variant 9B appears to have significantly more activity than any other variant including D12A THR. 10E appears to have the lowest activity at any concentration. This result was interesting, as, even though there is an apparent large abundance of N409del suggesting its potential enrichment upon selection, the E408_V409insEYW mutation of 9B appears to have more activity. However, the gel was heavily saturated and reactions were incubated for too long. HNA synthesis assays were repeated after protein standardization. As shown in Figure 3.8b, 9B appears to have slightly more activity than 3F but not significantly different from D12A THR. The primer extension assays were also carried out under selection conditions (c4), which revealed that D12A THR is more efficient at HNA synthesis under these stringent conditions than the variants. This suggests that one round of selection, regardless of the selection conditions, may not be enough to pool highly active rare variants and are thus less likely to be isolated during screening. It is also possible that the diversity introduced may not improve HNA synthesis and the wild-type is the most active polymerase. Deep mutational scanning can bypass this screening limitation and facilitate the identification of enriched rare variants after one round of selection. It can also help determine if the wild-type loop length is being enriched more than all other variants.

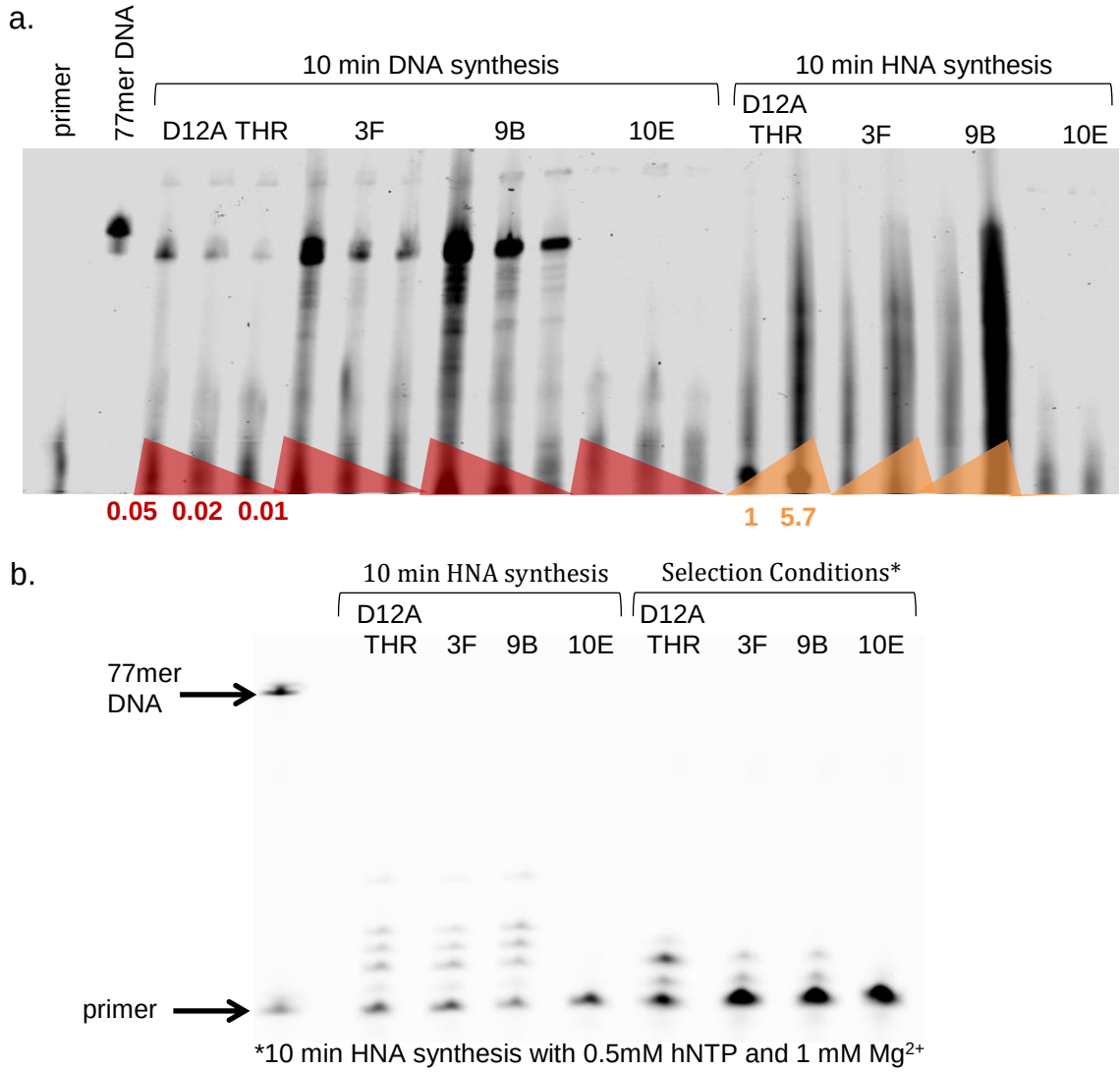


Figure 3.8: Screening TPR2 InDel R1c4 variants. (a) 10 min DNA and HNA primer extensions of 3F, 9B, 10E and D12A THR with decreasing or increasing protein concentrations represented with the red and orange triangles, values indicate μL used per reaction. (b) 10 min HNA primer extensions of 3F, 9B, 10E and D12A THR in standard conditions and with 'selection conditions' (0.5mM hNTP and 1 mM Mg2+) with standardized protein concentrations.

3.2.6 Deep Mutational Scanning of stringent selections

Exonuclease, TPR2 and thumb libraries subjected to the most stringent first round of selection (R1c4) along with their starting population (R0) were prepared and sent for deep sequencing as described in Section 2.4. Although the bulk activity of the selections shows lower activity than wild type and isolated clones shown wild type-like activity, it is still possible that rare clones displaying enhanced HNA synthesis are enriching but not enough after one round of selection to be detectable or easily isolated. Thus, by comparing the frequency of variants from R0 to R1, one should be able to quantify the enrichment of individual variants and compare it to the enrichment of the wild-type. Deep mutational scanning not only allows the identification of variants that were enriched more than the wild-type, with potentially enhanced HNA synthesis activity, but also variants that were significantly depleted, both of which contribute to our understanding of the functional sequence space of phi29 DNAP.

It was expected that the exonuclease library contained 1 - 3 amino acid insertions or deletions, the TPR2 InDel library 1 - 3 amino acid insertions or 1 - 2 deletions and the thumb InDel library 1 - 3 amino acid insertions or 1 - 4 deletions. However, as shown in tables 3.1 – 3.3, the deep sequencing results show a larger than expected variation in the population, including frame shifts and unplanned indels. In fact, a large proportion of variants in all libraries contained frameshifts due to the incorporation or deletion of a number of bases not divisible by three. Frameshifts could have partitioned as background due to unspecific binding during selection. All the frameshifts were combined as a single category termed “fs” in tables 3.1 – 3.3.

InDel size (aa)	Counts R0	Frequency R0	Counts R1	Frequency R1	Enrichment Ratio	Fitness Score	Z-score
2	9002	2.60E-01	1308	2.10E-01	8.08E-01	0.48	-8.370*
3	7969	2.30E-01	1072	1.72E-01	7.48E-01	0.45	-10.167*
-2	1441	4.16E-02	268	4.30E-02	1.03E+00	0.62	0.507
0	313	9.03E-03	94	1.51E-02	1.67E+00	1.00	4.425*
1	277	7.99E-03	39	6.25E-03	7.82E-01	0.47	-1.444
-3	128	3.69E-03	18	2.89E-03	7.82E-01	0.47	-0.984
-1	19	5.48E-04	4	6.41E-04	1.17E+00	0.70	0.286
-4	5	1.44E-04	1	1.60E-04	1.11E+00	0.67	0.097
4	2	5.77E-05	0	0.00E+00	0.00E+00	0.00	-0.600
fs	15517	4.48E-01	3435	5.51E-01	1.23E+00	**0.74**	15.027*
sum	34673	1	6239	1			

* p < 0.05; 5% critical value = ±1.96

Table 3.1: Exonuclease loop length enrichment scores for HNA synthesis. Counts and frequencies of amino acid codon insertions/deletions and frameshifts (fs) in the starting (R0) and selected (R1) library, enrichment ratios, fitness scores (enrichments divided by wild-type enrichment) and output Z-scores of two-tailed proportions Z-tests. In column 1, negative sign indicates deletion of codon; lack of sign indicates insertions and 0 represents wild-type length. Negative Z-score indicates depletion. Wild type estimates are shown in red. Highest fitness score excluding wild type is indicated in bold.

InDel size (aa)	Counts R0	Frequency R0	Counts R1	Frequency R1	Enrichment Ratio	Fitness Score	Z-score
-1	31341	6.37E-01	29770	6.19E-01	9.72E-01	0.35	-5.801*
1	6022	1.22E-01	5441	1.13E-01	9.24E-01	0.33	-4.478*
3	2834	5.76E-02	3598	7.48E-02	1.30E+00	0.47	10.805*
0	198	4.03E-03	535	1.11E-02	2.76E+00	1.00	12.805*
2	50	1.02E-03	52	1.08E-03	1.06E+00	0.39	0.313
-2	33	6.71E-04	59	1.23E-03	1.83E+00	**0.66**	2.822*
-3	23	4.68E-04	32	6.66E-04	1.42E+00	0.52	1.299
4	2	4.07E-05	0	0.00E+00	0.00E+00	0.00	-1.398
fs	8687	1.77E-01	8594	1.79E-01	1.01E+00	0.37	0.873
sum	49190	1	48081	1			

* p < 0.05; 5% critical value = ±1.96

Table 3.2: TPR2 loop length enrichment scores for HNA synthesis. Counts and frequencies of amino acid codon insertions/deletions and frameshifts (fs) in the starting (R0) and selected (R1) library, enrichment ratios, fitness scores (enrichments divided by wild-type enrichment) and output Z-scores of two-tailed proportions Z-tests. In column 1, negative sign indicates deletion of codon; lack of sign indicates insertions and 0 represents wild-type length. Negative Z-score indicates depletion. Wild type estimates are shown in red. Highest fitness score excluding wild type is indicated in bold.

InDel size (aa)	Counts R0	Frequency R0	Counts R1	Frequency R1	Enrichment Ratio	Fitness Score	Z-score
-4	8705	2.50E-01	11951	2.48E-01	9.93E-01	0.48	-0.612
-3	7334	2.11E-01	9521	1.98E-01	9.39E-01	0.46	-4.572*
3	4945	1.42E-01	6814	1.41E-01	9.96E-01	0.48	-0.219
-2	1300	3.73E-02	2045	4.24E-02	1.14E+00	0.55	3.705*
2	297	8.53E-03	381	7.91E-03	9.27E-01	0.45	-0.977
0	120	3.44E-03	342	7.10E-03	2.06E+00	1.00	6.982*
-1	97	2.78E-03	297	6.16E-03	2.21E+00	**1.07**	6.992*
1	71	2.04E-03	79	1.64E-03	8.04E-01	0.39	-1.335
-5	48	1.38E-03	81	1.68E-03	1.22E+00	0.59	1.094
4	0	0.00E+00	1	2.08E-05	0**	0**	0.850
fs	11920	3.42E-01	16674	3.46E-01	1.01E+00	0.49	1.158
sum	34837	1	48186	1			

* $p < 0.05$; 5% critical value = ±1.96; ** invalid - assumed as 0

Table 3.3: Thumb loop length enrichment scores for HNA synthesis. Counts and frequencies of amino acid codon insertions/deletions and frameshifts (fs) in the starting (R0) and selected (R1) library, enrichment ratios, fitness scores (enrichments divided by wild-type enrichment) and output Z-scores of two-tailed proportions Z-tests. In column 1, negative sign indicates deletion of codon; lack of sign indicates insertions and 0 represents wild-type length. Negative Z-score indicates depletion. Wild type estimates are shown in red. Highest fitness score excluding wild type is indicated in bold. The 3 amino acid deletion (P562-G564del) appears to have been significantly depleted compared to the wild-type with a Z-score ($p < 0.05$) of -4.572. 1, 2 and 5 amino acid deletions, 1, 2 and 4 amino acid insertions and frameshifts as well as the wild-type loop were enriched after 1 round of selection, from which only 1-2 amino acid deletions were enriched with statistical significance ($p < 0.05$).

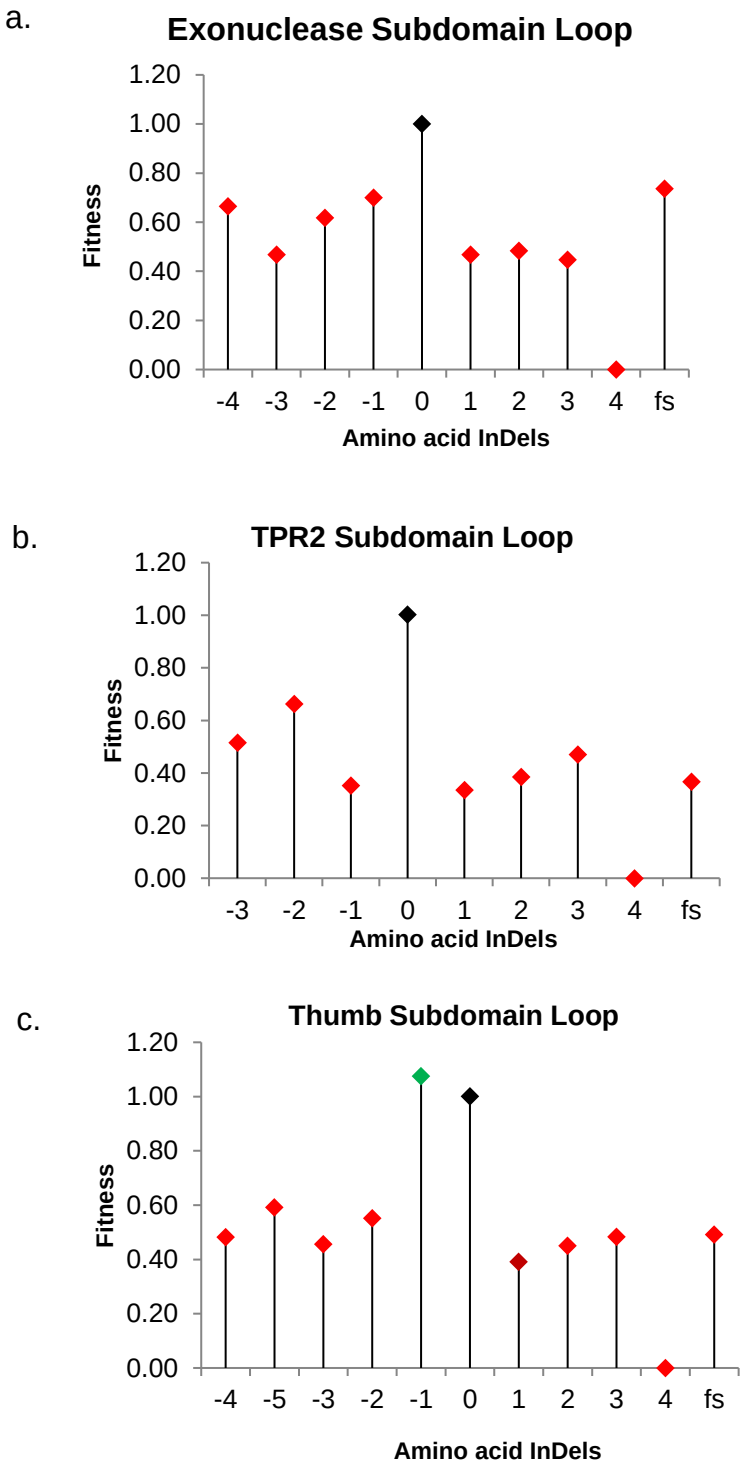


Figure 3.9: Phi29 DNAP loop length fitness for HNA synthesis. Exonuclease, TPR2 and thumb loop length fitness compared to the wild-type length. Negative sign indicates deletion of codons; lack of sign indicates insertions; 0 refers to the wild-type length. InDels with fitness scores lower than the wild-type are shown in red and scores higher than the wild-type are shown in green.

To account for the nonuniform abundance of variants in the input library and identify enriched mutations, enrichment ratios comparing the frequency of mutations post-selection to their frequency in the starting R0 population were calculated. An enrichment ratio >1 suggests enrichment and an enrichment ratio <1 suggests depletion.

Although the most sampled InDel mutations in the exonuclease loop post-selection was a 2 amino acid insertion (Table 3.1), enrichment ratios suggests that only 1, 2 and 4 amino acid deletions, wild-type, and frameshifts were enriched from R0. Comparing these enrichment ratios to that of the wild-type (Figure 3.9a), however, suggests that none of the variants have significantly improved fitness. Additionally, the Z-score test performed revealed that only the frameshifts and wild-type had significant ($p < 0.05$) enrichment. Interestingly, the test also revealed that 2 and 3 amino acid insertions were significantly depleted. Thus, the frameshift mutants and 2 amino acid mutations were chosen for further inspection.

The frameshifts (fs) were the second-most ‘active’ population after the wild-type with a fitness score of 0.74 and significant Z-score ($p < 0.05$) of 15.027. In the TPR2 or Thumb selections, however, significant frameshift enrichment is not observed. Analysing the frameshifts of the exonuclease domain library in greater detail (Figure 3.10), suggests that most of the fs mutations are actually depleting, with the exception of a 7bp deletion that appears to be significantly ($p < 0.05$) enriched with an enrichment ratio of 1.148 and significance score of 6.389 ($p < 0.05$). There is some diversity within the 7bp deletion population after selection (Figure 3.10c), but in all cases the mutations result in a premature opal (TGA) stop codon in a position that almost overlaps with the in-frame Met-30 (Figure 3.10b). It is possible protein synthesis reinitiates from the in-frame Met-30 resulting in an active split phi29 DNAP variant. This has been previously observed through the directed evolution of aminoacyl-tRNA synthetases and *pfu* DNAP where highly active split variants from frameshift mutations were preferentially selected [96, 123]. In particular, the split *pfu* DNAP showed improved incorporation of a bulky nucleotide *γ*-phosphate-*O*-linker-dabcyl substituent [123], thus a split phi29 DNAP could have improved HNA-DNA duplex stabilisation, thereby explaining its enrichment during selection. The 7bp deletion variant could be isolated and tested for

polymerase activity to corroborate that its selection is in fact due to enhanced HNA synthesis activity and then test polymerase activity reconstituted from separate fragments to ensure that enhanced activity (if any) is the result of the split nature of the protein and not read-through at the premature stop codon.

InDel (bp)	ExoR0	ExoR1c4	Enrichment Ratio	Z-scores
-7	0.399	0.458	1.148	6.389*
8	0.229	0.193	0.842	-2.304*
5	0.227	0.211	0.929	-4.382*
1	0.045	0.043	0.968	-0.363
2	0.027	0.027	1.001	0.013
4	0.022	0.020	0.917	0.183
-8	0.018	0.022	1.256	0.886
7	0.018	0.010	0.589	-3.045*

* $p < 0.05$; 5% critical value = ±1.96

Table 3.4: Enrichment for HNA synthesis of frameshifts in the exonuclease loop. Frequencies of frameshift in the starting (R0) and selected (R1) library with read count >100, enrichment ratios and output Z-scores of two-tailed proportions Z-tests. Negative Z-score indicates depletion.

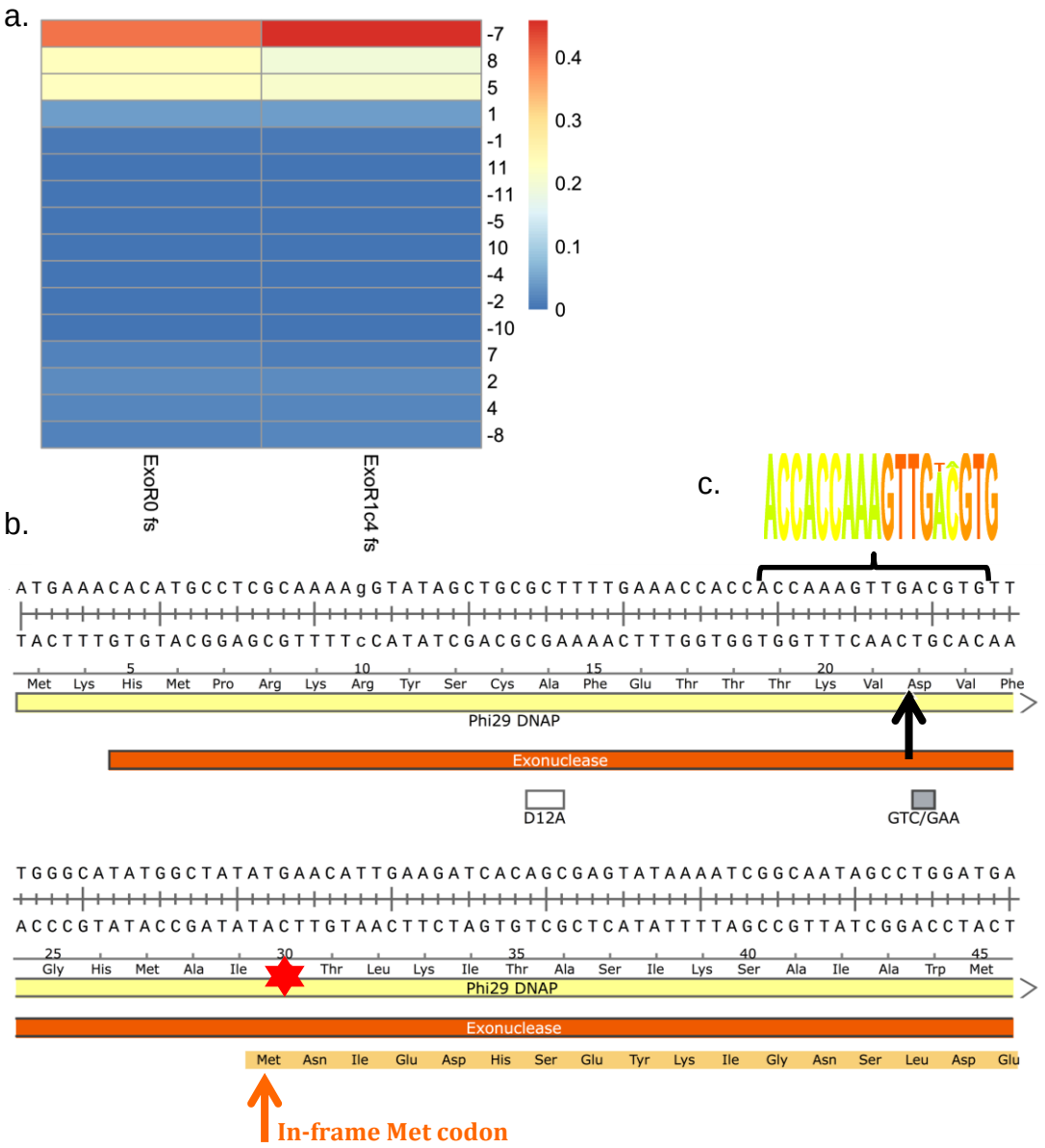


Figure 3.10: Exonuclease loop frameshifts scanning. (a) Heat map showing frequencies of each frameshift mutation pre- (ExoR0 fs) and post- (ExoR1c4 fs) selection for HNA synthesis. (b) Sequence of the fs mutation with highest enrichment score corresponding to a 7bp deletion. Black arrow indicates start of fs, red star indicates introduction of an early stop codon and orange arrow indicates the in-frame met codon. (c) Sequence logo displaying sequence conservation of the population with frameshift mutations post-HNA selection.

As mentioned previously, the exonuclease domain loop does not appear to tolerate 2 to 3 amino acid insertions as shown with enrichment ratios < 1 and significant Z-scores ($p < 0.05$) of -8.370 and -10.167 respectively. Insertions were generated with NNS codons, so it is possible that some amino acid insertions are more detrimental than others. Looking at the population carrying only 2 insertions in the exonuclease domain loop, a high degree of diversity and no conservation for a particular amino acid pre- or post-selection can be observed (Figure 3.11).

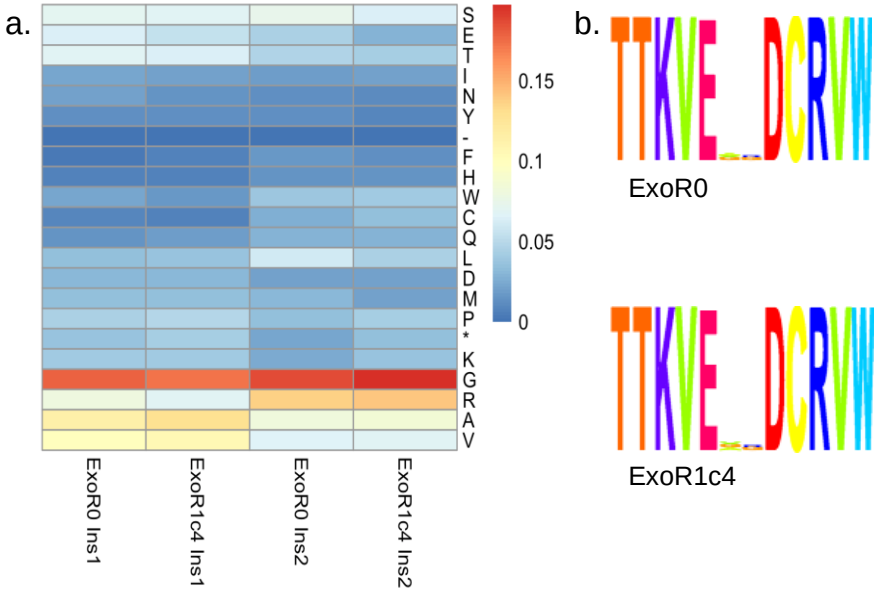


Figure 3.11: 2 amino acid insertion diversity in the exonuclease loop. (a) Heat map showing frequencies of each amino acid in the first (Ins1) and second (Ins2) position pre- (ExoR0) and post- (ExoR1c4) selection for HNA synthesis. (b) Sequence logo displaying sequence conservation of the population with 2 amino acid insertions pre- and post- selection.

Looking at the enrichment scores of 2 amino acid insertions (See Table 3.5), it appears that half of the mutations were depleted and the other half enriched, albeit with no statistical significance in most cases. The first insertion (Ins1) does not show any significant depletion of a particular residue, but the second insertion (Ins2) shows that glutamic acid, leucine and methionine residues are significantly depleted. Interestingly, some insertions also show significant enrichment; the first insertion appears to prefer a phenylalanine and the second insertion either a stop codon or a lysine. Although the 2 amino acid

insertion population was collectively depleted, these three insertions appear to be depleted to a lesser extent and could therefore be more tolerated.

Although 2 and 3 insertions were depleted, they still form the majority of the population after one round of selection, which could explain the poor activity of the library pre- and post-selection compared to wild-type. All other InDel mutations in the exonuclease domain loop including 1 and 4 amino acid insertions and 1, 2, 3, 4 amino acid deletions were not significantly enriched or depleted from R0 and thus appear to not interfere with the overall protein fitness significantly. It is possible that subsequent rounds of selection that further deplete variants with 2 and 3 insertions would result in an overall population activity closer to the wild-type. However, the population size after one round of selection analysed (6,239) is smaller than the theoretical library size (8,423), so it is possible that the observed frequencies of variants post-selection do not represent the true distribution of variants. Although the DNA concentration of all libraries was standardised prior to deep sequencing, DNA concentrations were quantified with a Nanodrop, which in hindsight, lacks specificity and sensitivity compared to other platforms such as Qubit or qPCR. Nonetheless, these observations suggest that amino acid insertions and deletions in the exonuclease domain loop of Phi29 DNAP do not appear to significantly improve HNA synthesis and 2 - 3 codon insertions negatively impact its HNA synthesis activity.

Insertion (aa)	ExoR0 Ins1	ExoR1c4 Ins1	Enrichment Ratio	Z-scores Ins1	ExoR0 Ins2	ExoR1c4 Ins2	Enrichment Ratio	Z-scores Ins2
*	0.037	0.042	1.13	0.84	0.022	0.034	1.53	2.64*
A	0.113	0.128	1.14	1.66	0.083	0.085	1.03	0.27
C	0.009	0.008	0.84	-0.53	0.027	0.034	1.26	1.47
D	0.030	0.030	0.99	-0.06	0.021	0.021	0.97	-0.16
E	0.064	0.054	0.84	-1.40	0.044	0.029	0.66	-2.55*
F	0.003	0.008	2.65	2.73*	0.017	0.012	0.74	-1.16
G	0.179	0.174	0.97	-0.45	0.186	0.197	1.06	0.97
H	0.007	0.006	0.89	-0.32	0.015	0.015	1.00	-0.01
I	0.022	0.021	0.98	-0.08	0.019	0.021	1.13	0.59
K	0.040	0.041	1.02	0.13	0.025	0.037	1.51	2.65*
L	0.034	0.037	1.09	0.57	0.060	0.044	0.74	-2.28*
M	0.034	0.034	1.02	0.10	0.031	0.021	0.68	-1.96*
N	0.020	0.015	0.76	-1.18	0.013	0.011	0.87	-0.52
P	0.043	0.049	1.13	0.90	0.035	0.042	1.21	1.30
Q	0.015	0.019	1.30	1.23	0.028	0.029	1.02	0.12
R	0.080	0.068	0.85	-1.50	0.137	0.141	1.03	0.34
S	0.071	0.067	0.95	-0.47	0.074	0.064	0.87	-1.27
T	0.067	0.065	0.97	-0.32	0.046	0.043	0.94	-0.46
V	0.099	0.106	1.08	0.86	0.065	0.067	1.03	0.28
W	0.022	0.016	0.72	-1.46	0.038	0.041	1.07	0.50
Y	0.012	0.012	0.99	-0.03	0.013	0.009	0.72	-1.07

* $p < 0.05$; 5% critical value = ±1.96

Table 3.5 Enrichment for HNA synthesis of 2 amino acid insertions in the exonuclease loop. Frequencies of each amino acid in the first (Ins1) and second (Ins2) position pre- (ExoR0) and post- (ExoR1c4) selection for HNA synthesis, enrichment ratios and output Z-scores of two-tailed proportions Z-tests (5% critical value = 1.96). Negative Z-score indicates depletion.

The TPR2 InDel library performed the best before and after selection in terms of HNA synthesis compared to the other two loops. From the small-scale screening, the N409del appeared to be the most frequent mutation, but had little impact in activity compared to the wild-type D12A THR. The deep sequencing results corroborated that the N409del was the most abundant mutation after selection, but also in the starting library (Table 3.2). N409del has an enrichment ratio below 1 and a significant Z-score ($p < 0.05$) of -5.801, which suggests that this mutation was in fact not enriched but significantly depleted. A 1 amino acid insertion in the TPR2 loop also appears to be significantly depleted post-selection. Variants that were significantly enriched include a 3 amino acid insertion and a 2 amino acid deletion. Still, comparing enrichment ratios to that of the wild-type, indicates that neither of these two variants have

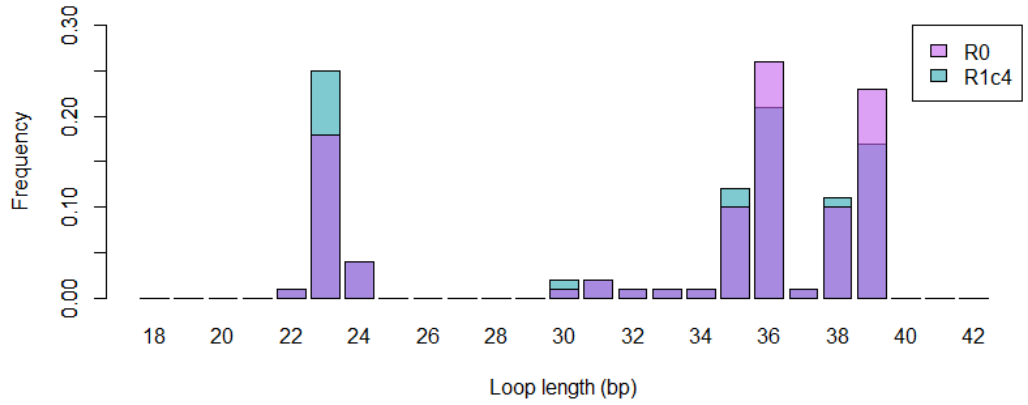
significant improved fitness (Figure 3.9b). The 2 amino acid deletion (N409-G410del) is the second-most ‘active’ variant after the wild-type with a fitness score of 0.66, which is even lower than that of the exonuclease domain loop frameshifts (0.74). The wild-type TPR2 loop appears to be significantly preferred over 1 - 4 amino acid insertions or 1 - 3 amino acid deletions. Thus, the InDels in the TPR2 loop investigated here do not appear to enhance HNA synthesis.

The thumb loop library selection displayed better activity than the exonuclease loop selection but less than the TPR2 library selection. Results and observations of the deep sequencing data of the thumb loop library are described under Figure 3.3. From these observations only a 1 amino acid deletion (P562del) displays enhanced fitness when compared to the wild-type (Figure 3.9c). The P562del of the thumb subdomain loop thus appear to be statistically more ‘active’ than the wild-type. This variant has not yet been isolated and its polymerase activity has not been tested, but it would be the next step to corroborate if the observed enrichment over the wild type is due to enhanced HNA synthesis activity.

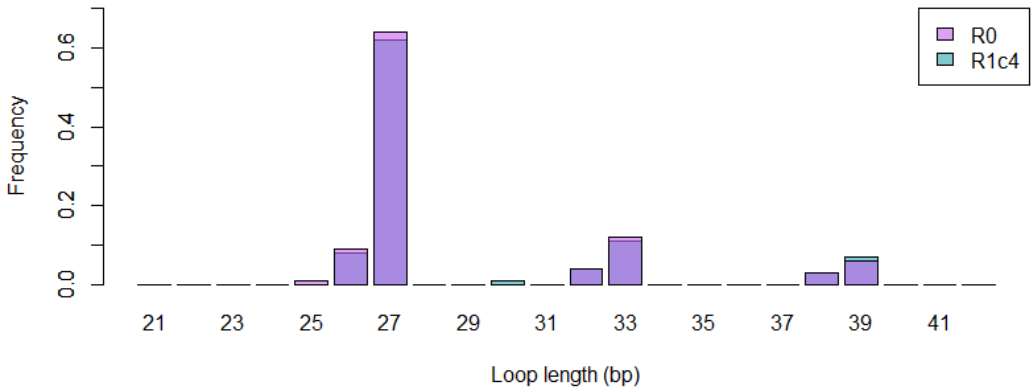
Although the deep sequencing analysis gave interesting insights into the functional landscape of phi29 DNAP, there are some caveats with regards to the statistical analysis of the data. As described in Section 2.4.3, two proportions Z-test [124] was used to score the significance of enrichment scores by directly comparing two proportions (proportions of each InDel mutation) from two independent populations (R0 and R1). A pooled version of the test was used assuming that the distribution of R0 and R1 was sufficiently similar to be averaged in the estimation of the standard error, which, as shown in Figure 3.12, turned out to be the case. This test can be used when the sampling distribution is approximately normal and since the sample size was sufficiently large (n>30) [97], the test was chosen. However, the density plots of loop lengths pre- and post- selection (Figure 3.12), show that the three libraries follow a multimodal distribution, which thus hinders the estimation of significant ratios based on the assumption of normality. While greater sampling may lead to a regression to normality, the sampling generated for at least some libraries suggest that the lack of normality is representative of the data. Thus, statistical analysis that do not make assumptions about population distribution, such as

the nonlinear least-squares method [98] or other non-parametric tests such as the Kruskal–Wallis test [99] by ranks are better suited for analysis. Due to time constraints these tests were not implemented. On the other hand, multimodality in the data could also arise from sampling bias. Thus, it is possible that we are in fact not observing an accurate representation of the full population. Sampling more variants of the population could help accurately determine rationale behind the observed distribution. Nonetheless, putting aside the statistical significance of enrichment, the enrichment and fitness scores obtained give a clear understanding of the overall fitness phi29 DNAP when introducing InDels to the exonuclease, TPR2 and Thumb loops studied here.

a. Exonuclease Loop Length Selection



b. TPR2 Loop Length Selection



c. Thumb Loop Length Selection

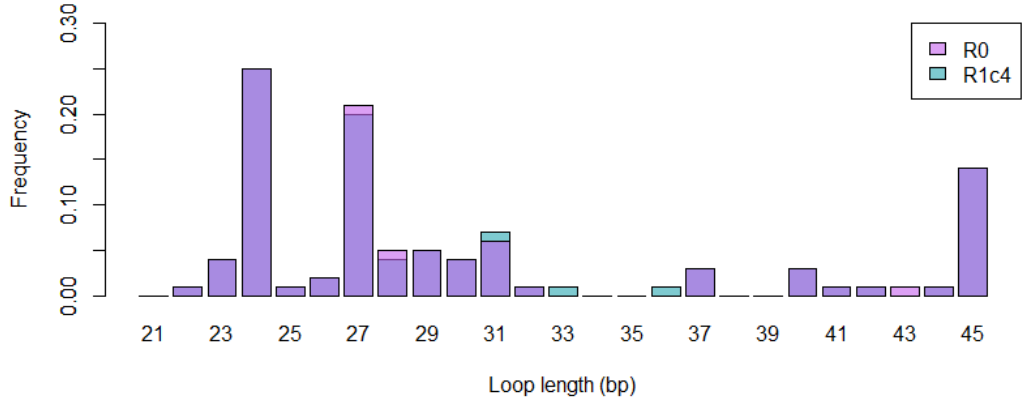


Figure 3.12: Histograms of exonuclease, TPR2 and thumb loop length selections. Frequencies pre- and post-selection for HNA synthesis of loop lengths in the (a) exonuclease (wild type length is 30bp), (b) TPR2 (wild type length is 30bp) and (c) thumb populations (wild type length is 36bp).

3.3 Conclusions

Three loops that come near the nascent duplex during HNA synthesis belonging to the exonuclease domain, involved in proofreading, TPR2 subdomain, involved in processivity and strand displacement, and thumb subdomain, involved in polymerase activity, were targeted through InDel mutagenesis. InDels were introduced to these loops with the aim of modifying phi29 DNAP dynamics to enhance its HNA synthesis activity. Shifts in polymerase dynamics that favour synthesis, disfavour exonuclease activity, facilitates template-binding or stabilises the nascent duplex during synthesis, all could contribute to enhanced HNA synthesis. From the libraries generated, the exonuclease domain loop appears to prefer its wild-type length as no HNA synthesis enhancement was observed and no variants appeared to be significantly enriched over the wild-type. On the other hand, insertions did appear to be significantly detrimental to function, something that was not as distinctly observed in the TPR2 or Thumb selections.

The TPR2 library, although performed the best during primer extension assays post-selection compared to the exonuclease and thumb loop libraries, the activity of isolated variants did not surpass that of the wild-type. Deep sequencing revealed that none of the variants were significantly enriched. However, the most enriched variant (N409-G410del) after the wild-type could be further characterised to see if the activity is in fact below the wild-type. Most of the variants randomly selected for screening were a N409del due to the fact that this mutation was the most frequent in the starting library but was in fact depleted during selection. This clearly suggests that deep sequencing could be carried out prior to screening to shortlist potential active variants. Overall, the InDels introduced to the TPR2 loop do not seem to instigate a significant detrimental effect but also does not appear to enhance the HNA activity of phi29 DNAP. Thus, the TPR2 loop appears to tolerate a wider range of lengths than the exonuclease and thumb loops analysed.

The thumb subdomain loop library performed with slightly less activity than the TPR2 library during HNA synthesis assays and contained a couple of mutations that were significantly enriched and depleted. One of these variants, P562del, showed enrichment above the wild-type, and thus should be further characterised to corroborate that it in fact enhances HNA synthesis. Both

deletions and insertions apart from the P562del, seem to impose a detrimental effect to polymerase function, particularly insertions albeit with no significant statistical significance and to a lesser extent than insertions in the exonuclease loop.

Overall, the introduction of InDels to the loops of the exonuclease domain and TPR2 and thumb subdomains resulted in varied effects, as expected, and gave an interesting insight into the mutational robustness (or mutational tolerance) of these loops. It appears that the exonuclease loop does not tolerate InDels, particularly insertions, suggesting that it is constrained to a limited conformational space and increasing its size potentially causes significant steric hindrance to the thumb. The thumb loop, positioned adjacent to the exonuclease loop, also did not tolerate insertions, which also corroborates the spatial constraint of the exonuclease loop. The thumb loop in fact seemed to tolerate or benefit from a one amino acid deletion, which potentially liberates some tension with the exonuclease loop/domain and could be stabilising the neighbouring TPR2 loop or nascent duplex. Modifying the TPR2 loop seems to not cause significant steric hindrance with the neighbouring loops, suggesting it does not significantly interact with the thumb subdomain or exonuclease domain. Further characterisation of the findings presented here should be carried out to identify the role of each of these loops in terms of polymerase dynamics to inform further phi29 DNAP engineering as well as further characterisation of variants that showed significant enrichment during selection for HNA synthesis. Additionally, the generality of this approach could be used to investigate other loop-loop interaction in phi29 DNAP, as well as other polymerases.

4. Multiple-site saturation mutagenesis of phi29 DNAP finger subdomain

4.1 Introduction

Most of our understanding of polymerase dynamics has originated from bulk biophysical assays and crystal structures at different stages of catalysis that may not capture the true or complete spectrum of properties and mechanisms of individual polymerases [125]. It has now been shown that instead of a rate-limiting step involving a thumb translocation that switches the polymerase conformation from ‘open’ to ‘closed’, multiple polymerase conformations exist, where, upon nucleotide triphosphate binding, the closed conformation is stabilised [125].

It has also been shown that non-canonical or incorrect nucleotides fail to stabilise the closed conformation, resulting in lower catalytic efficiency [125], thus a potential approach to enhance XNA synthesis and processivity is to increase the stability of the polymerases in its ternary catalytically active conformation and shift the equilibrium towards this state. Proteins that can adopt multiple thermodynamically stable conformations often require the formation and disruption of specific interdomain contacts [100]. It has been demonstrated that evolved thermostable polymerases able to synthesise or reverse-transcribe C2’-OMe modified oligonucleotides, are able to do so due to optimised interdomain interactions between the fingers and thumb subdomains that stabilise their catalytically active closed conformation [74]. Thus, optimising the interdomain contacts of phi29 DNAP to stabilise its ternary complex conformation could potentially enhance its HNA synthesis ability and help map the functional sequence landscape of phi29 DNAP.

Still, protein dynamics are complex, and residues can display synergistic epistasis (where the combined effect of mutations is stronger than the product of their individual effects), or antagonistic epistasis (where the combined effect of mutations is smaller than the product of their individual effects) [101, 102]. Targeting residues involved in interdomain contacts through traditional protein engineering approaches involving single-point mutations could allow the

determination of the immediate protein sequence landscapes but will still remain a very poor predictor of the epistatic effects of interactions and co-variation between residues. Instead, targeting multiple residues simultaneously coupled to DNA- and XNA-based selection platforms, could allow the reconstruction of networks of interaction that can help improve protein engineering while identifying more efficient XNA polymerases.

To identify residues involved in interdomain interactions, interdomain contacts in the apo structure (PDB: 1XHX) were compared to those in the ternary conformation (PDB: 2PYJ). Pymol and the WHAT IF Server were used to identify interdomain contacts (where a contact is defined as two atoms for which the distance between the Van der Waals surfaces is less than 0.25 Ångstrom) between all atoms of the finger, palm and thumb subdomains of phi29 DNAP. These subdomains form the internal clamp that accommodates and stabilises the nascent DNA duplex and thus are heavily involved in the polymerase dynamics and processivity [95]. More than 36 sites were identified to be involved in interdomain contacts at the palm, thumb and finger subdomains when Phi29 DNAP is in its ternary conformation (See Figure 4.1). To target all these sites and generate a library containing multiple mutations per gene, Darwin Assembly, a robust and efficient library assembly method, was used. Darwin assembly allows the generation of large, high quality and complex libraries with over 10^8 transformants targeting more than 10 distal sites [71]. From the identified targets, an initial library targeting 16 residues in the finger subdomain that appeared to make contacts with the catalytic palm subdomain was designed and constructed through Darwin Assembly (See Section 2.3.1) on a thermostabilised phi29 DNAP (D12A THR) background [87]. Positions where mutations had been previously identified to increase the thermostability of phi29 DNA were not targeted to preserve the thermostabilised phenotype as a background as it should make phi29 DNAP less susceptible to otherwise destabilising mutations.

4.2 Results and Discussion

4.2.1 Library design

As described in Section 2.3.1, Darwin Assembly requires three groups of oligonucleotides: inner (mutagenic), boundary and outnest oligonucleotides. Inner oligonucleotides were designed according to a site-saturation mutagenesis strategy that ensures an even distribution of amino acids and avoids stop or rare codons [75]. With this approach, four complementary oligonucleotides per target site containing either the codon NDT, VMA, ATG or TGG in place of the target codon are mixed in a ratio of 12:6:1:1 respectively. This results in one degenerate codon per amino acid. Three different groups of long inner oligonucleotides denoted F1, F2 and F3, were designed to cover all 16-target sites involved in interdomain contacts of the finger domain.

F1 inner oligonucleotides targeted 5 sites, F2 oligonucleotides targeted 5 sites and F3 oligonucleotides targeted 6 sites. To target all residues and meet the requirements for the inner oligonucleotide design, the designed F1 and F2 oligonucleotides partially overlap (see Figure 4.1). Thus, the library was expected to contain 2 mutations per gene and a size of 2.4^4 variants 2(20 residues x 5 sites) x (20 residues x 6 sites). Theta and biotinylated boundary oligonucleotides and corresponding outnesting oligonucleotides were designed as described in Section 2.3.1.

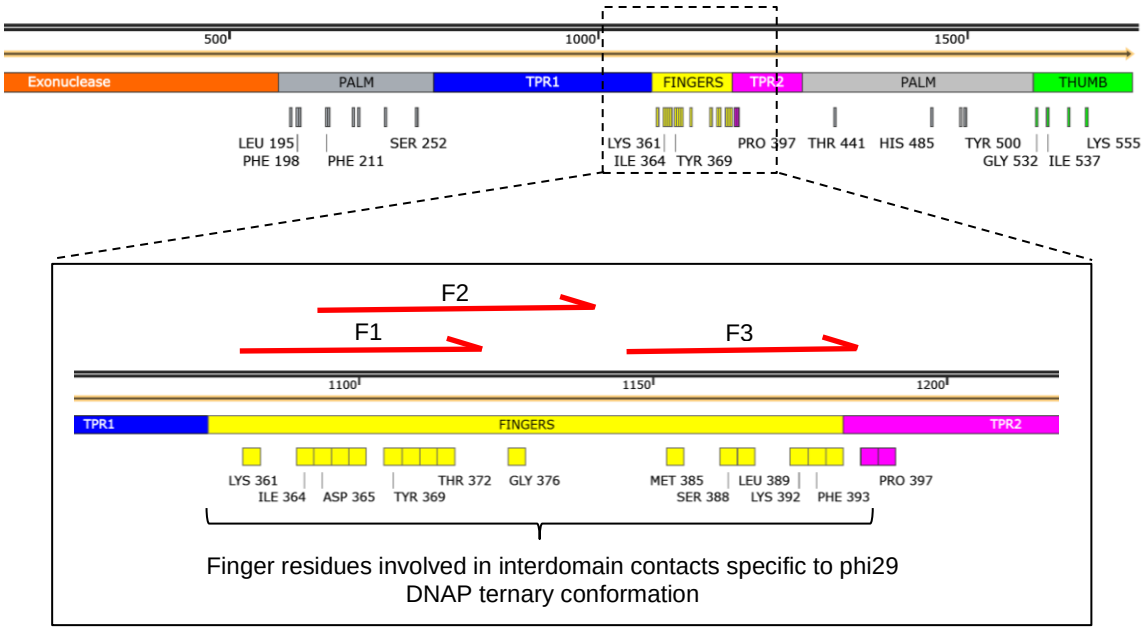


Figure 4.1: Interdomain contacts of Phi29 DNAP in its ternary conformation. Top panel shows residues involved in interdomain contacts (0.25 ångstrom distance from its pair) below the phi29 DNAP gene. Bottom panel zooms into the finger subdomain and additionally shows the three groups of long inner oligonucleotides (F1, F2 and F3) in red as well as the partial overlap between F1 and F2.

4.2.2 Library construction

The first step in Darwin Assembly is the single stranded plasmid generation. An Nt.BspQI nicking site was introduced through PCR as described in Section 2.3.1 (and Appendix B) in the pET23 plasmid harbouring the D12A THR encoding gene. The new plasmid and original plasmid were digested with Nt.BspQI and ExoIII as described in Section 2.3.1 and run on an agarose gel stained with SYBR gold (Thermo Fisher Scientific) for the visualisation of single stranded plasmid. As shown in Figure 4.2a, the original plasmid does not appear to have been affected by the nicking enzyme or exonuclease. The template carrying the nicking site, on the other hand, does appear to be nicked and degraded by the exonuclease as shown with the faint smear in Figure 4.2b. The pET23 BspQI D12 THR was used for the subsequent steps of Darwin Assembly.

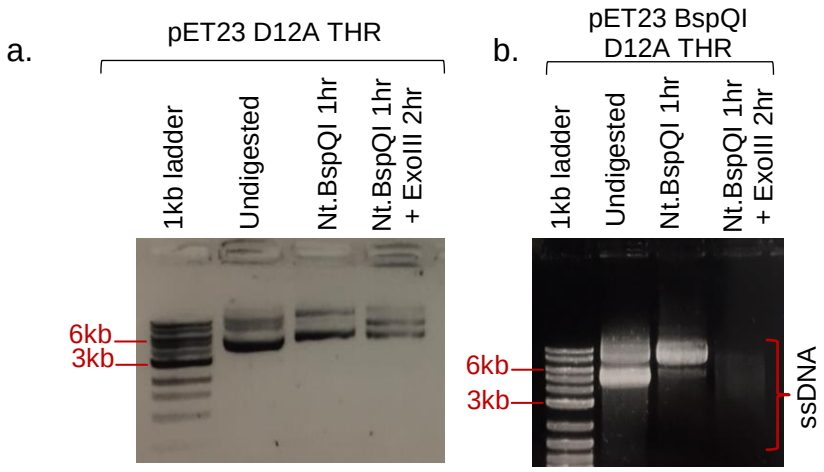


Figure 4.2: Nicking site introduction and single strand generation. Nt.BspQI and ExoIII digestion of the pET23 plasmid harbouring the phi29 DNAP D12A THR gene (expected 5.4kb band size) pre- (a) and post- (b) introduction of the Nt.BspQI restriction. Only ssDNA is observed after the introduction of the nicking.

4.2.2.1 Assembly with the theta oligonucleotide

Darwin Assembly was initially carried out with the theta oligonucleotide and a subset of inner oligonucleotides targeting 3 out of the 16 sites (see Figure 4.3a). As shown in Figure 4.3b the PCR amplification of the recovered product post-assembly appears to be larger (5kb) than the expected amplicon size (2kb). A negative control consisting of an assembly with no inner and no boundary oligonucleotides was also carried out to rule-out non-specific recovery from the outnest oligonucleotides; this negative control did not result in amplification (Figure 4.3a). Assembly with all inner oligonucleotides targeting all 16 sites was then attempted, but as shown in Figure 4.3c, no recovery was observed.

To optimise the assembly strategy, numerous modifications to the original Darwin Assembly protocol were tested. A small binding region of the F1 and F2 oligonucleotides overlapped, so it was hypothesised that partially bound oligonucleotides could be hindering KOD during amplification of the assembled product. Thus, before KOD amplification, the assembled products were additionally digested with Exonuclease VII and Taq ligase. The Exonuclease VII digests only single-stranded DNA, thus it would digest any partially bound oligonucleotide and the ligase should then seal the nick. It was also hypothesised that the boundary and/or inner oligonucleotides were not efficiently annealing to the template during the freezing-thawing cycle, thus a

heating and cooling oligonucleotide annealing approach was tested (See Section 2.3.1). This annealing approach was also tested in combination with the Exonuclease VII/Taq ligase clean-up step. Lastly, the assembly was also carried out only with F1 and F3 oligonucleotides to see if removing the overlapping oligonucleotides improved recovery.

As shown in Figure 4.3d, all the above-mentioned modifications to the assembly protocol were tested with all inner (F1, F2 and F3) oligonucleotides as well as the original Darwin Assembly conditions with all inner oligonucleotides, F1 and F3 oligonucleotides only or no inner oligonucleotides. Assembly with all inner oligonucleotides under original conditions appear to have resulted in a faint smear, which was not observed previously. The inclusion of Exonuclease VII/Taq ligase clean-up step decreased the smear slightly, indicating the presence of single-stranded template contamination due to excessive template degradation resulting in mispriming. Switching to heating and cooling for oligonucleotide annealing resulted in an even higher degree of non-specific recovery. Including the Exonuclease VII/Taq ligase clean-up step once again decreased the degree of the smear but did not improve the recovery of the desired product. This indicates that heating and cooling improved non-specific annealing onto degraded template. Carrying out the assembly with original conditions and only F1 and F3 oligonucleotides (avoiding overlaps between F1 and F2 oligonucleotides) did not differ significantly from assembly with all inner oligonucleotides. In contrast, assembly with no oligonucleotides and original assembly conditions resulted in the most pronounced smear.

Assembly with these modified parameters resulted in either no recovery or in a high degree of non-specific recovery. Despite optimisation of PCR, which included not only reaction conditions but also different DNA polymerases, no successful amplification of the desired library DNA was obtained. Although the observed smears could originate from DNA contamination, excessive template concentration or template degradation, it is also possible that mispriming is occurring due ineffective binding of the theta oligonucleotide. Rather than optimising the landing pads of the theta oligonucleotide, assembly with biotinylated oligonucleotides was tested.

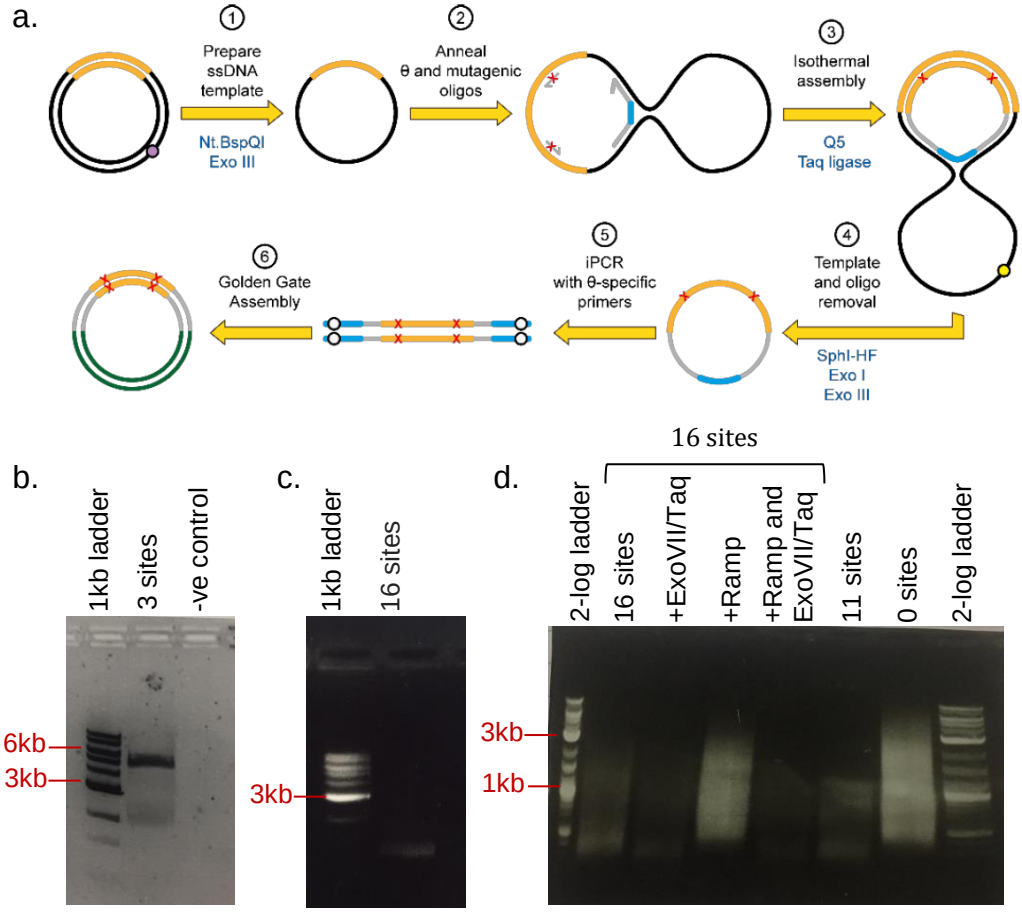
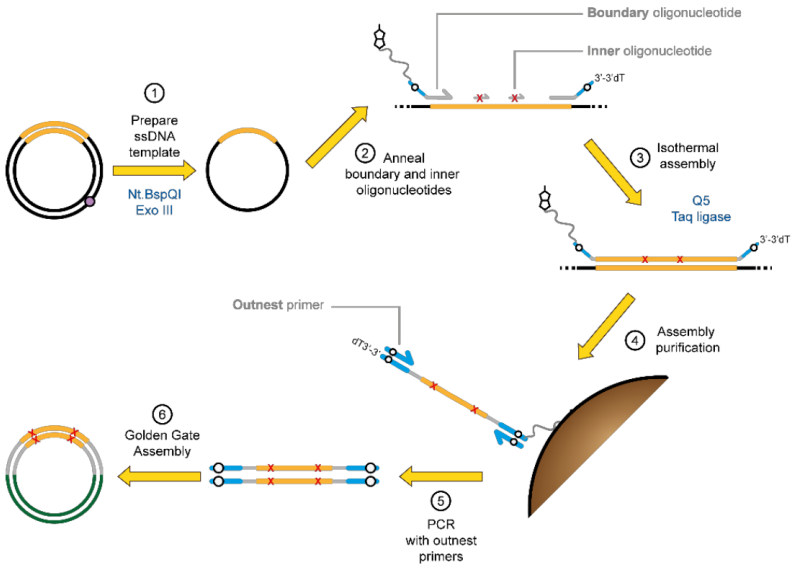


Figure 4.3: Optimising Darwin Assembly with the theta oligonucleotide. (a) Schematic representation of Darwin Assembly with the theta oligonucleotide (adapted from Cozens and Pinheiro 2017). KOD PCR recovery of a (b) 3-site saturation mutagenesis assembly (2 kb expected fragment size) and negative control of an assembly with no boundary and no inner oligonucleotides and (c) 16-site saturation mutagenesis assembly (2 kb expected fragment size). (d) Optimisation of a 16-site saturation mutagenesis assembly (2 kb expected fragment size) incorporating an ExoVII/Taq clean up step and annealing mutagenic oligonucleotides by heating-cooling, 11-site saturation assembly and a control assembly targeting 0 sites.

4.2.2.2 Assembly with biotinylated oligonucleotides

Darwin Assembly was executed with two different pairs of biotinylated boundary oligonucleotides as depicted in Figure 4.4a. As described in Section 2.3.1, assembly with two pairs of boundary oligonucleotides (F1/R1 and F2/R2) carrying distinct priming sites were tested. As shown in Figure 4.4b, assembly with both pairs of boundary oligonucleotides along with inner oligonucleotides targeting 16 sites as well as control assemblies with both pairs of biotinylated and no inner mutagenic oligonucleotides were all successful. Sequencing of a few clones corroborated the introduction of at least 2 mutations per gene at the correct target positions.

a.



b.

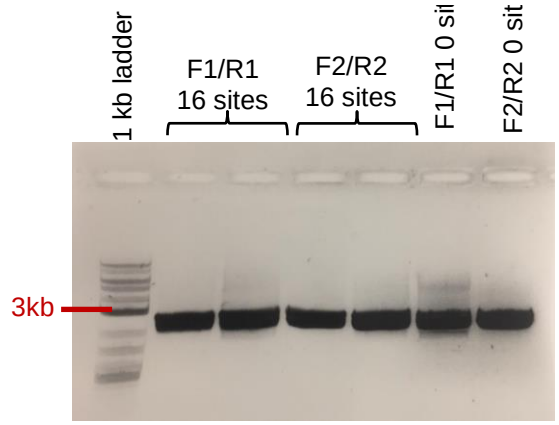


Figure 4.4: Darwin Assembly with biotinylated oligonucleotides. (a) Schematic representation of Darwin Assembly with biotinylated oligonucleotides (adapted from Cozens and Pinheiro, 2017). (b) PCR recovery (expected 2kb band size) of a 16-site saturation mutagenesis of the finger subdomain of phi29 DNAP using two different pairs of biotinylated oligonucleotides (F1/R1 and F2/R2) and inner oligonucleotides targeting 16 sites as well as control reactions with no inner oligonucleotides.

4.2.3 Library Selection

The library was subjected to a first round of selection of 3 hr as described in Section 2.4. In parallel, the highly stringent selection carried out in Chapter 3 consisting of a shorter extension time and lower nucleotide, selection primer and magnesium concentrations was carried out (Figure 4.5a). As mentioned in Chapter 3.2.4, lowering the concentration of hNTPs could push the selection of variants with potentially less substrate specificity able to recognise and incorporate hNTPs more readily. A potential issue with lowering the hNTP concentration, however, is that it could compete with cellular dNTP concentration, facilitating the rise of parasites (undesired phenotypes). Decreasing the selection primer concentration could also push the selection of highly processive variants while reducing the probability of non-specific binding of unextended/poorly-extended primers. As shown in previous primer extension optimisation assays (See Section 3.2.2), decreasing the magnesium concentration to 1 mM had a significant impact in the HNA synthesis activity of THR D12A phi29 DNAP, thus this concentration was used to increase the stringency of the selection. From previous activity assays, it also became clear that D12A THR can synthesise HNA within 10 min, thus 10 min incubation for synthesis during selection was used in combination with the other parameters. As shown in Figure 4.5, selections under standard conditions and stringent conditions both resulted with a broadly similar degree of recovery or both reactions reached saturation and a difference between the reactions is not observable. Activities were not assayed, but both selections and R0 were prepared and sent for deep sequencing.

Selection conditions	R1-3hr (Standard)	R1c4 (Stringent)
Time	3hr	10min
hNTPs	2.5mM	0.5mM
CST_4(7)	10µM	2.5µM
Mg^{2+}	10Mm	1mM

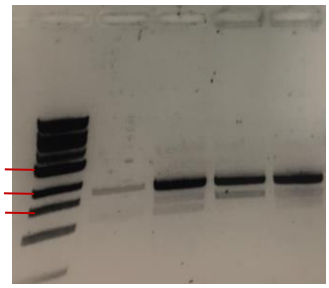


Figure 4.5: Selection for HNA synthesis of the multiple-site saturation mutagenesis of the finger subdomain of phi29 DNAP. (a) Conditions for standard (R1-3hr) and stringent (R1c4) HNA synthesis selection. (b) Q5 PCR recovery (in duplicate) of libraries after one round of selection for HNA synthesis with standard (R1-3hr) and stringent (R1c4) selection conditions (expected amplicon size of 1.7 kb). Duplicate PCR reactions were carried out on the same sample and bands of the correct size were jointly gel purified.

4.2.3 Deep Mutational Scanning

The multiple-site saturation mutagenesis of the finger subdomain through Darwin Assembly was expected to target 16 sites involved in interdomain contacts. The library pre-selection (R0) was prepared and sent for deep sequencing as described in Section 2.4.1.

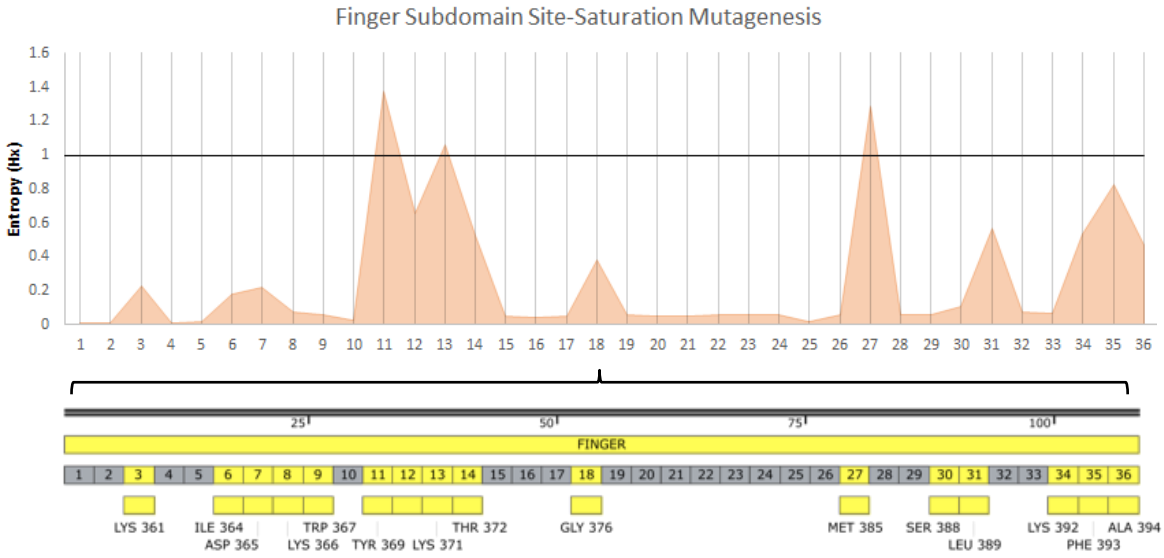


Figure 4.6: Entropy plot of the multiple-site saturation mutagenesis of the finger subdomain of phi29 DNAP. The top panel shows the variation along the multiple-site saturation mutagenesis library of the finger subdomain of phi29 DNAP as an entropy plot (See Section 2.4.4), x-axis refers to the alignment position and y-axis corresponds to the respective entropy values (nits). Bottom panel shows the residues of the finger subdomain represented in the entropy plot; sites targeted during assembly of the library are coloured in yellow.

Following data clean up and the alignment of sequences, entropies were calculated for each position of the finger subdomain to identify the degree of diversity introduced (See Section 2.4.4). The entropy value measures amino acid variation at each position in an alignment, where 0 indicates no variation and 3.04 nits indicates all possible 20 amino acids in equal frequency [103, 104]. As shown in Figure 4.6, 12 sites show increased diversity (residues 3, 6, 7, 11, 12, 13, 14, 18, 27, 31, 34, 35) out of the 16 sites targeted through Darwin Assembly indicated in yellow in Figure 4.6b. The positions depicting the closest to a perfect flat distribution between all 20 residues correspond to positions 11 (Y369), 13 (K371) and 27 (M385). The position with the highest entropy value,

position 11, displayed all 20 amino acids (not shown), but not in perfect ratios, which explains why the entropy value did not reach 3.04 nits.

The library was subjected to two independent selections with different stringency levels, both were prepared and sent for deep sequencing as described in Section 2.4.1. Following data clean up and the multiple sequence alignment of both selections, enrichment scores were calculated by comparing the frequency of each amino acid at each position to the R0 as described in Section 2.4.3 and residues with significant enrichment ratios ($p < 0.05$) for each position were identified. Since the exact distribution of each amino acid for each position was not known but the sample sizes were sufficiently large, the approach to statistically analyse the loop libraries in Section 3.2.6 was also used here but with a couple of alterations. An unpooled version of the two-proportions Z-test that uses the R0 and R1 proportions separately in the estimation of the standard error instead of averaging them was used. The scores were also corrected for multiple testing using the Bonferroni correction. Figure 4.7 depicts all the identified significantly enriched amino acids at specific residue positions of the finger subdomain; the colour scale refers to their respective z-score. Both selections show amino acid conservation at certain positions, the stringent selection in particular shows more enrichment of a wild-type genotype than the less stringent 3hr selection. Both selections, however, show conservation of the wild-type lysine in position 13, even though this position showed significant diversity in the starting R0 library (Figure 4.6). This position corresponds to K371 of phi29 DNAP, a highly conserved amino acid among A and B polymerase families (i.e. K486 in RB69, K464 in KOD-RI and K678 in Polδ) as it is a highly functionally relevant residue involved in the interaction with the γ-phosphate group of the incoming nucleotide during DNA synthesis [105, 18]. This suggests that modifying K371 of phi29 DNAP does not improve recognition and incorporation of sugar-modified nucleosides but remains a potential candidate for the backbone-modified substrates such as phosphorothioate [21] and boranophosphate [22] nucleotides.

The stringent selection (R1c4) enriched the wild-type isoleucine in position 6, corresponding to I364, a structurally relevant residue involved in DNA and dNTP binding [105], as well as I364Q, which has shown to reduce the exonuclease activity of phi29 DNAP [105]. In the longer and less stringent

selection (R1 3hr), only I364R and I364C were enriched significantly. I364R has shown to decrease the template/primer binding stability of phi29 DNAP [105]. A short and stringent selection such as R1c4, should favour enzymes that are fast at incorporating hNTPs, which explains why an exonuclease deficient mutation such as I364Q was enriched during selection for HNA synthesis. Longer incorporations enable less processive enzymes to catch up and can favour template-hopping. Longer reactions also penalize enzyme-catalysed phosphorolysis, exonuclease and low fidelity. The observation that a longer selection did not enrich the exonuclease deficient mutation I364Q or the wild-type, both of which should enable significant synthesis, suggests that the phosphorolysis rate is perhaps the most limiting factor of the reaction and mutations that favour template-hopping or decrease template binding stability such as I364R are favoured.

There are several positions that appear to tolerate a wide variety of residues such as the residue of position 35, corresponding to F393, in both selections and the residue of position 11, corresponding to Y369, in the R1c4 selection. F393 and Y369 have not been previously characterised, nor have shown significant structural or enzymatic roles. However, both selections show enrichment F393V, F393G and F393A in position 35 and Y369E in position 11. The F393G and Y369E show the highest statistically significant enrichment in both selections and thus should be experimentally assessed for HNA synthesis.

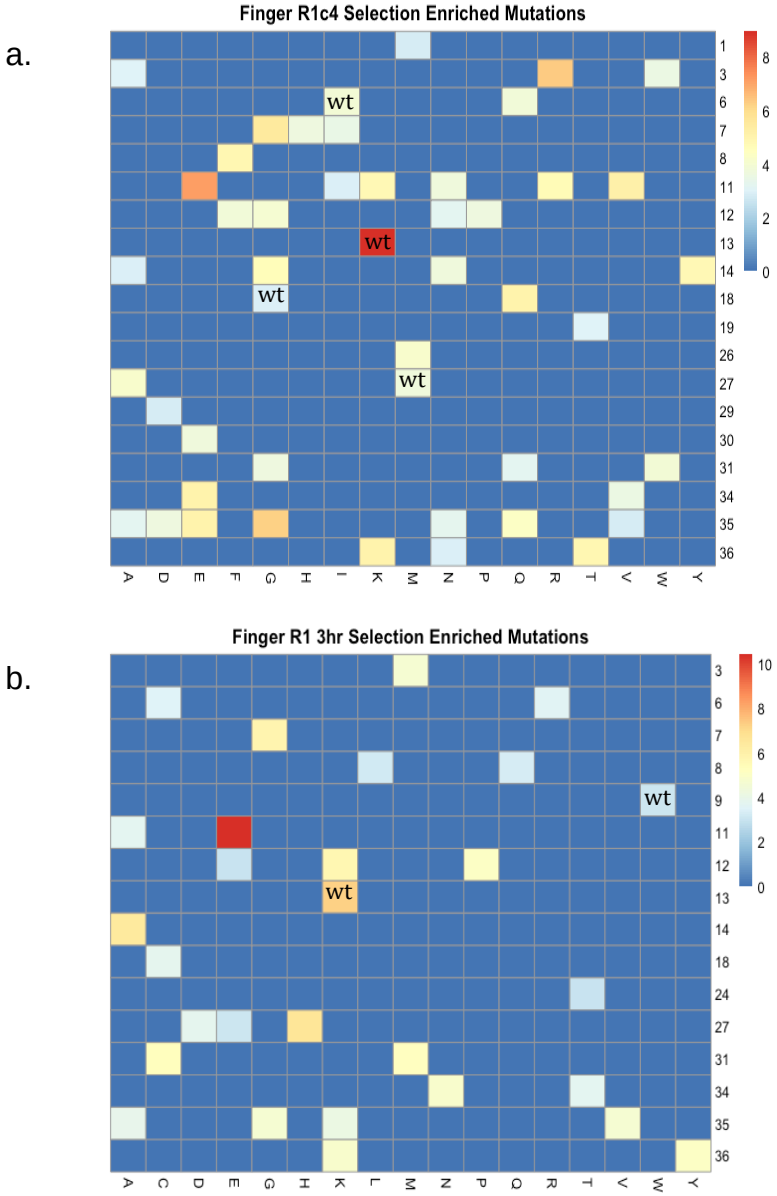


Figure 4.7: Enrichment plots of the multiple-site saturation mutagenesis library post-selection for HNA synthesis. Heat maps showing enrichment of amino acids at each position of the finger multiple-site saturation mutagenesis library after selection for HNA synthesis with standard (a) and stringent (b) selection conditions. The x-axis contains only the amino acids that were enriched for the particular selection. 5% critical value = 2.807 considering 20 multiple tests.

Both selections generally follow a similar pattern of enrichment with only a few and not drastically significant differences. It would be interesting to test the activity of variants observed in the R1c4 with highly significant enrichment scores not observed in the less stringent selection R1 3hr, such as K361R in position 3, to corroborate that the more stringent selection is more efficient at recovering active variants or if it facilitates the recovery of parasites. The less stringent selection also shows enrichment of M385H in position 27, whereas in the stringent selection the wild-type was preferred. Position 27 was one of the most diversified positions in the starting R0 library, thus significant enrichment of M385H at this position specific to the less stringent selection should be further analysed.

Overall, the stringent R1c4 selection resulted in more diversity per position but also in a larger proportion of wild-type genotype, which suggests potential contamination with parasites or that the wild-type is simply more efficient than most mutations when subjected to HNA synthesis or a combination of both. The wild-type could be very fast at incorporation but poor at extension so a fast selection favours enzymes that can put a couple of incorporations fast in that time window. The longer extensions may not put as much burden on how fast the enzyme can incorporate but on enzymes that can avoid side reactions such as phosphorolysis. Further rounds of selection coupled to activity assays should be carried out to see if the population diversity depletes further and individual variants can be isolated with more certainty.

Nonetheless, there are limitations with regards to the statistical analysis used to determine significant enrichment scores. Although The z-test used assumes a normal distribution of the data, however, subjecting the data to normality testing indicated that mutations in both selections and the starting library do not follow a normal distribution. Shapiro-Wilk tests were carried out to assess the distribution of mutations at each position of the finger subdomain pre- and post-selection and resulting p-values were plotted. The null-hypothesis of the Shapiro-Wilk test is that the distribution of amino acids per position is normally distributed. If the p-value is <= 0.05, the null hypothesis is of normality is rejected. As shown in Figure 4.8, none of the positions resulted in p-values >= 0.05, thus a normal distribution in the data cannot be assumed. The positions that showed the highest entropies (see Figure 4.6), are those with the

lowest p-values, which is expected as a high entropy indicates a flat distribution across all amino acids. The p-values at almost all positions decrease further after selection for HNA synthesis; this bias is expected, as some residues should be favoured over the rest. Further rounds of selection that bias the population towards an individual residue would most certainly result in p-values similar to those seen at positions with low diversity, which still do not show a normal distribution. Additional statistical analysis for multimodal distributions such as the nonlinear least-squares method [98] or other non-parametric tests such as the Kruskal–Wallis test [99] by ranks should be implemented to corroborate the previously mentioned observations.

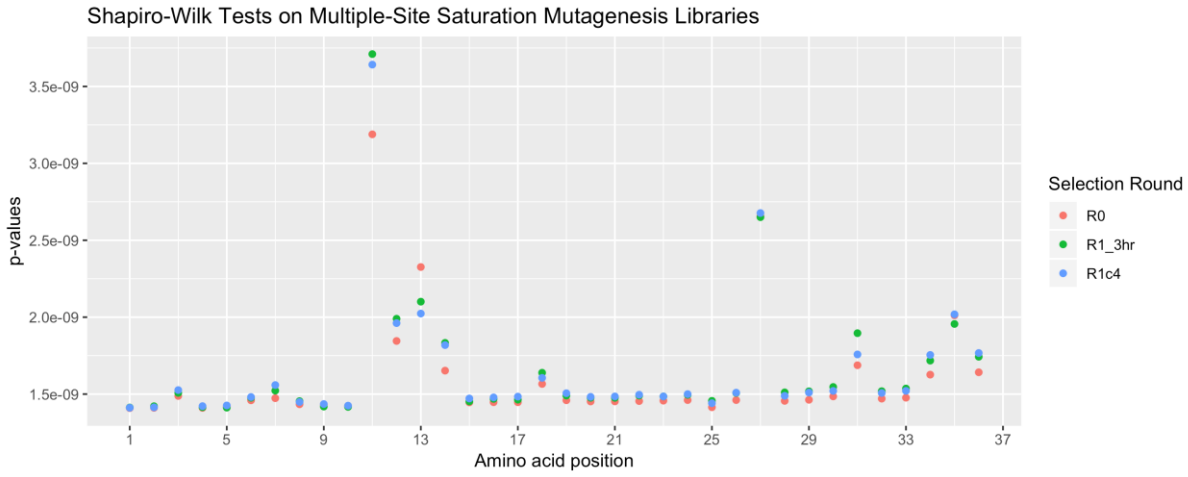


Figure 4.8: Shapiro-Wilk normality test on multiple-site saturation mutagenesis libraries. P-values for each position in the library pre- (R0) and post- (R1_3hr and R1c4) selection for HNA synthesis. Null hypothesis of normality is rejected when the p-value is <= 0.05.

Multiple-site saturation mutagenesis coupled to selection for HNA synthesis allows a more efficient identification of signs of coevolution and epistasis than the individual introduction of mutations, characterisation and subsequent combination. Mutual Information (MI) from information theory has been widely used to infer coevolutionary relationships between positions in protein families [106, 107]. When applied to multiple sequence alignments, MI measures the entropy reduction (or uncertainty reduction) of a position given the knowledge of another position [107, 108]. In other words, MI indicates the extent to which knowing the amino acid at one position can predict the identity

of another position [106]. Thus, MI could be used to reconstruct networks of interaction relevant to HNA synthesis by elucidating coevolving residues in libraries post-selection. As described in Section 2.4.4, the MISTIC server was used to generate MI networks of the finger saturation mutagenesis library selections (Figure 4.9). Each node represents a position in the MSA and connections (edges) represent significant MI >6.5 [109]. Amino acid conservation is also quantified using the Kullback-Leibler (KL) divergence, where red indicates conserved residues and blue less conserved residues. As shown in Figure 4.9, both selections show interactions between residues. The residue at position 9 corresponding to W367 shows significant conservation in both selections, which suggests it could play an important biological or structural role in phi29 DNAP that has not yet been identified. MI therefore can be used to reconstruct the network of interacting residues that are likely to be affected by epistasis.

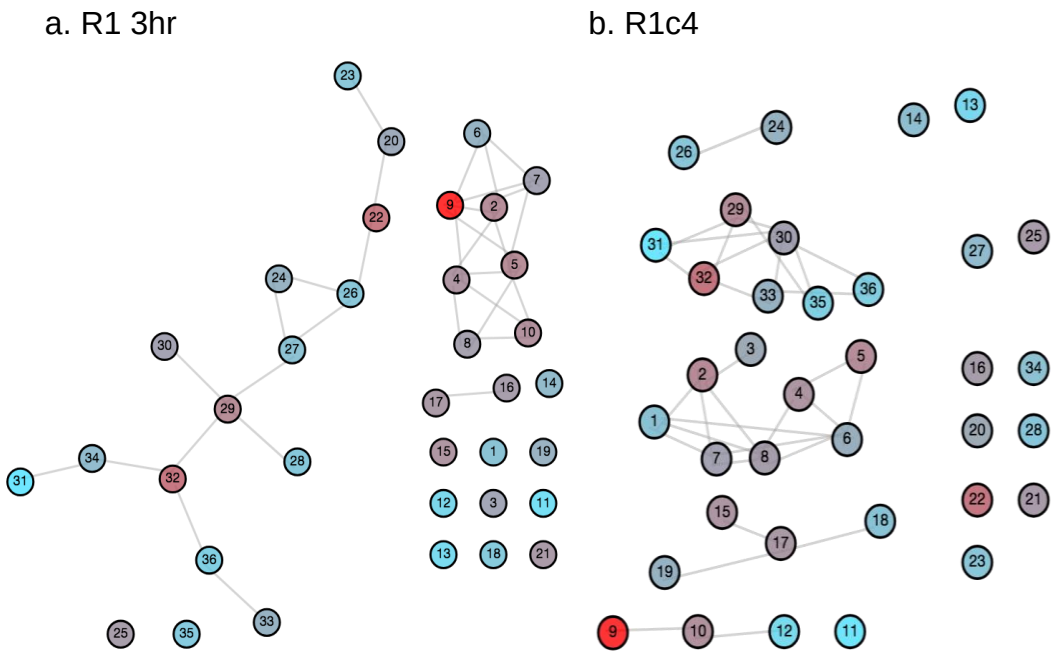


Figure 4.9: Residue coevolution networks in the finger subdomain specific to HNA synthesis. MI networks of (a) 3hr and (b) stringent (R1c4) selections. Nodes represent a position in the MSA and connections (edges) represent mutual information scores. Blue to red colouring represents low to high conservation.

Positions coevolving with multiple positions				Positions coevolving with 1-2 positions			
R1 3hr		R1c4 (stringent)		R1 3hr		R1c4 (stringent)	
Position	Coevolving positions	Position	Coevolving positions	Position	Coevolving positions	Position	Coevolving positions
2	4, 5, 6, 7, 9	1	2, 6, 7, 8	16	17	3	2
4	2, 5, 8, 9, 10	2	1, 3, 7, 8	17	16	5	4, 6
5	2, 4, 7, 8, 9, 10	4	5, 6, 8	20	22, 23	9	10
6	2, 7, 9	6	1, 4, 5, 7, 8	22	20, 26	10	9, 12
7	2, 5, 6, 9	7	1, 2, 6, 8	23	20	12	10
8	4, 5, 10	8	1, 2, 4, 6, 7	24	26, 27	15	17
9	2, 4, 5, 6, 7	17	15, 18, 19	26	24, 27	18	17
10	4, 5, 8	29	30, 31, 32, 35	27	24, 26	19	17
29	27, 28, 30, 32	30	29, 31, 32, 33, 35, 36	28	29	24	26
32	29, 34, 36	31	29, 30, 32	30	29	26	24
		32	29, 30, 31, 33	31	34	35	29, 30
		33	30, 32, 36	33	36		
				34	31, 32		
				36	32, 33		

Table 4.1: Residue coevolution in the finger subdomain specific to HNA synthesis. Coevolution is split between positions with multiple or 1-2 coevolving partners. Shaded in grey are positions showing similar coevolving partners in R1 3hr and R1c4, selections.

Two classes of coevolving positions are typically observed when analysing the MI in multiple sequence alignments, the first includes positions that only coevolve with one or two positions and typically display direct side-chain interactions and the second class includes positions that coevolve with several other positions and are typically located in functionally and structurally relevant regions [107]. Table 4.1 summarises the positions that coevolve with multiple (3+) positions and positions that coevolve with 1 - 2 positions. Residues at positions 24 and 26 corresponding to two non-polar amino acids, A382 and L384 respectively seem to coevolve in both selections. L384 is highly conserved as a non-polar amino acid in DNA-dependent DNA polymerases and has shown to play a significant role in positioning the templating nucleotide and nucleotide insertion fidelity in phi29 DNAP [110]. This position was not mutagenised but it would be unlikely that L384, being important for insertion fidelity, would be enriched during selection for HNA synthesis. It would be interesting to see the effect of disrupting the non-polar interaction between this coevolving pair.

Residues at positions 2, 4, 6, 7, 8, 32 corresponding to residues F360, D362, I364, D365, K366, Y390 respectively, appear to coevolve with multiple residues in both selections. From these residues, F360, D362 and Y390 were not targeted through Darwin Assembly, but seem to coevolve with at least 1 of the targeted residues. As mentioned previously, I364 is a structurally significant residue involved in DNA and dNTP binding [105], which explains why it appears to coevolve with multiple positions. The fact that mutations in this position were enriched and that it appears to interact with numerous other residues, places this position as a top candidate for further characterisation. K366 has shown to be involved in the stabilisation of the incoming nucleotide in terminal protein-primed reactions [111] but not in DNA-primed polymerisation reactions. Since the libraries were subjected to DNA-primed reactions during selection for HNA synthesis, it is possible that K366 has an alternative enzymatic or stabilisation that has not yet been identified. Y390 has also been previously characterised and it appears to have a role in nucleotide binding selection, fidelity and replication [112], which also explains why it appears to coevolve with multiple positions. It would be interesting to see if some of the coevolving pairs have a more significant influence in the structure and function of phi29 DNAP than others.

4.3 Conclusions

Darwin assembly proved to be an efficient approach to generate large multiple-site directed mutagenesis libraries. This approach allowed targeting 12 out of 16 sites in the finger subdomain and introducing significant diversity in at least 3 of these sites. On average, two mutations per gene were observed, as expected from the partial overlap between two of the three inner oligonucleotide groups. Pairwise interactions between residues are thought to significantly contribute to the encoding of protein folds [127] and can thus provide a strong source of functional information on phi29 DNAP and help identify signs of covariation and epistasis. The library was selected for HNA synthesis under different degrees of stringency, and although the activity of these has not been experimentally assessed, their deep sequencing allowed elucidating interesting signs of enrichment and covariation that can be used to improve the engineering of a more efficient HNA synthetase. For instance, the I364Q

substitution appears to be significantly enriched in one of the selections. I364 is a structurally relevant residue involved in DNA and dNTP binding, but the I364Q mutation has shown to decrease the exonuclease activity of phi29 DNAP [105] and could therefore improve the efficiency of HNA synthesis by reducing the rate of phosphorolysis. This mutation is one of the enriched residues that should be considered for further characterisation. Interesting observations were also noted when analysing the data. For instance, K371, a residue involved in interacting with the phosphate of the incoming nucleotide during DNA synthesis was significantly diversified pre-selection, but the wild-type lysine was enriched post-selection for HNA synthesis. This indicates that K371 is essential for phosphate-mediated substrate recognition regardless of the sugar moiety of the incoming nucleotide. Signs of covariation between residues were also identified; from these I364, K366 and Y390 appear to form interactions with multiple residues. The next step is to develop an approach to process all the observed interactions and the effect of single- and double-point mutations on these interactions.

5. Random mutagenesis of the thumb subdomain

5.1 Introduction

Another approach to enhance HNA synthesis of Phi29 DNAP is through the expansion of its substrate specificity by targeting the subdomains involved in substrate recognition and processing. Based on the available crystal structures, the common folding pattern of B-family of DNA polymerases involves a polymerization active site at the C- terminus, a 3'-5' exonuclease active site at the N-terminus and the palm, thumb and finger subdomains that fold into a partially open right hand with a U-shaped groove where the primer-template DNA gets positioned [113]. In particular, the thumb subdomain appears to play a significant role in coordinating the polymerisation and exonuclease activities of Phi29 DNAP [113]. Thus, targeting residues at the thumb subdomain could potentially enhance HNA synthesis by enhancing polymerisation rate and lowering the exonuclease activity of the thumb. The thumb subdomain plays a significant role in the substrate specificity of other replicative DNA polymerases such as Tgo from the archaeon *Thermococcus gorgonarious*. Modifying a substrate specificity checkpoint within its thumb subdomain allows the recognition of rNTPs and the processive synthesis of protein-coding RNAs [114] as well as HNA, CeNA, LNA, TNA, ANA and FANA synthesis and FANA replication [115]. This thumb specificity checkpoint, located at the periphery of the primer-template interaction interface of Tgo, could potentially have a similar role in Phi29 DNAP. However, the structure of the thumb subdomain of Phi29 DNAP differs significantly when compared to other replicative polymerases; it is much smaller and with little helical character [117], making it difficult, if not impossible, to map the same mutations from Tgo. To identify residues in the thumb subdomain of Phi29 DNAP that could enhance XNA synthesis, without making any structural or functional assumptions, random mutagenesis through epPCR was carried out. Additionally, library synthesis through epPCR was already established in the group and I wanted to explore data analysis from epPCR libraries.

5.2 Results and Discussion

5.2.1 Library construction, expression and activity

The random mutagenesis of the thumb subdomain was generated through error-prone PCR (epPCR) amplification as described in Section 2.3.3, initially on a D12A Phi29 DNAP background (D12A). After one round of CST selection the library shows enrichment of HNA synthesis activity (Figure 5.1a), however sequencing variants from the library revealed that a high proportion (90%) of the library was wild-type (D12A variant with an unmodified thumb subdomain). Since the D12A can already posses some HNA synthesis activity, this indicates that the standard 3 hour window is too long to favour enzymes that are faster at incorporating hNTPs than the D12A but too short to allow enzymes that are possibly less processive but have reduced side reactions such as pyrophosphorolysis to catch up. The stringency of the selection was modified by carrying round a first round of selection with a shorter primer extension incubation (30 min) and a longer incubation (12+) hrs to pool non-WT variants.

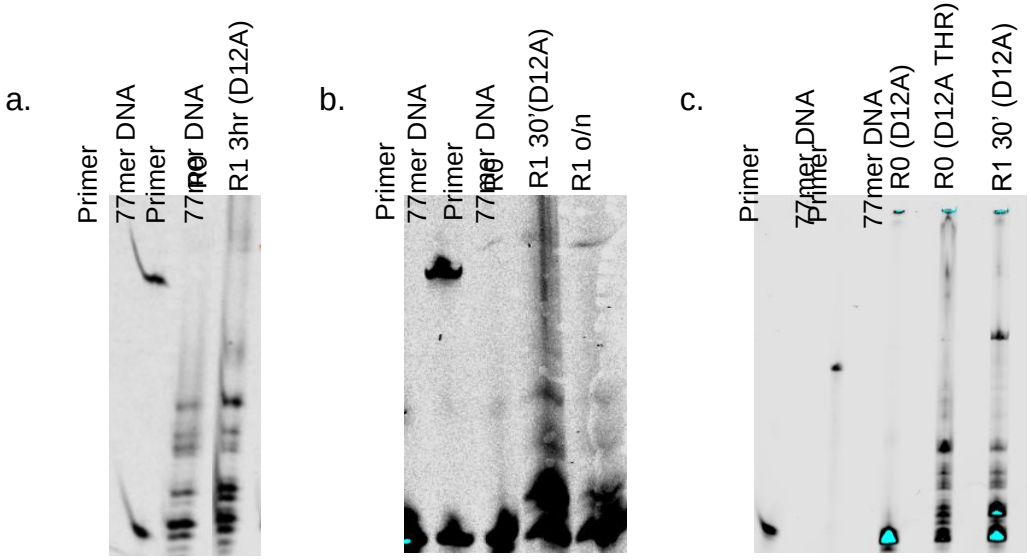


Figure 5.1: HNA synthesis of epPCR thumb libraries on D12A and D12A THR backgrounds. (a) 3 hr HNA primer extension assays of the thumb subdomain epPCR library pre- (R0) and post- a 3 hr selection (R1 3hr) on a D12A background. (b) 3 hr HNA primer extension assays of the thumb subdomain epPCR library pre- (R0) and post- a 30min selection (R1 30') and an overnight selection (R1 o/n) on a D12A background. (c) 3 hr HNA primer extension assays of the thumb subdomain epPCR library pre-selection on a D12A (R0 D12A) and D12A THR (R0 D12A THR) backgrounds and the 30min selection on a D12A background.

Comparing the HNA synthesis activity of purified protein (Figure 5.1b) suggests that the shorter selection appears to have significantly more activity than the overnight selection albeit with a higher degree of background and template-independent synthesis (shown as bands above the full extension mark), suggesting protein saturation. The epPCR Thumb library was also generated on a thermostabilised D12A Phi29 DNAP background (D12A THR), which as shown on Figure 5.1c, appears to have a significantly higher HNA synthesis activity than the epPCR R0 library generated on the D12A background. The 30 min selection on the D12A background library still demonstrates more activity than the R0 generated on the D12A THR, which further demonstrates enrichment, nonetheless this suggests that the D12A THR R0 could potentially show greater activity upon selection.

The 30 min and overnight R1 selections on the D12A background and a mixture of both were subjected to a second round of selection with a primer extension incubation time of 30 min. The R0 library generated on the D12A THR background was subjected also to a 30 min first round of selection (See Table 5.1).

Round 0	Round 1	Round 2	Selection ID
Thumb epPCR on a D12A background	30 min selection	30 min selection	**A2**
	o/n selection	30 min selection	**B4**
	Mix o/n and 30 min selections	30 min selection	**C2**
Thumb epPCR on a THR D12A background	30 min selection	-	**D2**

Table 5.1: Summary of epPCR libraries on D12A and D12A THR backgrounds. R0 indicates epPCR libraries pre-selection; R1 and R2 indicate libraries after 1 and 2 rounds of selection for HNA synthesis on the corresponding R0 and R1 backgrounds respectively.

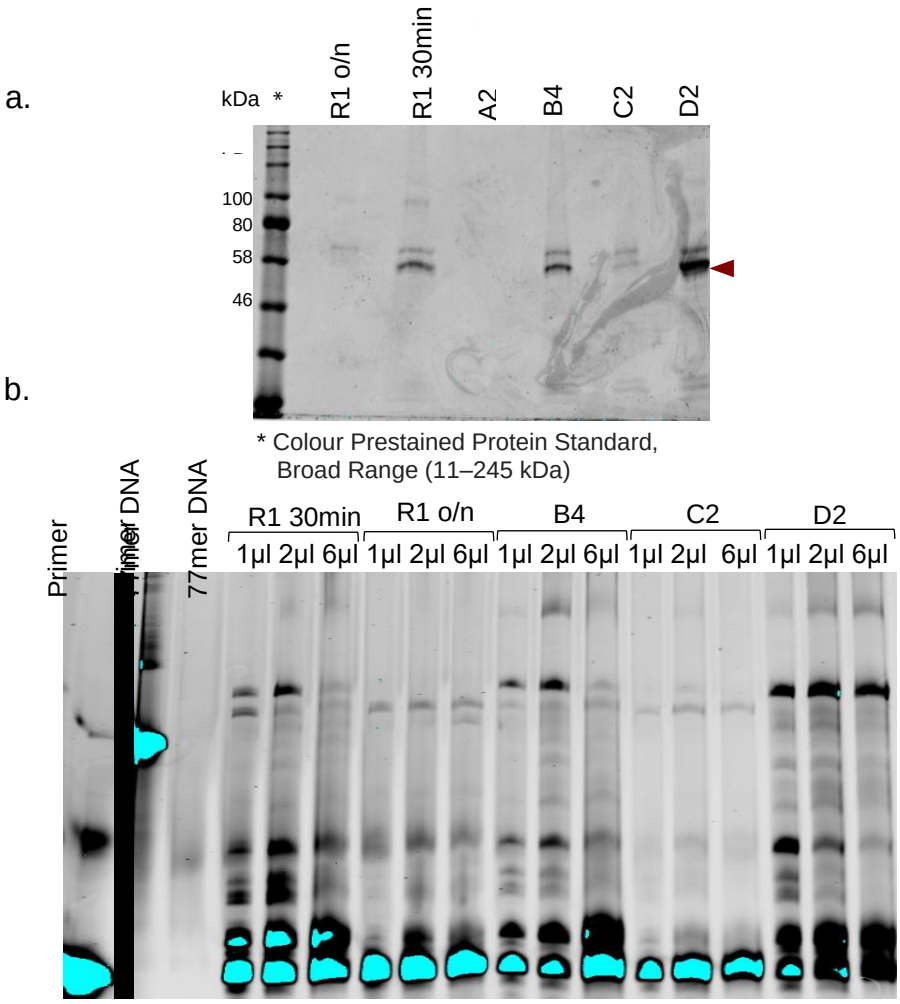


Figure 5.2: HNA synthesis epPCR selections on D12A and D12A THR backgrounds. (a) Protein expression of overnight and 30 min first rounds of selection of the epPCR library on D12A backgrounds (R1 o/n and R1 30 min) as well as protein expression of a 30 min second round of selection on the R1 30 min (A2), on the R1 o/n (B4) and on a combination of R1 30 and R1 o/n (C2), and the protein expression of a 30 min first round of selection of the epPCR library on a D12A THR background. (b) 30 min HNA synthesis from R1 and R2 selections described in (a).

As shown in figure 5.2a, the A2 selection (second round of selection on the 30 min R1) did not express sufficiently to be visualised and was thus not tested for HNA synthesis. The protein yield from the R1 of the library generated on the D12A THR appears to be significantly higher than the rest. Primer extension assays of 30 min were carried out on the selections (see Figure 5.2b). With the shorter extension time, a clear difference between the 30 min and overnight R1 on the epPCR library generated on the D12A background, can be observed. The 30 min selection appears to be more efficient than the overnight selection, nonetheless, B4 (the second round of selection on the overnight R1) does appear to have enriched the population and acquire activity similar to that of the 30 min R1. Still, due to the discrepancies of protein concentration, 1x, 2x and 6x protein volume may not be sufficient to make accurate comparisons. Thus, the experiment should be repeated with standardized protein concentration. Despite this, the D2 selection (first round of selection on the epPCR library generated on the D12A THR background) appears to have significant HNA synthesis activity. Although, it could be due to an artefact of protein concentration, the thermostabilised library seems to be much more stable and express better than the non-thermostabilised library, which could be explained by its ability to withstand inactivating mutations to a greater degree without resulting in unfolded proteins that can lead to cell toxicity. Additionally, despite excess protein of the thermostabilised selection (D2), no significant template degradation was observed, indicating that the mutations could also be reducing the rate of phosphorolysis of phi29 DNAP which could also contribute to enhanced HNA synthesis.

Nonetheless, the sequencing of clones from this first and second rounds of selection revealed that the selected populations were mostly wild-type, which indicates that selection has not been optimised to enrich variants with higher activity. Thus, optimisation of the selection conditions was carried out on the epPCR library generated on the D12A THR background, the seemingly more stable library.

5.2.2 Optimising primer extension assay conditions

Selection conditions tested in the loop libraries and multiple-site saturation mutagenesis libraries (See Sections 3.2.4 and 4.2.3) were used to optimise the selection of the epPCR Thumb library. Selection time was reduced to 10 minutes with the addition of different selection pressures, including: lower hNTP concentration (c1) to push the selection of variants with potentially less substrate specificity able to recognise and incorporate hNTPs more readily, lower selection primer concentration (c2) to push the selection of highly processive variants while reducing the probability of non-specific binding of unextended/poorly-extended primers or lower magnesium concentrations (c3) or a combination of the three (c4) as described in Figure 5.3b. Due to lack of time, the c1, c2 and c3 repertoires were combined and expressed together and compared to the most stringent selection, c4. If any of the c1, c2 or c3 selections enriched the population more than the combined effect of all the selection pressures, a difference in primer extension activity should be observed. 30 min HNA synthesis assays with standardised protein concentrations were carried out on the selections and compared to the starting population of R0. As shown in Figure 5.3d, the R0 and R1c4 appear to have similar activity, which is higher than the combination of c1, c2 and c3 selections. Sequencing colonies from each selection also indicated that the c4 selection contained a higher proportion of non-wild-type variants. Thus it appears that the c4 selection was more effective, although there is no significant enrichment in activity compared to the starting R0 population.

a.

Selection Conditions	Standard	R1c1	R1c2	R1c3	R1c4
Time	3hr	**10min**	**10min**	**10min**	**10min**
hNTPs	2.5mM	**0.5mM**	2.5mM	2.5mM	**0.5mM**
CST_4(7)	10uM	10uM	**2.5uM**	10uM	**2.5uM**
Mg+	10mM	10mM	10mM	**1mM**	**1mM**

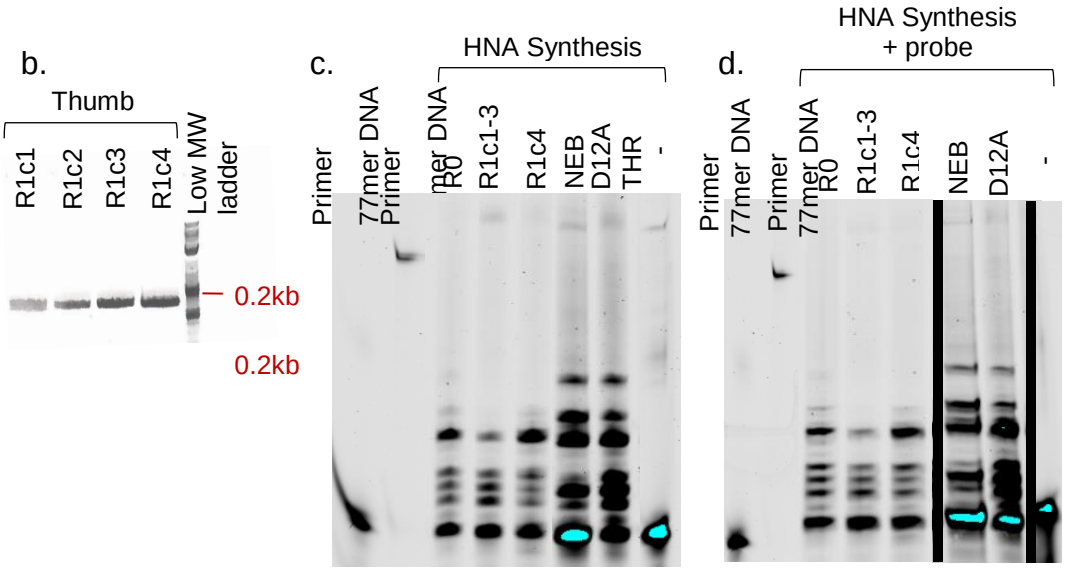


Figure 5.3: Optimising stringency of epPCR library selections and strand displacement assay. (a) Table of standard (black) and stringent selection conditions (red). (b) PCR recovery of thumb libraries (expected 162 bp band size) subjected to stringent selection conditions described in (a). (c) 30 min HNA synthesis of R1c1, R1c2 and R1c3 combined, the most stringent selection (R1c4), phi29 DNAP (NEB) and wild-type D12A THR with standardised protein concentrations. (d) 30 min HNA synthesis and strand displacement assay as described in (c) but with strand displacement probe.

The D12A mutation in the phi29 DNAP background used to generate the library not only abolishes its exonuclease activity but it also affects strand displacement during DNA synthesis [93]. The thumb subdomain plays significant roles in the polymerisation and exonuclease activities of phi29 DNAP, so it could be possible that modifications in the thumb could recover the strand displacement activity of the D12A mutant. To test the strand displacement activity of phi29 DNAP, primer extension assays with an additional probe (TempNblock+20-ExoR) as described in Section 2.5.3 were carried out on the selections. The probe should bind 20bp upstream from the

primer-binding site so that extension past the probe only occurs if phi29 DNAP displaces it from the template. As shown in figure 5.3d, the probe did not appear to interfere with HNA synthesis, as primer extensions look almost identical to those without a probe. Thus, it is possible that the probe was efficiently displaced or that probe binds at a position in the template where HNA synthesis does not reach. Therefore, a probe with a binding site closer to the 3' end of the strand but sufficiently apart from the primer-binding site (to prevent its usage as a primer) should be tested. Alternatively, introducing a hairpin into the template to see if phi29 DNAP can open it as it extends the primer could be tested.

Early rounds of selection may not significantly alter the bulk activity of a population [85], thus although none of the first rounds of selection demonstrated significant activity enrichment from the R0, a second round of selection was carried out with the same R1c4 conditions but with an additional TBT2 20% formamide wash and longer bead capture, which should help reduce background by removing loosely bound plasmids after capture. HNA synthesis assays were then carried out for 3 hr. As shown in Figure 5.4, the R2 activity dropped from the R1c4. 90% of colonies sequenced from the R2 selection were were truncated sequences, potentially arising from non-specific binding during plasmid recovery after selection. Still, even if a small of proportion of enzymes in the R2c4 are functional, the fact that they can collectively almost recapitulate the activity of R0 and R1, indicates that this selection could still contain highly functioning variants. Despite this, since R1c4 displays higher HNA synthesis activity than any other selection, it was chosen for deep mutational scanning to try to identify active variants with minimal non-specific background.

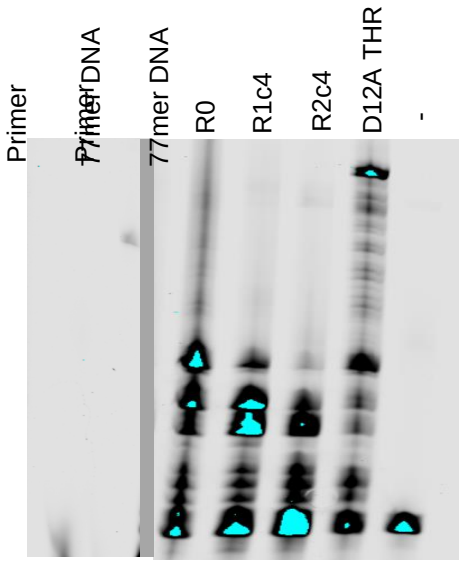


Figure 5.4: Stringent R1 and R2 selections of epPCR thumb library on a D12A THR background. 3 hr HNA primer extension assays of thumb epPCR library before selection (R0), after a stringent selection (See Figure 5.3 for conditions, R1c4), after a second round of selection on the R1c4 selection (R2c4) with the same stringent conditions as R1c4 but with a 3 hr capture and 2 formamide washes and the wild-type D12A THR.

5.2.3 Deep Mutational Scanning

The random mutagenesis library of the thumb subdomain generated through epPCR and its stringent selection for HNA synthesis (R1c4), were both prepared and sent for deep-sequencing as described in Section 2.4.1. The data was cleaned-up and trimmed as described in Section 2.4.2. Since the exact distribution of each amino acid for each position was not known but the sample sizes were sufficiently large, significant enrichment ratios were calculated with the same approach used to analyse the multiple-site saturation mutagenesis data in Section 4.2.3, which involved implementing the unpooled version of the two-proportions Z-test with a correction for multiple testing. As shown in Figure 5.5, eleven positions of the thumb subdomain showing significant enrichment ratios ($p < 0.05$) were identified and plotted in a heatmap with the colour scale referring to their respective z-score. Positions 13 and 42 displayed enrichment of the wild-type residue, corresponding to E540 and D569 respectively. Neither of these two residues have been previously shown to play a significant structural role in phi29 DNAP but could be involved in salt-bridge formation with the neighbouring S551 and K555 of the thumb subdomain as shown in Figure 5.6a.

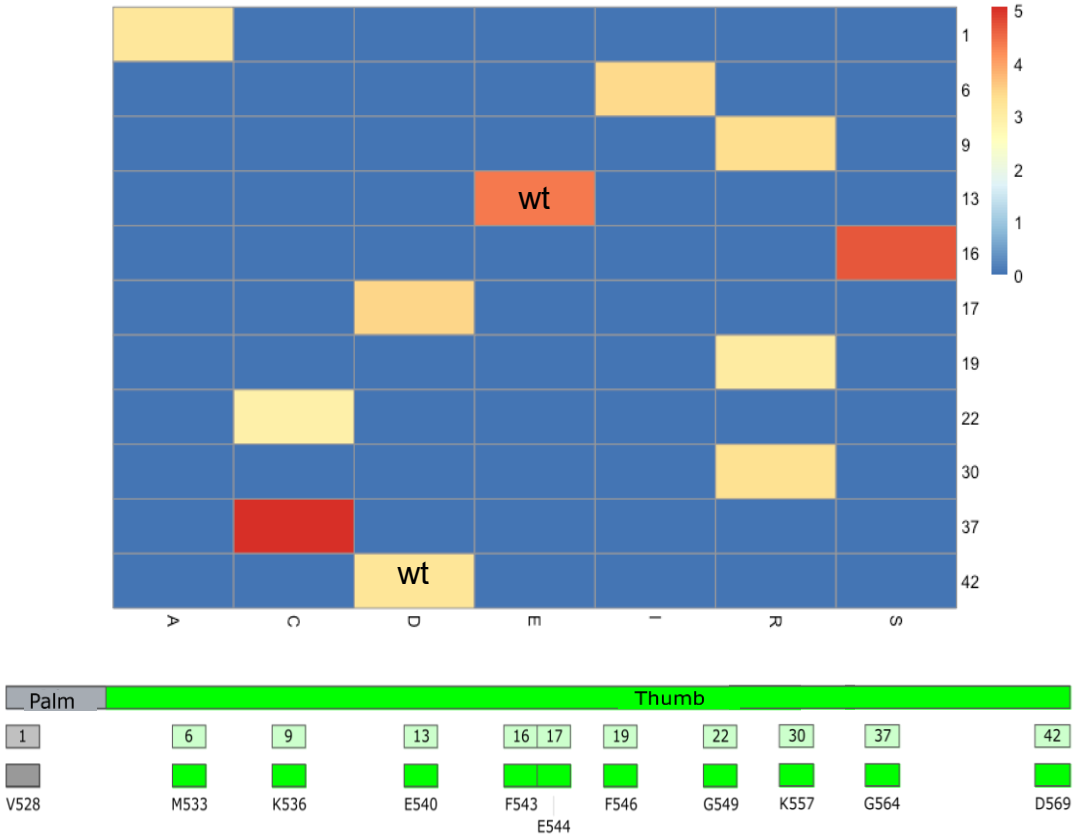


Figure 5.5: Enrichment plots of the epPCR thumb library post-selection for HNA synthesis. Heat maps (top) showing enrichment of amino acids at each position of the thumb after selection for HNA synthesis with stringent conditions (R1c4). 5% critical value = 2.807 considering 20 multiple tests. Thumb map displaying position number with their corresponding amino acid number (bottom).

a.

b.

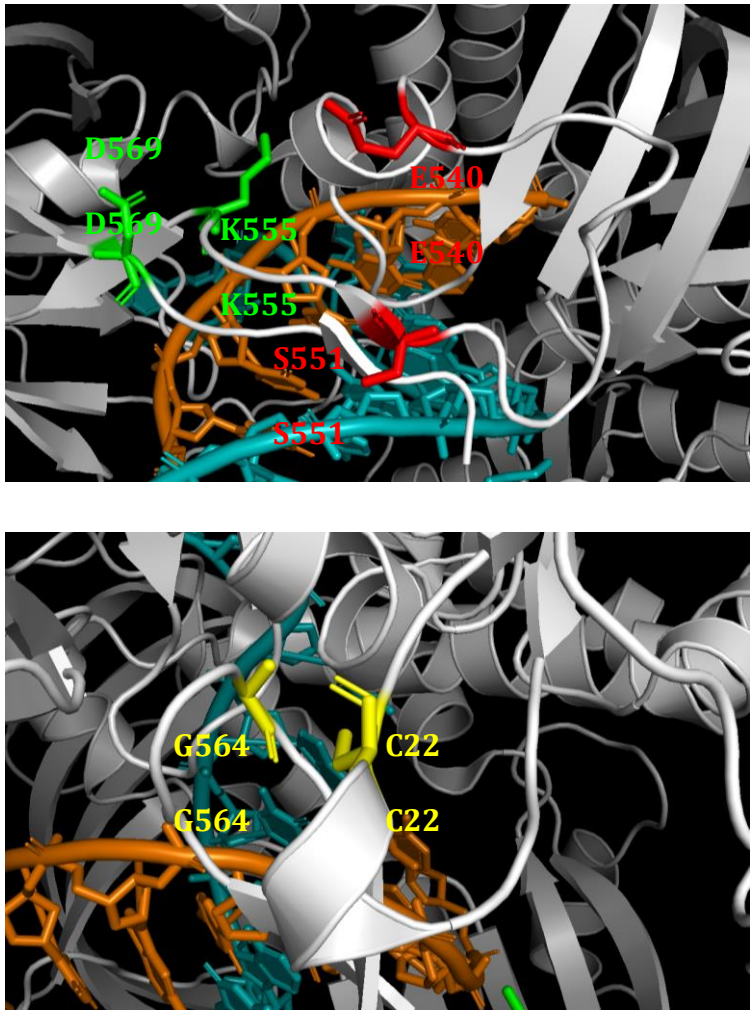


Figure 5:6: Crystal structure of the thumb subdomain of Phi29 DNAP. (a) Phi29 DNAP (PDB ID 2PYJ) showing potential salt-bridge formation between residues D569 and K555 and between residues E540 and S551. (b) Phi29 DNAP (PDB ID 2PYJ) showing a potential disulphide linkage between residues G564 and C22.

Mutations at positions 16 and 37 were the most enriched and correspond to F543S and G564C substitutions respectively. Phenylalanine is a large and highly hydrophobic amino acid and the phenylalanine at position 543 is located near the protein's surface, thus it is possible that its substitution with a small polar amino acid such as serine could be contributing to enhanced solubility and stability. Sub-setting the dataset to variants containing the F543S substitution did not reveal any other significantly enriched mutation that could compensate for the loss of a large hydrophobic residue. Nonetheless, F543S is located in a loop of the thumb subdomain rather than in a hydrophobic core that could otherwise disrupt protein stability. While it is not possible to rule out F543S being a parasite without further characterisation, its strong enrichment suggests it could be contributing to enhanced HNA synthesis. G564C was the most enriched mutation. G564 forms part of the thumb-exonuclease domain interface so the G564C may result in a stabilising disulphide link to the neighbouring C22 in the exonuclease domain (Figure 5.6b) and thus should be experimentally characterised for HNA synthesis.

As shown in Figure 5.7, the Shapiro-Wilk test for normality was carried out on the selections, which indicates that the data is in fact not normally distributed. Thus, as mentioned in Section 4.2.3, the same limitations of the statistical analysis used to determine significant enrichment scores here should be taken into account. Interestingly, there appears to be a clear difference between the p-values of the library generated through epPCR with those of the library generated Darwin Assembly (See Section 4.2, Figure 5.8). In the starting library generated through Darwin Assembly, most p-values fall into 10^{-8}, whereas in the starting library generated through epPCR, most values fall within 10^{-16} and 10^{-17}. It was expected from Darwin Assembly to generate a library with peaks of profound diversity at targeted positions with a close to normal distribution, whereas from epPCR a wider range of residues should be targeted but with less depth and potentially a more skewed distribution such as the one observed here. Further analysis of the data could be carried out to determine the specific differences in distribution of amino acids during mutagenesis and identify the degree of biases introduced with these two mutagenesis methods.

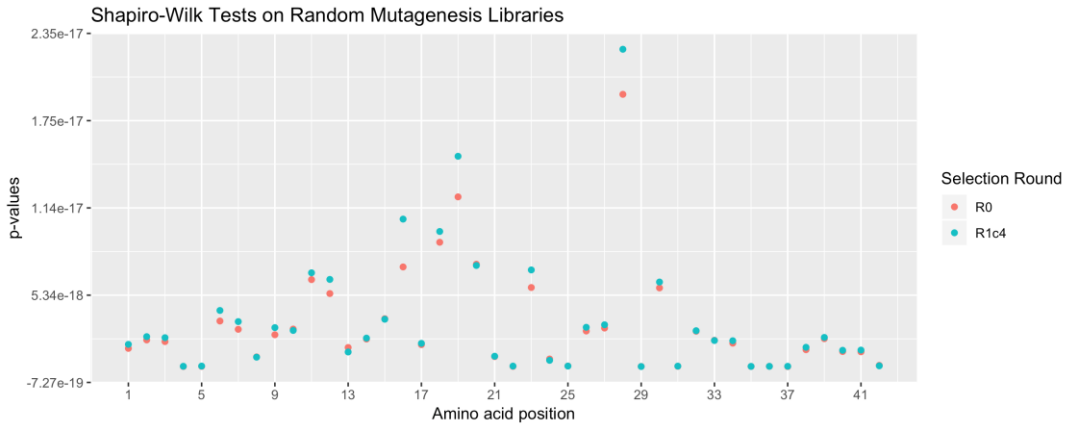


Figure 5.7: Shapiro-Wilk normality test on random mutagenesis libraries. P-values for each position in the library pre- (R0) and post- (R1c4) selection for HNA synthesis. Null hypothesis of normality is rejected when the p-value is <= 0.05.

Random mutagenesis through epPCR can also be used to identify signs of coevolution and epistasis more efficiently than the individual introduction of mutations, characterisation and subsequent combination of mutations. As described in Section 4.2.3, Mutual Information (MI) from information theory has been widely used to infer coevolutionary relationships between positions in protein families [106, 107], by measuring the entropy reduction of a position given the knowledge of another position in a MSA [107, 108]. Thus, MI could be used to reconstruct networks of interaction relevant to HNA synthesis by elucidating coevolving residues in libraries post-selection. As described in Section 2.4.4, the MISTIC server was used to generate MI networks were generated of the random mutagenesis library of the thumb subdomain post-HNA synthesis selection (Figure 5.8). As shown in Figure 5.8, residues at positions 3, 6, 27, 29, 31, 33 and 35 corresponding to C530, M533, M554, P556, P558, Q560 and P562 respectively, show significant conservation. The highly conserved positions detected suggests significant enrichment of a specific residue at the respective position but this would also be observed if these positions were not significantly targeted during epPCR mutagenesis. From the identified conserved positions, only position 6 (M533I) showed enrichment but not very significantly. As mentioned previously, the thumb subdomain of Phi29 DNAP differs significantly when compared to other replicative polymerases; it is much smaller and with little helical character [117],

making it difficult to map these residues onto other polymerases to corroborate their degree of conservation.

There appears to be two clusters of residues coevolving with multiple residues. From these two clusters, the residues at positions 27 and 8, corresponding to residues M554 and D535 respectively have the most connections to other residues. No significant enrichment at either of these two positions was observed post-selection. Nonetheless these two residues should be further characterised for HNA synthesis as positions that coevolve with several other positions are typically located in functionally and structurally relevant regions [107] and could therefore contribute to the enhancement of HNA synthesis.

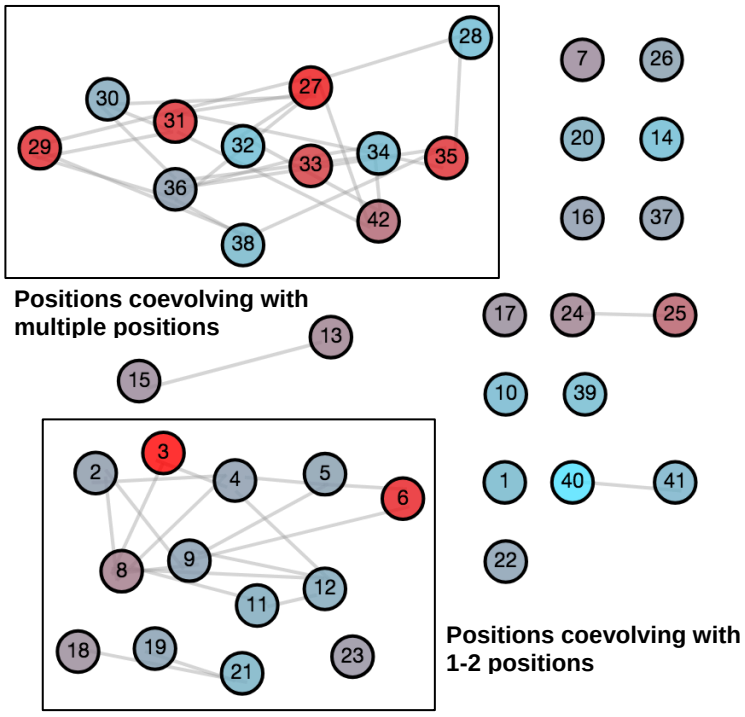


Figure 5.8: Residue coevolution networks in the thumb subdomain specific to HNA synthesis. Nodes represent a position in the MSA and connections (edges) represent mutual information scores >6.5 [109]. Blue to red colouring represents low to high conservation.

5.3 Conclusions

The random mutagenesis of the thumb subdomain through epPCR, proved to be an effective approach for targeting multiple sites simultaneously. Although the diversity at each position was not high, this approach still produced data that can help improve our understanding of the sequence-function relationship of the thumb subdomain of phi29 DNAP with respects to HNA synthesis and overall protein structure and stability. Potential structurally relevant residues that have not been previously characterised, E540 and D569, were enriched; these two residues could be involved in salt-bridge formation with the neighbouring S551 and K555 respectively.

The most enriched mutations post-selection for HNA synthesis, F543S and G564C, should be experimentally characterised. G564C is interesting in particular, as it could be introducing a disulphide link to the neighbouring C22 in the exonuclease domain. A disulphide between polymerase domains, if demonstrated, could contribute to protein stability or significantly affect polymerase dynamics, contributing to enhanced HNA synthesis under the tested selection conditions.

The MI analysis also revealed 2 residues, M554 and D535, seemingly coevolving with multiple residues. Although these 2 residues were not significantly enriched post-selection, these two positions or any of their coevolving partners should be further characterised through site-saturation mutagenesis.

6. Conclusions and Perspectives

The aim of the project was to develop and optimise a platform for the directed evolution and deep mutational scanning of phi29 DNAP to map its sequence-function relationships in the process of evolving a more efficient HNA polymerase. Three different approaches to introduce genetic diversity at different subdomains of phi29 DNAP were implemented and coupled to a functional selection platform for XNA synthetases. A thermostabilised mutant (D12A) was used as a background onto which the libraries were constructed make phi29 DNAP less susceptible to otherwise destabilising mutations and explore a wider range of sequence diversity. Three engineering approaches were carried out aiming to shift phi29 DNAP dynamics and favour synthesis, disfavour proofreading (exonuclease) activity, facilitate template-binding and/or stabilise the nascent duplex during synthesis. Selection parameter optimisation also gave an insight into polymerase dynamics and how it can be implemented to favour different polymerase properties. Overall, short selections seemed to favour enzymes that are fast at incorporating hNTPs. Longer reactions, which should penalize enzyme-catalysed phosphorolysis, exonuclease and low fidelity, resulted in enrichment of enzymes that should be less processive but with enhanced template-hopping and reduced side reactions such as pyrophosphorolysis.

The first approach described was the InDel mutagenesis of three loops belonging to the exonuclease domain and TPR2 and thumb subdomains that come in close proximity to the nascent duplex, with the aim of stabilising it during HNA synthesis and enhance its efficiency. Even after increasing the selection stringency, a single round of selection for HNA synthesis did not alter the overall activity of the libraries sufficiently to be observed during HNA primer extension assays. Nonetheless, deep mutational scanning allowed getting a better understanding of the mutational tolerance and spatial constraints of these loops and their impact in polymerase function. The interaction between the thumb and exonuclease domain couples the polymerisation to the proofreading activities of phi29 DNAP, which is essential for its function [113]. Increasing the length of the exonuclease loop was detrimental to function and is most certainly

causing steric hindrance and/or interfering with the function or its interaction with the thumb. The P562del in the thumb loop, positioned adjacent to the exonuclease loop, was enriched in the stringent or ‘fast’ selection for HNA synthesis, which indicates that this deletion is potentially minimising the thumb-exonuclease interaction resulting in reduced proofreading activity that would favour enzymes in a quick selection for HNA synthesis.

The second mutagenesis approach was a multiple-site saturation mutagenesis strategy targeting residues involved in interdomain contacts with the aim of stabilising the active ternary conformation of phi29 DNAP. Although the activity of this library pre- and post-selection has not yet been experimentally assayed, the deep mutational scanning showed the enrichment of residues and mutations that have been previously characterised and shown significant roles in substrate recognition. The I364Q substitution in particular, has shown to reduce exonuclease activity in phi29 DNAP and it showed significant enrichment during a stringent/quick selection for HNA synthesis. This once again depicts the influence that selection conditions have on the properties of polymerases that are favoured during selection. A long (3 hr) selection resulted in the enrichment of I364R, which has previously shown to decrease template binding stability, depicting the favouring of variants with increased template-hopping and potentially a reduced rate of pyrophosphorolysis in longer selections. Residues coevolving with multiple or a few residues were also identified through mutual information. I364 is one of the identified positions appearing to coevolve with multiple positions, which further confirms that this position or any of its coevolving partners should be further characterised.

The third approach for introducing diversity was random mutagenesis through epPCR of the thumb subdomain. Although the thumb is involved in substrate specificity in numerous replicative polymerases, its structure significantly differs to that of phi29 DNAP, making the mapping of mutations from other polymerases to phi29 DNAP difficult and random mutagenesis a more suitable approach. Although the diversity at each position was not high, this approach still proved to be effective for targeting multiple sites and sampling the sequence space of phi29 DNAP. Even though a single round of selection for HNA synthesis did not alter the overall activity of the library

sufficiently to be observed during HNA primer extension assays, the deep mutational scanning of the library post-selection for HNA synthesis revealed enrichment of potentially structurally relevant residues that have not been previously identified, such as E540 and D569 likely forming salt bridges with the neighbouring S551 and K555 respectively or the G564C mutation which could be introducing a stabilising disulphide link to the neighbouring C22 in the exonuclease domain.

Altogether, the results gave insight into the functional landscape of phi29 DNAP. Further characterisation of enriched and depleted variants identified through deep mutational scanning could be used to expand the shallow phylogenetic depth of phi29 DNAP currently available in open databases. Still, under the conditions tested, in most cases, no significant enrichment or depletion compared to the wild type was observed, which suggests that most variants retained wild type-like activity. This therefore suggests that the functional landscape of phi29 DNAP is flat with a wild type local maxima. It is possible that strong epistatic interactions are preventing the departure from the wild type phenotype. Still, the libraries tested did not sample the complete sequence space of phi29 DNAP; introducing more sequence diversity could give access to a sequence space with a global maxima for HNA synthesis. More complex libraries (i.e. triple mutant libraries), approaches of epistasis and covariation analysis and modelling should be investigated to facilitate reaching the local or global maxima peak for HNA synthesis. Enzyme stability may also be a limiting factor during selection and efficient but unstable HNA synthetases are being not recovered. Thus, polymerase stability should also be tested in enriched and depleted variants as well as variants with wild type-like activity and results should be taken in consideration when optimising selection parameters.

Ultimately, approaches investigated here could be used as a platform to evolve other XNA polymerases and quantify the functional consequence of sequence variation to further improve their engineering.

Appendix A – List of Reagents

Reagent	Manufacturer
1kb DNA ladder	New England Biolabs, USA
2-propanol (isopropanol)	VWR, USA
20% Sodium Dodecyl Sulfate solution (SDS) 2-Butanol	Fisher Scientific, USA
4-(2-hydroxyethyl)-1-piperazineethanesulfonic acid (HEPES)	Sigma-Aldrich, USA
Acrylamide/Bisacrylamide 37.5:1, 40% solution	Alfa Aesar, USA
Absolute Ethanol (Ethanol)	VWR, USA
Agarose	Fisher Scientific, USA
Amicon Ultra-0.5 Centrifugal Filter Unit with Ultracel membrane	Merck Millipore, Germany
Ampicillin, sodium salt (100 mg/mL solution in dH2 O)	AppliChem GmbH, Germany
Bacto Tryptone (Tryptone)	Becton Dickinson, USA
Bacto Yeast Extract (Yeast Extract)	Becton Dickinson, USA
Betaine	Sigma-Aldrich, USA
Bromophenol Blue	Alfa Aesar, USA
Bovine Serum Albumin (BSA)	Sigma-Aldrich, USA
Color Protein Standard, Broad Range	New England Biolabs, USA
Difco Agar (Agar)	Becton Dickinson, USA
Dithiothreitol (DTT)	Thermo Scientific, USA
Dynabeads MyOne Streptavidin C1	Thermo Scientific, USA
Ethylenediaminetetraacetic acid (EDTA), trisodium salt	Sigma-Aldrich, USA
GeneJET Plasmid Miniprep Kit	Thermo Scientific, USA
GeneJET PCR Purification Kit	Thermo Scientific, USA
Glycerol	Fisher Scientific, USA
Glycine	Sigma-Aldrich, USA
Glycogen Azure	Sigma-Aldrich, USA
hNTPs	Oxeltis, France
InstantBlue	Expedeon Ltd, UK
Isopropyl b-D-1-thiogalacto-pyranoside (IPTG)	Glycon Bioch. GmbH, Germany
Lysozyme from chicken egg white	Thermo Scientific, USA
Monarch DNA Gel Extraction Kit	New England Biolabs, USA
Ni-NTA Resin	Thermo Scientific, USA
Orange G	TCS Biosciences, UK
Phenol-Chloroform-Isoamyl Alcohol (25:24:1)	Acros Organics, USA
Polymixin	Sigma-Aldrich, USA
Sodium Chloride (NaCl)	Fisher Scientific, USA
Span 80	Sigma-Aldrich, USA
SYBR Safe DNA Gel Stain	Life Technologies, USA
SYBR Gold DNA Gel Stain	Thermo Scientific, USA
Tetramethylethylenediamine (TEMED)	Sigma-Aldrich, USA
Triton X-100	Sigma-Aldrich, USA
Trypsin	Sigma-Aldrich, USA
Tween 80	Sigma-Aldrich, USA

Appendix B – List of Oligonucleotides used in this work

Oligonucleotides used for constructing InDel libraries

Purpose	Name	Sequence
Exonuclease domain Insertions and deletions	Exo_loop_R	TTCAACTTTGGTGGTGGTTTC
	Exo_loop_INS1	NNSGATTGTCGTGTTTGGGCATATG
	Exo_loop_INS2	NNSNNSGATTGTCGTGTTTGGGCATATG
	Exo_loop_INS3	NNSNNSNNSGATTGTCGTGTTTGGG
	Exo_loop_DEL1	TGTCGTGTTTGGGCATATGG
	Exo_loop_DEL2	CGTGTTTGGGCATATGGCTATATG
	Exo_loop_DEL3	GTTTGGGCATATGGCTATATGAAC
TPR2 subdomain Insertions and deletions	TPR2_loop_R	TTCTTTCAGATAAGGAACTTTACC
	TPR2_loop_INS2	NNSNNSAATGGTGCACTGGGT
	TPR2_loop_INS1	NNSAATGGTGCACTGGG
	TPR2_loop_INS3	NNSNNSNNSAATGGTGCACTGGG
	TPR2_loop_DEL1	GGTGCACTGGGTTTTC
	TPR2_loop_DEL2	GCACTGGGTTTTCGTC
Thumb subdomain Insertions and deletions	Thumb_loop_R	AACCTGAACCGGTTTCGGTTTC
	Thumb_loop_INS2	NNSNNSCCGGGTGGTGTTGTTC
	Thumb_loop_INS1	NNSCCGGGTGGTGTTGTTCTG
	Thumb_loop_INS3	NNSNNSNNSCCGGGTGGTGTTGTTC
	Thumb_loop_DEL1	GGTGGTGTTGTTCTGGTTGATGATAC
	Thumb_loop_DEL2	GGTGTTGTTCTGGTTGATGATACCTTTAC
	Thumb_loop_DEL3	GTTGTTCTGGTTGATGATACCTTTACGATC
	Thumb_loop_DEL4	GTTCTGGTTGATGATACCTTTACGATCAAA

Oligonucleotides used for Darwin Assembly

Category	Name	Sequence
Nicking site introduction	iPCR_Nt.BspQI_F	gctcttcaACGTTCGCTCGCGTATC
	iPCR_Nt.BspQI_R	GAAGCGACTGCTGCTGC
Biotinylated	P2_DA_F1_SapI	/5BiotinTEG/ACAGACGGCATGATGAACCTGAaaagctcttcaGGGTCC TCAACGACAGGAGCAC
	P2_DA_R1_SapI	CAAAGCCCGAAAGGAAGCTGAggaagagctttTTGAGATCGTTTTGGT CTGCGC/3InvdT/
	P2_DA_F2_SapI	/5BiotinTEG/GATGAGGGTGTCAGTGAAGaaagctcttcaGGGTCCTCA ACGACAGGAGCAC
	P2_DA_R2_SapI	CAAAGCCCGAAAGGAAGCTGAggaagagctttGATATAGGCGCCAGC AACC/3InvdT/
Theta	Theta2(F1R2)	CAAAGCCCGAAAGGAAGCTGAGGAAGAGCTTTGATATAGGCGCC AGCAACCTTTTTACAGACGGCATGATGAACCTGAAAAGCTCTTCA GGGTCCTCAACGACAGGAGCAC
	pET23P2_NotI_F	CATTAAGCGCGGCCGCTGTGG
Outnest	P2_DA_F1_SapI_Out	ACAGACGGCATGATGAACCTGA
	P2_DA_R1_SapI_Out	GCGCAGACCAAAACGATCTCAA
	P2_DA_F2_SapI_Out	GATGAGGGTGTCAGTGAAG

	P2_DA_R2_SapI_Out	GGTTGCTGGCGCCTATATC
Backbone amplification	DA_iPCR_pET23_SapI_FWD	AAAGCTCTTCCTGAGTTGGCTGCTGCCAC
	DA_iPCR_pET23_SapI_RV2	aaaGCTCTTCcCCCGGCTAGGCTGGC
Inner Mutagenic	p2_f1_1NDT	CCGGTCTGTTCNDTGACTTCATTGATAAATGGACCTATATCAAAACC
	p2_f1_1VMA	CCGGTCTGTTCVMAGACTTCATTGATAAATGGACCTATATCAAAACC
	p2_f1_1ATG	CCGGTCTGTTCATGGACTTCATTGATAAATGGACCTATATCAAAACC
	p2_f1_1TGG	CCGGTCTGTTCTGGGACTTCATTGATAAATGGACCTATATCAAAACC
	p2_f1_2NDT	CCGGTCTGTTCAAAGACTTCNDTGATAAATGGACCTATATCAAAACC
	p2_f1_2VMA	CCGGTCTGTTCAAAGACTTCVMAGATAAATGGACCTATATCAAAACC
	p2_f1_2ATG	CCGGTCTGTTCAAAGACTTCATGGATAAATGGACCTATATCAAAACC
	p2_f1_2TGG	CCGGTCTGTTCAAAGACTTCTGGGATAAATGGACCTATATCAAAACC
	p2_f1_3NDT	CCGGTCTGTTCAAAGACTTCATTNDTAAATGGACCTATATCAAAACC
	p2_f1_3VMA	CCGGTCTGTTCAAAGACTTCATTVMAAAATGGACCTATATCAAAACC
	p2_f1_3ATG	CCGGTCTGTTCAAAGACTTCATTATGAAATGGACCTATATCAAAACC
	p2_f1_3TGG	CCGGTCTGTTCAAAGACTTCATTTGGAAATGGACCTATATCAAAACC
	p2_f1_4NDT	CCGGTCTGTTCAAAGACTTCATTGATNDTTGGACCTATATCAAAACC
	p2_f1_4VMA	CCGGTCTGTTCAAAGACTTCATTGATVMATGGACCTATATCAAAACC
	p2_f1_4ATG	CCGGTCTGTTCAAAGACTTCATTGATATGTGGACCTATATCAAAACC
	p2_f1_4TGG	CCGGTCTGTTCAAAGACTTCATTGATTGGTGGACCTATATCAAAACC
	p2_f1_5NDT	CCGGTCTGTTCAAAGACTTCATTGATAAANDTACCTATATCAAAACC
	p2_f1_5VMA	CCGGTCTGTTCAAAGACTTCATTGATAAAVMAACCTATATCAAAACC
	p2_f1_5ATG	CCGGTCTGTTCAAAGACTTCATTGATAAAATGACCTATATCAAAACC
	p2_f1_5TGG	CCGGTCTGTTCAAAGACTTCATTGATAAATGGACCTATATCAAAACC
	p2_f2_1NDT	GATAAATGGACCNDTATCAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_1VMA	GATAAATGGACCVMAATCAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_1ATG	GATAAATGGACCATGATCAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_1TGG	GATAAATGGACCTGGATCAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_2NDT	GATAAATGGACCTATNDTAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_2VMA	GATAAATGGACCTATVMAAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_2ATG	GATAAATGGACCTATTGGAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_2TGG	GATAAATGGACCTATATCAAAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_3NDT	GATAAATGGACCTATATCNDTACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_3VMA	GATAAATGGACCTATATCVMAACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_3ATG	GATAAATGGACCTATATCATGACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_3TGG	GATAAATGGACCTATATCTGGACCACCTCCGAAGGTGCAATTAAACAG
	p2_f2_4NDT	GATAAATGGACCTATATCAAANDTACCTCCGAAGGTGCAATTAAACAG
	p2_f2_4VMA	GATAAATGGACCTATATCAAAVMAACCTCCGAAGGTGCAATTAAACAG
	p2_f2_4ATG	GATAAATGGACCTATATCAAAATGACCTCCGAAGGTGCAATTAAA

	CAG
p2_f2_4TGG	GATAAATGGACCTATATCAAATGGACCTCCGAAGGTGCAATTAAA CAG
p2_f2_5NDT	GATAAATGGACCTATATCAAAACCACCTCCGAANDTGCAATTAAA CAG
p2_f2_5VMA	GATAAATGGACCTATATCAAAACCACCTCCGAAVMAGCAATTAAA CAG
p2_f2_5ATG	GATAAATGGACCTATATCAAAACCACCTCCGAAATGGCAATTAAA CAG
p2_f2_5TGG	GATAAATGGACCTATATCAAAACCACCTCCGAATGGGCAATTAAA CAG
p2_f3_1NDT	CTGGCAAAACTGNDTCTGAATTCCCTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_1VMA	CTGGCAAAACTGVMACTGAATTCCCTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_1ATG	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_1TGG	CTGGCAAAACTGTGGCTGAATTCCCTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_2NDT	CTGGCAAAACTGATGCTGAATNDTCTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_2VMA	CTGGCAAAACTGATGCTGAATVMACTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_2ATG	CTGGCAAAACTGATGCTGAATATGCTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_2TGG	CTGGCAAAACTGATGCTGAATTGGCTGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_3NDT	CTGGCAAAACTGATGCTGAATTCCNDTTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_3VMA	CTGGCAAAACTGATGCTGAATTCCVMATATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_3ATG	CTGGCAAAACTGATGCTGAATTCCATGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_3TGG	CTGGCAAAACTGATGCTGAATTCCTGGTATGGTAAATTTGCAAGC AATCCGGATG
p2_f3_4NDT	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTNDTTTTGCAAGC AATCCGGATG
p2_f3_4VMA	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTVMATTTGCAAGC AATCCGGATG
p2_f3_4ATG	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTATGTTTGCAAGC AATCCGGATG
p2_f3_4TGG	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTTGGTTTGCAAGC AATCCGGATG
p2_f3_5NDT	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAANDTGCAAGC AATCCGGATG
p2_f3_5VMA	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAAVMAGCAAG CAATCCGGATG
p2_f3_5ATG	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAAATGGCAAGC AATCCGGATG
p2_f3_5TGG	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAATGGGCAAG CAATCCGGATG
p2_f3_6NDT	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAATTTNDTAGC AATCCGGATG
p2_f3_6VMA	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAATTTVMAAGC AATCCGGATG
p2_f3_6ATG	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAATTTATGAGC AATCCGGATG
p2_f3_6TGG	CTGGCAAAACTGATGCTGAATTCCCTGTATGGTAAATTTTGGAGC AATCCGGATG

Oligonucleotides used for epPCR

Category	Name	Sequence
Thumb randomisation	P2_HisTAG_SapI_Rv	aaaGCTCTTCcGTGTTTGATCGTAAAGGTATCATC
	Lib7_Fw	AAAGCTCTTCATACACCGATATCAAATTTAGC
	Lib7_Fw_ThSTP2	AAAGCTCTTCATACACCGATATCAAAcTTAGC
Backbone amplification	iPCR pET23_Lib7_Fw	AAAGCTCTTCACACCACCACCACCACCACTGAGATC
	iPCR_Lib7_rv	AAAGCTCTTCAGTAATCATCCGGACTACCTTCAACC
	iPCR_Lib7Rv_ThSTP2	AAAGCTCTTCAGTAATCATCCGGACTACCTgCAACC

Oligonucleotides used for cloning selections and activity assays

Category	Name	Sequence
Cloning selections	p2_SapI_Exo_Fwd	aaaGCTCTTCcCACATGcctcgcaaaaggtatag
	P2_SapI_Thumb_Rv	aaaGCTCTTCcGTGTTTGATCGTAAAGGTATCATC
	p2_SapI_Exo_Rv	aaaGCTCTTCcGTGtttcatgtatgactatctcctgtgtgG
	p2_SapI_ThumbHistag_Fwd	AAAGCTCTTCACACCACCACCACCACCACTG
Selection oligonucleotide	CST_04(7)exoR	5biotinTEG/ACC*G*C*A
Strand displacement	PH_TempNblock+20-ExoR	TGCTGTTCGGTAATCG/3InvdT/
HNA synthesis	Tag01F3-exoR	/5IRD700/CGGATCCGTTTAAGC*T*A*G*G
	TempN	TGGTCCAGCATCGTGAGATCGATTACCGAACAGCACTAC GTGGCTAAGTGCTTATCTCCTAGCTTAAACGGATCCG

Oligonucleotides used for amplifying libraries for deep sequencing

Category	Name	Sequence
Amplifying exonuclease	Seq_Exo_F1	GAGATCTCGATCCCGCGAAATT
	Seq_Exo_R3	CCATTGCGTTCCAGCCAGTTAA
Amplifying TPR2	Seq_TPR2_F1	TGAAATTCAAAGCAACCACCGGT
	Seq_TPR2_R2	CGGAATTTCGGTGCCGGTC
Amplifying thumb	Seq_Thumb_F1	CATCTGACCGGCACCGAAATTC
	Seq_Thumb_R1	CAGCCAACTCAGCTTCCTTTCG
Amplifying finger library	Seq_Finger_F1	CTGAAAAGCAGCGGTGGTGAAA
	Seq_Finger_R2	ACGTGCCCATGCGGTAATAAAC

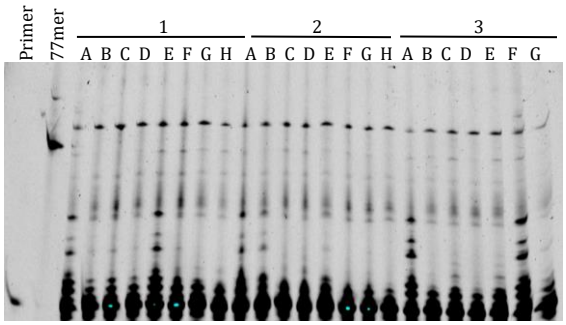
Appendix C – Scripts written for data analysis

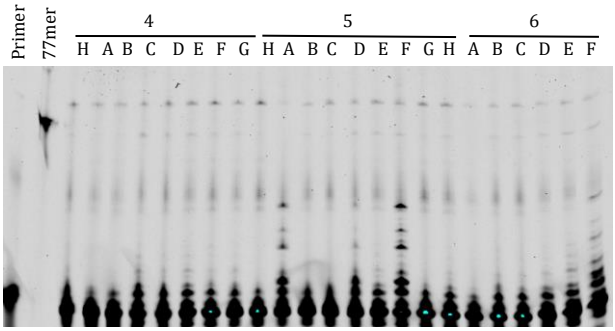
```
[Inp_head, Inp_seq] = fastaread ('MSAR0.fas');  % Input MSA R0
Input_MSA = char(Inp_seq);
[Out_head, Out_seq] = fastaread ('MSAR1.fas');  % Input MSA R1
Output_MSA = char(Out_seq);
WT = char('VKCAGMTDKIKKEVTFENFKVGFSRKMKPKPVQVPGGVVLVD');  % Input WT sequence
% Alignment into integer matrix
Input_MSA2 = aa2int(Input_MSA);
Output_MSA2 = aa2int(Output_MSA);
WT2 = aa2int(WT);
% aa2int converts all unknown symbols into 0s, the loops convert 0s into 25 fo
r gaps and 24 for stop codons
for a = 1 : size(Input_MSA, 2)
    for b = 1 : size(Input_MSA, 1)
        if Input_MSA2(b,a) == 0
            Input_MSA2(b,a) = 25;
        end
    end
end

for a = 1 : size(Output_MSA, 2)
    for b = 1 : size(Output_MSA, 1)
        if Output_MSA2(b,a) == 0
            Output_MSA2(b,a) = 25;
        end
    end
end

for a = size(Input_MSA,1) : -1 : 1
    for b = 1 : size(Input_MSA,2)
        if Input_MSA2(a,b) == 24
            Input_MSA2(a,:) = [];
        end
    end
end

for a = size(Output_MSA,1) : -1 : 1
    for b = 1 : size(Output_MSA,2)
        if Output_MSA2(a,b) == 24
            Output_MSA2(a,:) = [];
        end
    end
end
```

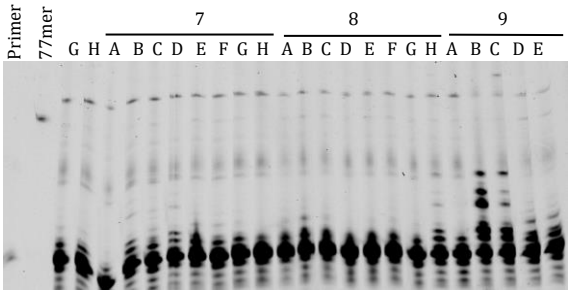
```
42.
43. % Alignments into a frequency tables
44. Input_total = size(Input_MSA2,1);
45. Output_total = size(Output_MSA2, 1);
46. Input_count = zeros(20,size(Input_MSA2, 2));
47. Output_count = zeros(20,size(Output_MSA2, 2));
48.
49. for a = 1 : size(Input_MSA2, 1)
50.     for b = 1 : size(Input_MSA2, 2)
51.         Input_count(Input_MSA2(a,b),b) = Input_count(Input_MSA2(a,b),b) + 1;
52.     end
53. end
54.
55. Input_freq = Input_count/Input_total;
56.
57. for a = 1 : size(Output_MSA2, 1)
58.     for b = 1 : size(Output_MSA2, 2)
59.         Output_count(Output_MSA2(a,b),b) = Output_count(Output_MSA2(a,b),b) +
    1;
60.     end
61. end
62. Output_freq = Output_count/Output_total;
63.
64. %% Comparing the two ratios
65. for a = 1 : size(Output_freq, 1)        % number of residues
66.     for b = 1 :size(Output_freq, 2)     % number of amino acids
67.         Pi = Input_freq(a,b);
68.         Qi = 1 - Input_freq(a,b);
69.         Po = Output_freq(a,b);
70.         Qo = 1 - Output_freq(a,b);
71.         Ni = Input_total;
72.         No = Output_total;
73.         Zbin(a,b) = (Po - Pi) / sqrt((Pi*Qi)/Ni + (Po*Qo)/No);
74.     end
75. end
```

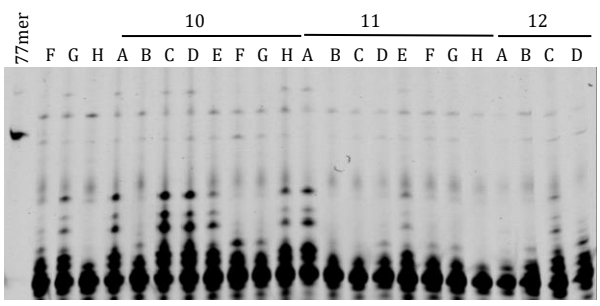
Appendix D – Small-scale screening

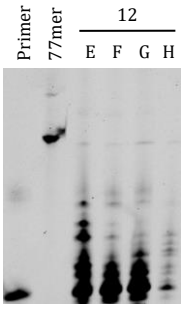
3 hr HNA small-scale activity assays of lysates from 96 colonies of the InDel mutagenesis TPR2 R1c4 selection











Bibliography

1. Benner S. Understanding Nucleic Acids Using Synthetic Chemistry. Accounts of Chemical Research. 2004;37(10):784-797.
2. Herdewijn P, Marlière P. Toward Safe Genetically Modified Organisms through the Chemical Diversification of Nucleic Acids. Chemistry & Biodiversity. 2009;6(6):791-808.
3. Pinheiro V, Holliger P. The XNA world: progress towards replication and evolution of synthetic genetic polymers. Current Opinion in Chemical Biology. 2012;16(3-4):245-252.
4. Pinheiro V, Loakes D, Holliger P. Synthetic polymers and their potential as genetic materials. BioEssays. 2012;35(2):113-122.
5. Ma Q, Lee D, Tan Y, Wong G, Gao Z. Synthetic genetic polymers: advances and applications. Polymer Chemistry. 2016;7(33):5199-5216.
6. Wojciechowski F, Leumann C. Alternative DNA base-pairs: from efforts to expand the genetic code to potential material applications. Chemical Society Reviews. 2011;40(12):5669.
7. Kjellberg J, Johansson N. Characterization of N7 and N9 alkylated purine analogues by 1H and 13C nmr. Tetrahedron. 1986;42(23):6541-6544.
8. Duschinsky R, Pleven E, Heidelberger C. THE SYNTHESIS OF 5-FLUOROPYRIMIDINES. Journal of the American Chemical Society. 1957;79(16):4559-4560.
9. Lu H, He K, Kool E. yDNA: A New Geometry for Size-Expanded Base Pairs. Angewandte Chemie. 2004;116(43):5958-5960.
10. Wagner E, Oberhauser B, Holzner A, Brunar H, Issakides G, Schaffner G et al. A simple procedure for the preparation of protected 2′-O-methyl or 2′-O-ethyl ribonucleoside-3′-O-phosphoramidites. Nucleic Acids Research. 1991;19(21):5965-5971.
11. Wilds C, Damha M. 2'-Deoxy-2'-fluoro-beta-D-arabinonucleosides and oligonucleotides (2'F-ANA): synthesis and physicochemical studies. Nucleic Acids Research. 2000;28(18):3625-3635.
12. Burkart M, Vincent S, Düffels A, Murray B, Ley S, Wong C. Chemo-enzymatic synthesis of fluorinated sugar nucleotide: useful mechanistic Probes for glycosyltransferases. Bioorganic & Medicinal Chemistry. 2000;8(8):1937-1946.
13. Koshkin A, Singh S, Nielsen P, Rajwanshi V, Kumar R, Meldgaard M et al. LNA (Locked Nucleic Acids): Synthesis of the adenine, cytosine, guanine, 5-methylcytosine, thymine and uracil bicyclonucleoside monomers, oligomerisation, and unprecedented nucleic acid recognition. Tetrahedron. 1998;54(14):3607-3630.
14. Obika S, Nanbu D, Hari Y, Andoh J, Morio K, Doi T et al. ChemInform Abstract: Stability and Structural Features of the Duplexes Containing Nucleoside Analogues with a Fixed N-Type Conformation, 2′-O,4′-C-Methyleneribonucleosides. ChemInform. 2010;29(42):no-no.
15. De Bouvere B, Kerreinans L, Hendrix C, De Winter H, Schepers G, Van Aerschot A et al. Hexitol Nucleic Acids (HNA): Synthesis and Properties. Nucleosides and Nucleotides. 1997;16(7-9):973-976.
16. Nauwelaerts K, Lescrinier E, Sclep G, Herdewijn P. Cyclohexenyl nucleic acids: conformationally flexible oligonucleotides. Nucleic Acids Research. 2005;33(8):2452-2463.

17. Schoning K, Scholz P, Guntha S, Wu X, Krishnamurthy R, Eschenmoser E. Chemical Etiology of Nucleic Acid Structure: The alpha -Threofuranosyl-(3' -> 2') Oligonucleotide System. Science. 2000;290(5495):1347-1351.
18. Chim N, Shi C, Sau S, Nikoomanzar A, Chaput J. Structural basis for TNA synthesis by an engineered TNA polymerase. Nature Communications. 2017;8(1).
19. Zhang L, Peritz A, Meggers E. A Simple Glycol Nucleic Acid. Journal of the American Chemical Society. 2005;127(12):4174-4175.
20. Joyce G, Schwartz A, Miller S, Orgel L. The case for an ancestral genetic system involving simple analogues of the nucleotides. Proceedings of the National Academy of Sciences. 1987;84(13):4398-4402.
21. Guga P, Koziołkiewicz M. Phosphorothioate Nucleotides and Oligonucleotides - Recent Progress in Synthesis and Application. Chemistry & Biodiversity. 2011;8(9):1642-1681.
22. Li P, Sergueeva Z, Dobrikov M, Shaw B. Nucleoside and Oligonucleoside Boranophosphates: Chemistry and Properties. Chemical Reviews. 2007;107(11):4746-4796.
23. Nielsen P, Egholm M. An Introduction to Peptide Nucleic Acid. Current Issues in Molecular Biology. 1999;.
24. Wilson S. Application of nucleic acid-based technologies to the diagnosis and detection of disease. Transactions of the Royal Society of Tropical Medicine and Hygiene. 1993;87(6):609-611.
25. Tenover F. Diagnostic deoxyribonucleic acid probes for infectious diseases. Clinical Microbiology Reviews. 1988;1(1):82-101.
26. Cheng J, Zhang Y, Li Q. Real-time PCR genotyping using displacing probes. Nucleic Acids Research. 2004;32(7):e61-e61.
27. Wang L, Yang C, Medley C, Benner S, Tan W. Locked Nucleic Acid Molecular Beacons. Journal of the American Chemical Society. 2005;127(45):15664-15665.
28. Tyagi S, Kramer F. Molecular Beacons in Diagnostics. F1000 Medicine Reports. 2012;4.
29. Kratschmer C, Levy M. Effect of Chemical Modifications on Aptamer Stability in Serum. Nucleic Acid Therapeutics. 2017;27(6):335-344.
30. Ni S, Yao H, Wang L, Lu J, Jiang F, Lu A et al. Chemical Modifications of Nucleic Acid Aptamers for Therapeutic Purposes. International Journal of Molecular Sciences. 2017;18(8):1683.
31. Ellington A, Szostak J. In vitro selection of RNA molecules that bind specific ligands. Nature. 1990;346(6287):818-822.
32. Tuerk C, Gold L. Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. Science. 1990;249(4968):505-510.
33. Ni X, Castanares M, Mukherjee A, Lupold S. Nucleic Acid Aptamers: Clinical Applications and Promising New Horizons. Current Medicinal Chemistry. 2011;18(27):4206-4214.
34. Tolle F, Mayer G. Dressed for success – applying chemistry to modulate aptamer functionality. Chem Sci. 2013;4(1):60-67.
35. Keefe A, Cload S. SELEX with modified nucleotides. Current Opinion in Chemical Biology. 2008;12(4):448-456.
36. Taylor A, Holliger P. Selecting Fully-Modified XNA Aptamers Using Synthetic Genetics. Current Protocols in Chemical Biology. 2018;10(2):e44.
37. Wilds C, Wawrzak Z, Krishnamurthy R, Eschenmoser A, Egli M. Crystal Structure of a B-Form DNA Duplex Containing (l)-α-Threofuranosyl (3‘→2‘) Nucleosides: A Four-Carbon Sugar Is Easily Accommodated into the Backbone of DNA. Journal of the American Chemical Society. 2002;124(46):13716-13721.

38. Rangel A, Chen Z, Ayele T, Heemstra J. In vitro selection of an XNA aptamer capable of small-molecule recognition. Nucleic Acids Research. 2018;46(16):8057-8068.
39. Friedman A, Kim D, Liu R. Highly stable aptamers selected from a 2′-fully modified fGmH RNA library for targeting biomaterials. Biomaterials. 2015;36:110-123.
40. Hagiwara K, Fujita H, Kasahara Y, Irisawa Y, Obika S, Kuwahara M. In vitro selection of DNA-based aptamers that exhibit RNA-like conformations using a chimeric oligonucleotide library that contains two different xeno-nucleic acids. Molecular BioSystems. 2015;11(1):71-76.
41. Patra A, Paolillo M, Charisse K, Manoharan M, Rozners E, Egli M. 2′-Fluoro RNA Shows Increased Watson-Crick H-Bonding Strength and Stacking Relative to RNA: Evidence from NMR and Thermodynamic Data. Angewandte Chemie. 2012;124(47):12033-12036.
42. Guerrier-Takada C, Gardiner K, Marsh T, Pace N, Altman S. The RNA moiety of ribonuclease P is the catalytic subunit of the enzyme. Cell. 1983;35(3):849-857.
43. Kruger K, Grabowski P, Zaug A, Sands J, Gottschling D, Cech T. Self-splicing RNA: Autoexcision and autocyclization of the ribosomal RNA intervening sequence of tetrahymena. Cell. 1982;31(1):147-157.
44. Illangasekare M, Sanchez G, Nickles T, Yarus M. Aminoacyl-RNA synthesis catalyzed by an RNA. Science. 1995;267(5198):643-647.
45. LEVY M, GRISWOLD K, ELLINGTON A. Direct selection of trans-acting ligase ribozymes by in vitro compartmentalization. RNA. 2005;11(10):1555-1562.
46. Zaher H, Unrau P. Selection of an improved RNA polymerase ribozyme with superior extension and fidelity. RNA. 2007;13(7):1017-1026.
47. Wochner A, Attwater J, Coulson A, Holliger P. Ribozyme-Catalyzed Transcription of an Active Ribozyme. Science. 2011;332(6026):209-212.
48. Breaker R, Joyce G. A DNA enzyme that cleaves RNA. Chemistry & Biology. 1994;1(4):223-229.
49. Chandra M, Sachdeva A, Silverman S. DNA-catalyzed sequence-specific hydrolysis of DNA. Nature Chemical Biology. 2009;5(10):718-720.
50. Taylor A, Arangundy-Franklin S, Holliger P. Towards applications of synthetic genetic polymers in diagnosis and therapy. Current Opinion in Chemical Biology. 2014;22:79-84.
51. Larsen A, Dunn M, Hatch A, Sau S, Youngbull C, Chaput J. A general strategy for expanding polymerase function by droplet microfluidics. Nature Communications. 2016;7(1).
52. Chaput J, Yu H, Zhang S. The Emerging World of Synthetic Genetics. Chemistry & Biology. 2012;19(11):1360-1371.
53. Shin H, Cho B. Rational Protein Engineering Guided by Deep Mutational Scanning. International Journal of Molecular Sciences. 2015;16(9):23094-23110.
54. Dunn M, Otto C, Fenton K, Chaput J. Improving Polymerase Activity with Unnatural Substrates by Sampling Mutations in Homologous Protein Architectures. ACS Chemical Biology. 2016;11(5):1210-1219.
55. Arnold F. Engineering proteins for nonnatural environments. The FASEB Journal. 1993;7(9):744-749.
56. Gardner A, Jack E. Acyclic and dideoxy terminator preferences denote divergent sugar recognition by archaeon and Taq DNA polymerases. Nucleic Acids Research. 2002;30(2):605-613.
57. Ichida J, Zou K, Horhota A, Yu B, McLaughlin L, Szostak J. An in Vitro Selection System for TNA. Journal of the American Chemical Society. 2005;127(9):2802-2803.
58. Steitz T. DNA Polymerases: Structural Diversity and Common Mechanisms. Journal of Biological Chemistry. 1999;274(25):17395-17398.

59. Chen T, Romesberg F. Directed polymerase evolution. FEBS Letters. 2013;588(2):219-229.
60. Romero P, Arnold F. Exploring protein fitness landscapes by directed evolution. Nature Reviews Molecular Cell Biology. 2009;10(12):866-876.
61. Araya C, Fowler D. Deep mutational scanning: assessing protein function on a massive scale. Trends in Biotechnology. 2011;29(9):435-442.
62. Fowler D, Fields S. Deep mutational scanning: a new style of protein science. Nature Methods. 2014;11(8):801-807.
63. Packer M, Liu D. Methods for the directed evolution of proteins. Nature Reviews Genetics. 2015;16(7):379-394.
64. Leung D, Chen E, Goeddel D. A method for random mutagenesis of a defined DNA segment using a modified polymerase chain reaction. Technique. 1989;1:11-15.
65. Cadwell R, Joyce G. Randomization of genes by PCR mutagenesis. Genome Research. 1992;2(1):28-33.
66. Vanhercke T, Ampe C, Tirry L, Denolf P. Reducing mutational bias in random protein libraries. Analytical Biochemistry. 2005;339(1):9-14.
67. Zaccolo M, Williams D, Brown D, Gherardi E. An Approach to Random Mutagenesis of DNA Using Mixtures of Triphosphate Derivatives of Nucleoside Analogues. Journal of Molecular Biology. 1996;255(4):589-603.
68. Stephens D, Singh S, Permaul K. Error-prone PCR of a fungal xylanase for improvement of its alkaline and thermal stability. FEMS Microbiology Letters. 2009;293(1):42-47.
69. Yang H, Swartz A, Park H, Srivastava P, Ellis-Guardiola K, Upp D et al. Evolving artificial metalloenzymes via random mutagenesis. Nature Chemistry. 2018;10(3):318-324.
70. Lutz S. Beyond directed evolution—semi-rational protein engineering and design. Current Opinion in Biotechnology. 2010;21(6):734-743.
71. Cozens C, Pinheiro V. Darwin Assembly: fast, efficient, multi-site bespoke mutagenesis. Nucleic Acids Research. 2018;46(8):e51-e51.
72. Dosztanyi Z, Magyar C, Tusnady G, Simon I. SCide: identification of stabilization centers in proteins. Bioinformatics. 2003;19(7):899-900.
73. Magyar C, Gromiha M, Sávoly Z, Simon I. The role of stabilization centers in protein thermal stability. Biochemical and Biophysical Research Communications. 2016;471(1):57-62.
74. Chen T, Hongdilokkul N, Liu Z, Adhikary R, Tsuen S, Romesberg F. Evolution of thermophilic DNA polymerases for the recognition and amplification of C2'-modified DNA. Nature Chemistry. 2016;8(6):556-562.
75. Tang L, Gao H, Zhu X, Wang X, Zhou M, Jiang R. Construction of “small-intelligent” focused mutagenesis libraries using well-designed combinatorial degenerate primers. BioTechniques. 2012;52(3).
76. Siloto R, Weselake R. Site saturation mutagenesis: Methods and applications in protein engineering. Biocatalysis and Agricultural Biotechnology. 2012;1(3):181-189.
77. Higuchi R, Krummel B, Saiki R. A general method ofin vitropreparation and specific mutagenesis of DNA fragments: study of protein and DNA interactions. Nucleic Acids Research. 1988;16(15):7351-7367.
78. Byrappa S, Gavin D, Gupta K. A highly efficient procedure for site-specific mutagenesis of full-length plasmids using Vent DNA polymerase. Genome Research. 1995;5(4):404-407.
79. Kadkhodaei S, Memari H, Abbasiliasi S, Rezaei M, Movahedi A, Shun T et al. Multiple overlap extension PCR (MOE-PCR): an effective technical shortcut to high throughput synthetic biology. RSC Advances. 2016;6(71):66682-66694.
80. Peng R, Xiong A, Yao Q. A direct and efficient PAGE-mediated overlap extension PCR method for gene multiple-site mutagenesis. Applied Microbiology and Biotechnology. 2006;73(1):234-240.

81. Herman A, Tawfik D. Incorporating Synthetic Oligonucleotides via Gene Reassembly (ISOR): a versatile tool for generating targeted libraries. Protein Engineering, Design and Selection. 2007;20(5):219-226.
82. Seyfang A, Huaqian Jin J. Multiple site-directed mutagenesis of more than 10 sites simultaneously and in a single round. Analytical Biochemistry. 2004;324(2):285-291.
83. Esvelt K, Carlson J, Liu D. A system for the continuous directed evolution of biomolecules. Nature. 2011;472(7344):499-503.
84. Packer M, Rees H, Liu D. Phage-assisted continuous evolution of proteases with altered substrate specificity. Nature Communications. 2017;8(1).
85. Pinheiro V, Arangundy-Franklin S, Holliger P. Compartmentalized Self-Tagging for In Vitro-Directed Evolution of XNA Polymerases. Current Protocols in Nucleic Acid Chemistry. 2014;:9.9.1-9.9.18.
86. Ghadessy F, Ong J, Holliger P. Directed evolution of polymerase function by compartmentalized self-replication. Proceedings of the National Academy of Sciences. 2001;98(8):4552-4557.
87. Ghadessy F, Ramsay N, Boudsocq F, Loakes D, Brown A, Iwai S et al. Generic expansion of the substrate spectrum of a DNA polymerase by directed evolution. Nature Biotechnology. 2004;22(6):755-759.
88. Povilaitis T, Alzbutas G, Sukackaite R, Siurkus J, Skirgaila R. In vitro evolution of phi29 DNA polymerase using isothermal compartmentalized self replication technique. Protein Engineering, Design and Selection. 2016;29(12):617-628.
89. Salas M, Holguera I, Redrejo-Rodríguez M, de Vega M. DNA-Binding Proteins Essential for Protein-Primed Bacteriophage Φ29 DNA Replication. Frontiers in Molecular Biosciences. 2016;3.
90. Dean F, Nelson J, Giesler T, Lasken R. Rapid Amplification of Plasmid and Phage DNA Using Phi29 DNA Polymerase and Multiply-Primed Rolling Circle Amplification. Genome Research. 2001;11(6):1095-1099.
91. Blanco L, Bernad A, Lharo J, Martins G, Garmendia C, Salas M. Highly Efficient DNA Synthesis by the Phage φ DNA Polymerase. The Journal of Biological Chemistry. 1989;254(15):8935-8940.
92. Torres L, Pinheiro V. Xenobiotic Nucleic Acid (XNA) Synthesis by Phi29 DNA Polymerase. Current Protocols in Chemical Biology. 2018;10(2):e41.
93. Bernad A, Blanco L, Lázaro J, Martín G, Salas M. A conserved 3′→5′ exonuclease active site in prokaryotic and eukaryotic DNA polymerases. Cell. 1989;59(1):219-228.
94. Wu J, de Paz A, Zamft B, Marblestone A, Boyden E, Kording K et al. DNA binding strength increases the processivity and activity of a Y-Family DNA polymerase. Scientific Reports. 2017;7(1).
95. Rodriguez I, Lazaro J, Blanco L, Kamtekar S, Berman A, Wang J et al. A specific subdomain in 29 DNA polymerase confers both processivity and strand-displacement capacity. Proceedings of the National Academy of Sciences. 2005;102(18):6407-6412.
96. Bryson D, Fan C, Guo L, Miller C, Söll D, Liu D. Continuous directed evolution of aminoacyl-tRNA synthetases. Nature Chemical Biology. 2017;13(12):1253-1260.
97. Sprinthall R. Basic statistical analysis. 9th ed. Harlow : Pearson; 2014.
98. Clark M. Some methods for statistical analysis of multimodal distributions and their application to grain-size data. Journal of the International Association for Mathematical Geology. 1976;8(3):267-282.
99. Bewick V, Cheek L, Ball J. Statistics review 10: Further nonparametric methods. Critical Care. 2004;8(3):196.
100. Ramírez-Sarmiento C, Noel J, Valenzuela S, Artsimovitch I. Interdomain Contacts Control Native State Switching of RfaH on a Dual-Funneled Landscape. PLOS Computational Biology. 2015;11(7):e1004379.

101. Desai M, Weissman D, Feldman M. Evolution Can Favor Antagonistic Epistasis. Genetics. 2007;177(2):1001-1010.
102. Otwinowski J, Plotkin J. Inferring fitness landscapes by regression produces biased estimates of epistasis. Proceedings of the National Academy of Sciences. 2014;111(22):E2301-E2309.
103. Hall T. BioEdit, version 5.0.6. North Carolina State University: Ibis Biosciences; 2001.
104. Bruni R, Prosperi M, Marcantonio C, Amadori A, Villano U, Tritarelli E et al. A computational approach to identify point mutations associated with occult hepatitis B: significant mutations affect coding regions but not regulative elements of HBV. Virology Journal. 2011;8(1):394.
105. Truniger V, Lazaro J, Esteban F, Blanco L, Salas M. A positively charged residue of phi29 DNA polymerase, highly conserved in DNA polymerases from families A and B, is involved in binding the incoming nucleotide. Nucleic Acids Research. 2002;30(7):1483-1492.
106. Simonetti F, Teppa E, Chernomoretz A, Nielsen M, Marino Buslje C. MISTIC: mutual information server to infer coevolution. Nucleic Acids Research. 2013;41(W1):W8-W14.
107. Gloor G, Martin L, Wahl L, Dunn S. Mutual Information in Protein Multiple Sequence Alignments Reveals Two Classes of Coevolving Positions†. Biochemistry. 2005;44(19):7156-7165.
108. Villaverde A, Ross J, Morán F, Banga J. MIDER: Network Inference with Mutual Information Distance and Entropy Reduction. PLoS ONE. 2014;9(5):e96732.
109. Buslje C, Santos J, Delfino J, Nielsen M. Correction for phylogeny, small number of observations and data redundancy improves the identification of coevolving amino acid pairs using mutual information. Bioinformatics. 2009;25(9):1125-1131.
110. Truniger V, Lázaro J, de Vega M, Blanco L, Salas M. ϕ29 DNA Polymerase Residue Leu384, Highly Conserved in Motif B of Eukaryotic Type DNA Replicases, Is Involved in Nucleotide Insertion Fidelity. Journal of Biological Chemistry. 2003;278(35):33482-33491.
111. Truniger V, Lázaro J, Salas M. Two Positively Charged Residues of φ29 DNA Polymerase, Conserved in Protein-primed DNA Polymerases, are Involved in Stabilisation of the Incoming Nucleotide. Journal of Molecular Biology. 2004;335(2):481-494.
112. Saturno J, Blanco L, Salas M, Esteban J. A Novel Kinetic Analysis to Calculate Nucleotide Affinity of Proofreading DNA Polymerases:. Journal of Biological Chemistry. 1995;270(52):31235-31243.
113. Perez-Arnaiz P, Lazaro J, Salas M, de Vega M. Involvement of 29 DNA polymerase thumb subdomain in the proper coordination of synthesis and degradation during DNA replication. Nucleic Acids Research. 2006;34(10):3107-3115.
114. Cozens C, Pinheiro V, Vaisman A, Woodgate R, Holliger P. A short adaptive path from DNA to RNA polymerases. Proceedings of the National Academy of Sciences. 2012;109(21):8067-8072.
115. Pinheiro V, Taylor A, Cozens C, Abramov M, Renders M, Zhang S et al. Synthetic Genetic Polymers Capable of Heredity and Evolution. Science. 2012;336(6079):341-344.
116. Liu C, Cozens C, Jaziri F, Rozenski J, Maréchal A, Dumbre S et al. Phosphonomethyl Oligonucleotides as Backbone-Modified Artificial Genetic Polymers. Journal of the American Chemical Society. 2018;140(21):6690-6699.
117. Kamtekar S, Berman A, Wang J, Lázaro J, de Vega M, Blanco L et al. Insights into Strand Displacement and Processivity from the Crystal Structure of the Protein-Primed DNA Polymerase of Bacteriophage φ29. Molecular Cell. 2004;16(6):1035-1036.

118. Fowler D, Araya C, Fleishman S, Kellogg E, Stephany J, Baker D et al. High-resolution mapping of protein sequence-function relationships. Nature Methods. 2010;7(9):741-746.
119. Espadaler J, Querol E, Aviles F, Oliva B. Identification of function-associated loop motifs and application to protein function prediction. Bioinformatics. 2006;22(18):2237-2243.
120. Yu H, Yan Y, Zhang C, Dalby P. Two strategies to engineer flexible loops for improved enzyme thermostability. Scientific Reports. 2017;7(1).
121. Lantto J, Ohlin M. Functional Consequences of Insertions and Deletions in the Complementarity-determining Regions of Human Antibodies. Journal of Biological Chemistry. 2002;277(47):45108-45114.
122. Berman A, Kamtekar S, Goodman J, Lázaro J, de Vega M, Blanco L et al. Structures of phi29 DNA polymerase complexed with substrate: the mechanism of translocation in B-family polymerases. The EMBO Journal. 2007;26(14):3494-3505.
123. Hansen C, Wu L, Fox J, Arezi B, Hogrefe H. Engineered split in Pfu DNA polymerase fingers domain improves incorporation of nucleotide-phosphate derivative. Nucleic Acids Research. 2010;39(5):1801-1810.
124. Tests for Two Proportions [Internet]. NCSS;. Available from: https://ncss-wpengine.netdna-ssl.com/wp-content/themes/ncss/pdf/Procedures/PASS/Tests_for_Two_Proportions.pdf
125. Pinheiro V. Engineering-driven biological insights into DNA polymerase mechanism. Current Opinion in Biotechnology. 2019;60:9-16.
126. Tizei P, Csibra E, Torres L, Pinheiro V. Selection platforms for directed evolution in synthetic biology. Biochemical Society Transactions. 2016;44(4):1165-1175.
127. Gupta A, Adami C. Strong Selection Significantly Increases Epistatic Interactions in the Long-Term Evolution of a Protein. PLOS Genetics. 2016;12(3):e1005960.
128. Dunn S, Wahl L, Gloor G. Mutual information without the influence of phylogeny or entropy dramatically improves residue contact prediction. Bioinformatics. 2007;24(3):333-340.
129. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Science. 2008;1(3):409-417.