

Title:

Directed evolution of mesophilic HNA polymerases providing insight into DNA polymerase mechanisms.

Authors and affiliations:

Paola Handal-Marquez^{1,2}, Leticia L. Torres¹ and Vitor B. Pinheiro^{2*}

¹ University College London, Department of Structural and Molecular Biology, Gower Street, WC1E 6BT, London, UK

² KU Leuven, Rega Institute for Medical Research, Department of Pharmaceutical and Pharmacological Sciences, Herestraat, 49 – box 1041, 3000 Leuven, Belgium

* Corresponding author: e-mail: v.pinheiro@kuleuven.be

Keywords: Directed evolution, XNA synthase, Phi29 DNA polymerase, high-throughput biochemistry

Abstract

Detailed biochemical characterization of natural and mutant enzymes provides essential clues to understand their mechanisms. There are, however, limits to the throughput of such approaches and they are not without errors. DNA polymerases have benefited from over 50 years of detailed study and remain not fully understood. As such, methods that allow high-throughput interrogation of variants, and viable analysis pipelines to identify relevant variants, become an important tool to accelerate research. Using the DNA polymerase from *B. subtilis* Phi29 bacteriophage as a model, we demonstrate how coupling focused libraries, selection and deep sequencing can be combined to identify variants of interest for characterization. As selection parameters can be controlled, different areas of an enzyme's mechanism can be explored. Focusing selection on faster HNA (1,5-anhydrohexitol nucleic acid) synthesis, we identified P562del as a variant of interest, enriching significantly between rounds. Characterization confirmed its faster HNA synthesis initiation but lower processivity and fidelity. P562 is a non-conserved residue, unlikely to be selected by more traditional approaches, but its deletion recapitulates knowledge on how Phi29 exonuclease, thumb and TPR2 subdomains regulate polymerase function. Our data further support the hypothesis that Phi29 shows a two-state binding to its template: a fast non-replicative complex that transitions to a replication-competent state.

Introduction

Information storage and replication is an essential process in all living organisms, and it is carried out by specialist enzymes: template-dependent DNA (or RNA) polymerases. Despite their central role, the ancient origin of those proteins has given evolution time and opportunity to explore a vast array of characteristics and specializations: from high-fidelity highly processive replicative polymerases (e.g. *Bacillus subtilis* Phi29 bacteriophage DNA polymerase (Garmendia et al., 1992) to low-fidelity, low processivity DNA translesion repair enzymes (e.g. the mammalian Pol eta (Waters et al., 2009); from monomeric polymerases to large complexes; from mesophilic to hyperthermophilic enzymes.

Those characteristics and specializations are the result of functional and sequence constraints imposed during the natural evolution of DNA polymerases, and they are not yet fully understood. Replicative DNA-dependent DNA polymerases are often multifunctional enzymes (harboring both polymerase and exonuclease activities) with complex dynamics and operating under multiple, and sometimes overlapping, layers of regulation. For instance, DNA polymerases can often discriminate between their natural substrate (deoxyribose nucleotide triphosphates) and other natural analogues present in the cellular milieu (ribonucleotides, partially phosphorylated nucleotides, and damaged nucleobases), even though the analogues are present at a higher concentration than the substrates (Wang et al., 2012). This is done not only at incorporation in the active site of the enzyme (Tsai & Johnson, 2006; Wang et al., 2012), but also at multiple points during the elongation process (Freund et al., 2022; Khare & Eckert, 2002).

A consequence of their complexity and as a result of their evolution, specific mutations can have different impacts on the function of homologous DNA polymerases. For instance, the *Escherichia coli* bacteriophage T4 has evolved around a high-fidelity DNA polymerase for its successful infection of the bacterial host and even small changes in function, such as loss of 3'→5' exonuclease activity, can result in a 10-fold reduction of bacteriophage viability (Frey et al., 1993). The same significant drop in viability is not observed in the T4-related RB69 bacteriophage (Bebenek et al., 2001). In another example, mutation of a β hairpin in both T4 and RB69 polymerases leads to loss of fidelity but a similar mutation on the *S. cerevisiae* DNA polymerase δ (also a B-family

polymerase) does not. Therefore, the mutational space naturally accessible to each enzyme can differ.

Detailed biochemical and mutational analysis has been very successful at identifying some of the sequence and structural determinants for the different DNA polymerase functions, but they are limited to the models we use to understand polymerases, to how those can be efficiently tested, and to how generalizable they are.

DNA polymerase engineering, whether for DNA applications (e.g. PCR or DNA sequencing) or for the synthesis of xenobiotic nucleic acids (XNAs), has identified multiple new mutations that modulate enzymatic function and further contribute to our understanding of those enzymes (Laos et al., 2014; Pinheiro, 2019). Nonetheless, the gain in functional understanding is not the goal of those efforts, instead being a byproduct of selection and rational design methodologies, and often used as a tool to rationalize the contribution of isolated mutants.

However, those same directed evolution platforms, that rely on high-throughput screening, have the potential to be refocused towards improving our mechanistic understanding of DNA polymerases. While the design of the libraries used in directed evolution creates a boundary for the sequence space that can be explored in a selection campaign, the high-throughput screening used in selection has the power to bypass knowledge gaps: it can operate without the need of a prior hypothesis, and it can overcome previous misinterpretations. The first reduces the likelihood of confirmation biases, while the latter is ideal to challenge established paradigms. In addition, and particularly in the case of polymerases, the use of unnatural substrates that cannot be efficiently incorporated has additional benefits, such as slowing down reaction rates (changing the shape of the sequence space that can be experimentally isolated) and potentially not equally affecting all functional parameters in the polymerases (thus allowing deconvolution of complex interactions).

For example, in engineering a hyperthermophilic DNA polymerase from *Thermococcus gorgonarius* for the synthesis of XNAs, a variant, named 6G12, was isolated for its ability to efficiently synthesize HNA (1,5-anhydrohexitol nucleic acids) (Pinheiro et al., 2012). It

harbored a total of 18 mutations: four introduced to facilitate engineering (e.g. disabling its exonuclease activity), and 14 isolated from selection. At the time, we rationalized that reshaping of the enzyme thumb domain was needed to accommodate the distorted hybrid nascent HNA-DNA helix. Nonetheless, later efforts to engineer a mesophilic XNA polymerase from the *B. subtilis* bacteriophage Phi29, determined that reducing (N62D mutation) or disabling (D12A) the exonuclease activity of the polymerase was already sufficient to enable HNA synthesis – albeit substantially less efficient than its DNA activity. An updated hypothesis is, therefore, that stability of the polymerase XNA/DNA duplex is the driving parameter for efficient XNA synthesis, as observed for a DNA polymerase engineered for processive RNA synthesis (Cozens et al., 2012).

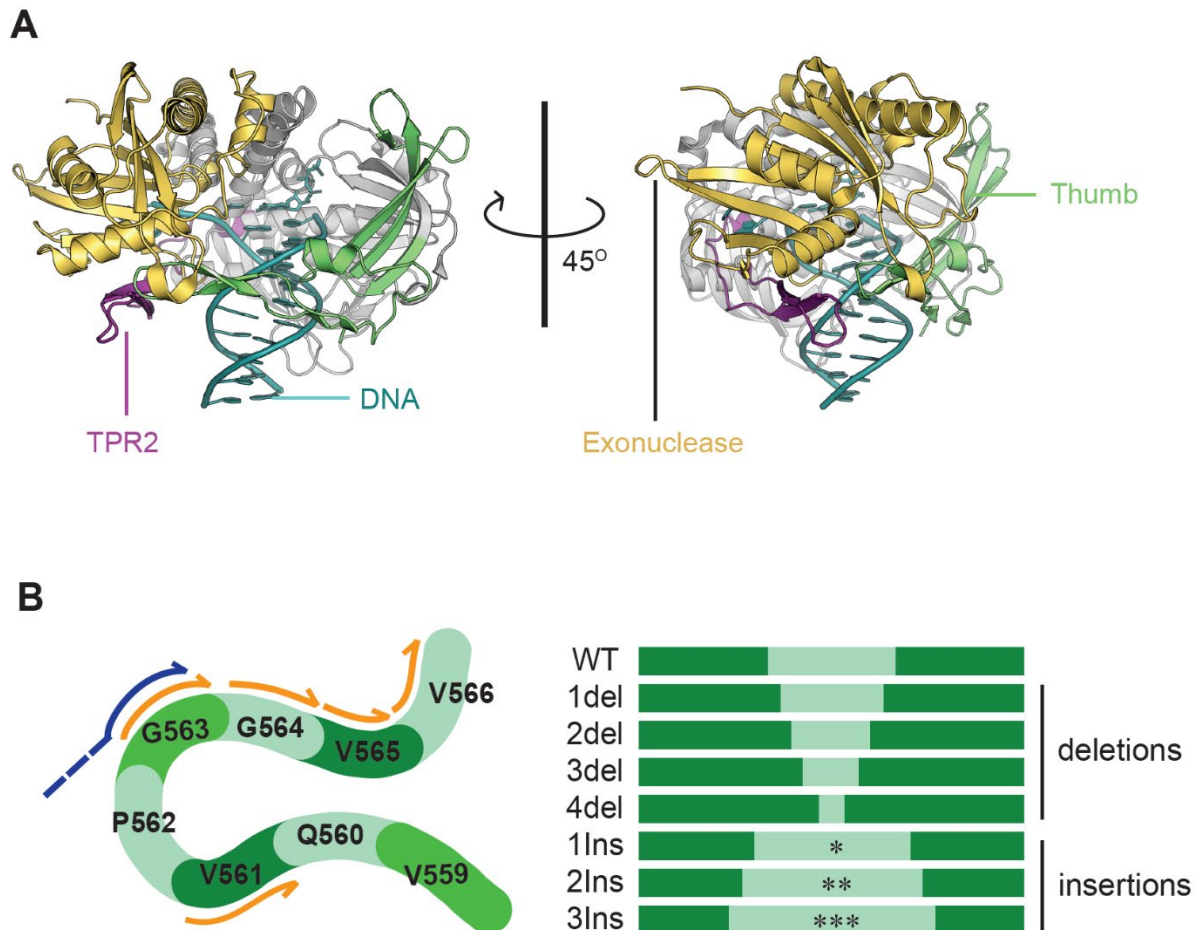


Figure 1. Structure of the closed ternary complex of Phi29 DNA Polymerase and InDel library design. (A) Three domains in the phi29 DNA polymerase (PDB: 2PYJ) are involved in creating a clamp around the nascent DNA duplex: the polymerase thumb

(green), the exonuclease domain (yellow) and the TPR2 domain (purple). The DNA duplex is shown in cyan. (B) InDel mutagenesis of the thumb loop through inverse PCR (iPCR). A single reverse primer in combination with forward primers harbouring 1-3 NNS codons or with priming sites that skipped 1-4 codons were used to generate focused libraries or deletion mutants respectively. The same approach was used to target the TPR2 and exonuclease loops.

Phi29 DNA polymerase is known to be a highly processive enzyme (processivity defined as number of nucleotides incorporated by a single polymerase in a single DNA-binding event), capable of up to 70 kb synthesis per binding event (Lieberman et al., 2010). Phi29 processivity is closely associated with the terminal protein region 2 (TRP2) subdomain, which forms a structure akin to a sliding clamp around template and nascent DNA, and which harbours multiple contacts with the thumb (along residues V559-V566) and exonuclease (along residues T15-V24 and F65-N72) subdomains (see **Figure 1A**).

Deletion of the TRP2 subdomain, does not affect the ratio between exonuclease and polymerase activities of the polymerase but significantly decreases polymerase affinity for DNA and its processivity (Rodriguez et al., 2005). This is analogous to what is observed in the T7 bacteriophage DNA polymerase (gp5), where high processivity is only observed when a clamp-like structure is formed around the DNA by the binding of the thioredoxin processivity factor (Tran et al., 2012).

The close contacts between TRP2, thumb and exonuclease can, in principle, affect other polymerase processes, including exonuclease activity (given the packing against an alpha helix harbouring one of the catalytic exonuclease residues – D66), exonuclease sampling (DNA switching between polymerase and exonuclease catalytic sites, by modulating how exonuclease and thumb interact), strand displacing activity (through the interaction TRP2 with template DNA) and DNA binding.

This three-way junction is therefore also a relevant region for the engineering of efficient HNA polymerases based on the Phi29 DNA polymerase chassis, as the optimisation of the sequence length and composition in these loops could potentially create new

molecular interactions that stabilize protein binding to the nascent heteroduplex or avoid steric clashes that could be contributing to the lower processivity observed during HNA synthesis (Handal Marquez, 2019). If any of the other processes described above can be affected by changes in this region, that could also lead to improvements in HNA synthesis.

The potential role of enzyme processivity on HNA synthesis, together with the biochemical evidence of how TRP2 is involved in modulating it, establish a tractable system for investigating whether selection platforms can be used to gain biological insights into complex enzymatic mechanisms. Libraries can be designed with minimal assumptions, selection conditions can be tailored towards investigating specific parameters, and deep sequencing analysis of the selection output rapidly identifies variants of interest: enzymes enriched by the selection conditions.

Here, we demonstrate how this system can be assembled and we use it to identify a single residue in the DNA polymerase thumb (P562) implicated in Phi29 DNA polymerase processivity. We characterise the isolated enzyme to demonstrate the impact of the mutation on DNA affinity, enzyme processivity, fidelity and strand displacement activities. We confirm that Phi29 DNA polymerase needs not only to bind DNA but also to switch to a stable conformation to achieve highly processive DNA synthesis.

New Approaches

We demonstrate that high-throughput selection can be harnessed for high-throughput biochemistry, allowing novel mechanistic insights in enzyme function. The approach relies on an integrated approach to library design, selection and data analysis (**Figure 2**), so that relevant enzyme variants can be identified and further characterised. The approach is scalable but hinges on the effectiveness and specificity of the selection strategy being implemented.

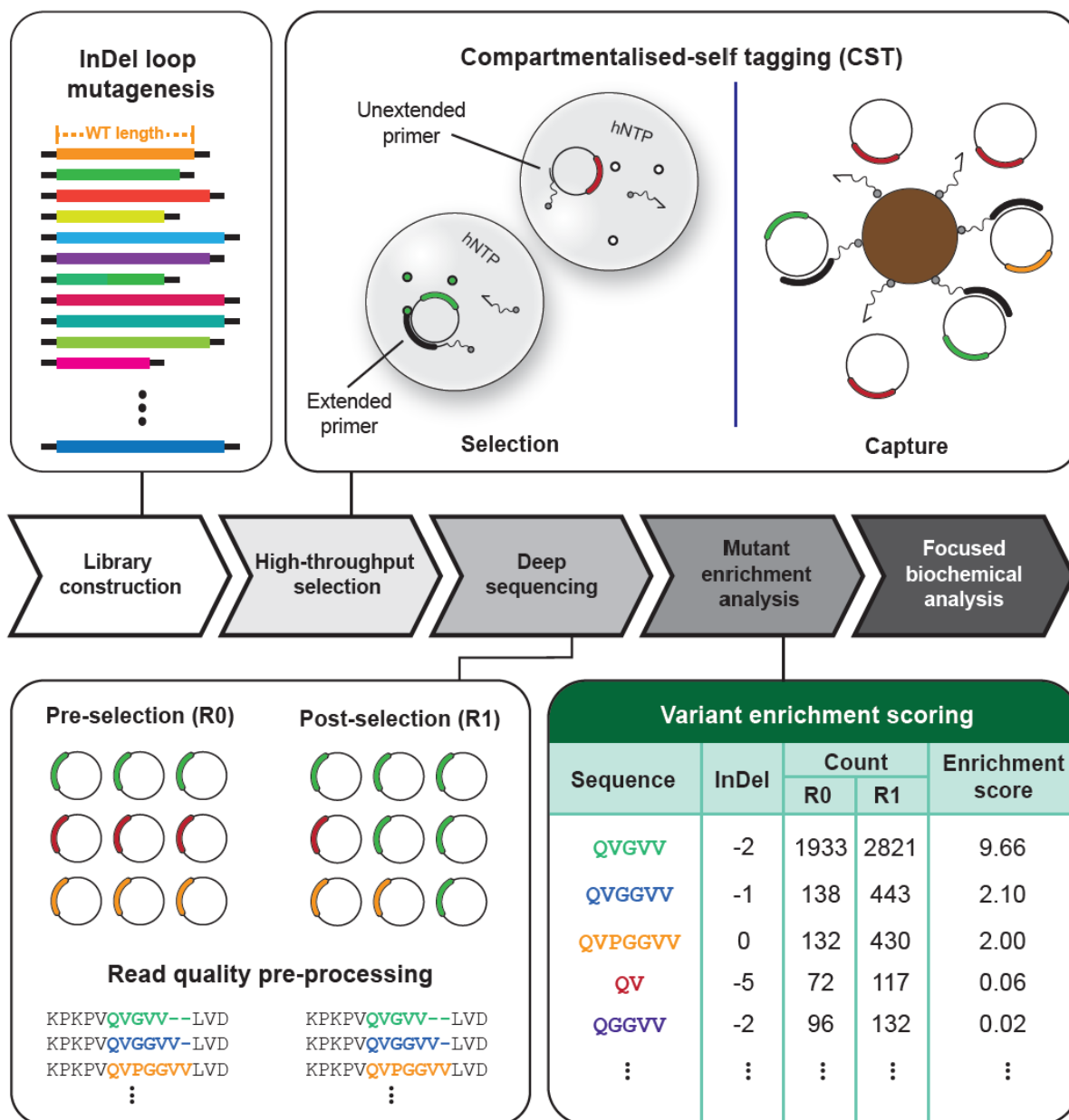


Figure 2: Mesophilic XNA polymerase directed evolution pipeline. A library is initially constructed – represented (top left panel) by different color/lengths compared to the wild-type. The library is then subjected to a round(s) of selection through compartmentalized self-tagging (top right panel). The library pre- and post-selection is sequenced through NGS (bottom left panel). The proportion of highly active variants (e.g., green) should increase whereas that of variants with lower activity (e.g., orange, and red) should decrease. NGS data is filtered for quality. Pre-processed data is then analyzed (bottom right). The frequency of each mutant corresponds to its abundance in the library pre- and

post-selection. The enrichment score is used to evaluate the increase or decrease in abundance of specific mutants within the population and serves as starting point for further analysis.

Results

Library design: focused libraries sampling beyond biochemical knowledge.

There are four interconnected steps that are crucial for harnessing selection as a tool to investigate biology: library design, selection methodology, selection conditions and downstream data analysis.

The wide range of available DNA library assembly strategies allows length and compositional variation to be introduced, randomly or in a targeted manner, across a gene of interest (Chica et al., 2005; Lutz, 2011), such that technical bottlenecks in library generation are no longer relevant. Instead, design is constrained by the number of variants that can be generated and sampled (linking design to selection methodology), considering the likely destabilizing effect of a high mutational burden to the protein being engineered.

Although there are multiple possible contacts between TRP2, thumb and exonuclease subdomains, we hypothesised that contact between the three loops (TRP2: L406-L412, thumb: V559-V566, and exonuclease T15-V24) is likely functional, capable of providing the dynamic response needed for polymerase-exonuclease balance and enzyme processivity. Only limited diversity can be introduced in the 29 residues, given the rapid rise in the number of variants and the poor stability of the Phi29 DNA polymerase (half-life of 37 ± 4.5 min at 40°C, when in the presence of stabilizing DNA template - (Povilaitis et al., 2016).

Starting from a previously reported thermostabilized Phi29 DNA polymerase (Povilaitis et al., 2016) ($t_{1/2} > 16$ h at 40°C in the presence of DNA), harboring the additional D12A mutation (which disables exonuclease activity) to allow HNA synthesis, we opted to concomitantly explore loop length in the three loops. Firstly, because insertions and

deletions (indels) tend to be least disruptive to protein stability when they occur in loops. Secondly, we hypothesized that loop distortions, due to indels, would be a more sensitive tool to study loop-loop interaction than patterns of substitution, while also minimizing library diversity. That is further supported by sequence alignment with Phi29 DNA polymerase homologues (**Figure S1**), showing that these regions can accept indels.

Secondly, given that the HNA-DNA heteroduplex is expected to be wider than a DNA duplex, we hypothesized that an insertion could lead to stabilizing interactions between the three subdomains, disrupted by the wider substrate, being re-established – as reported for the engineering of the *Chlorella* virus DNA ligase towards an XNA ligase (Vanmeert et al., 2019).

The tips of the loops (residues D19, N409 and P562) were used as seed in design, with up to three insertions being introduced upstream of those seeds, or up to four deletions targeting the seed and downstream residues, as shown in **Figure 1B**. Exonuclease, TRP2 and thumb libraries had each a theoretical maximum diversity of approximately 8425 variants – well within the sampling range of the chosen selection platform (approximately 2×10^7 per experiment) and analysis pipeline (between 5.0 and 7.0×10^4 of sequences per run).

Although guided by the polymerase structures and phylogeny, the libraries are agnostic to the biochemical knowledge that TRP2 (Rodriguez et al., 2005; Rodríguez et al., 2009) and the Phi29 DNA polymerase C-terminus (Kamtekar et al., 2004; Truniger et al., 2004) are crucial for processivity, with a single previous report targeting one of those residues (Rodríguez et al., 2009).

Using enrichment to identify relevant mutations in selection

High-throughput selections rely on systems that link the phenotype being sought with efficient recovery of the responsible genotype, so that at each round of selection, the genotype of active enzymes becomes more frequent.

A variety of selection methodologies have been demonstrated for the selection of DNA and XNA polymerases, including *in vivo* (Patel & Loeb, 2000), phage display (Xia et al.,

2002), and emulsion-based methods such as CSR (compartmentalised self-replication) (Ghadessy et al., 2001), CST (compartmentalised self-tagging) (Pinheiro et al., 2014), and CBL (compartmentalised bead-labelling) (Houlihan et al., 2020). Technical characteristics of each method shape the range of polymerase activities that can be selected and what activities can emerge as selection parasites (active enzymes that co-enrich with the desired activity).

CST relies on polymerases capable of stabilizing the binding of the selection primer against their own gene through XNA synthesis (**Figure 2**). Given the diminishing return per XNA incorporation to the overall primer stabilization (i.e. analogous to oligonucleotide length effects on primer melting temperature), the pressure of CST selection is highest on the first few incorporations, with CST selections working akin to a high-pass filter (i.e. selection is unable to discriminate between polymerases that can synthesise beyond a certain number of incorporations). Expected parasites in this CST selection include polymerases that can make use of the low concentrations of natural cellular dNTPs still present, polymerases that express to significantly higher levels or that are more thermostable.

For the selection of HNA polymerases with higher processivity, extension time is a key parameter in selection, since shorter extension times are likely to benefit highly processive enzymes as well as polymerases that can start XNA synthesis faster. The latter was not expected to be a significant possibility, considering Phi29 high affinity for DNA ($K_D = 1\mu\text{M}$ in conditions close to the selection conditions (Lázaro et al., 1995)) and in view of our experience engineering thermostable XNA polymerases.

A single round of selection for HNA synthesis was carried out for each library, followed by deep sequencing of naïve and selected libraries for analysis. Frequency counting post-selection can be a suitable route to identify best candidates in circumstances where libraries are not biased, for selection methodologies that have very low background (i.e. recovered genotypes that are randomly recovered) and for late-stage selection campaigns (i.e. libraries that have already been carried through multiple selection rounds and that are populated by few variants) (Wrenbeck et al., 2017). As our libraries were pooled (by site across all designed lengths) during assembly, we expected them to show

significant coverage biases. In addition, emulsion-based polymerase selection platforms show significant background due to PCR amplification post-selection. Therefore, we opted to focus on enrichment, rather than simply frequency, as a measure of selection.

Sequencing errors and suboptimal sampling of the libraries also contribute to the complexity of traditional analyses, with most simple enrichment estimators unable to cope with sequences that appear only in the selected pool. Treating the observation of individual sequence variants as independent Poisson distributions permits pre- and post-selection populations to be tested for statistically significant differences, via an E-test (Krishnamoorthy & Thomson, 2004). However, multiple testing across all variants rapidly limits the usefulness of the approach.

Instead, we filtered the datasets to focus on variants that were present in both datasets and that increased in frequency between pre- and post-selection. In addition, we scored them not only on their change in frequency but also on their average number of observations, to minimise false positive detection by concentrating analysis on better sampled variants. Significantly enriched variants are shown in **Table 1**.

Thumb loop variants	InDel	R0 Count (n= 35849)	R1 Count (n= 46907)	EC	E-test
KPKPVQV--GVVLVD	-2	1933	2821	9.7E+00	0.0E+00
KPKPVQV-GGVVLVD	-1	138	443	1.1E+00	0.0E+00
KPKPVQVPPGGVVLVD	0	132	430	1.0E+00	0.0E+00
KPKPVQV-----LVD	-5	72	117	4.6E-02	1.4E-01
KPKPVQ--GGVVLVD	-2	96	132	1.6E-02	7.2E-01
TPR2 loop variants	InDel	R0 Count (n= 75695)	R1 Count (n= 73519)	EC	E-test
YLKE NGALGFRLGEE	0	123	730	3.5E+00	0.0E+00
YLKEGRNGALGFRLGEE	3	56	94	4.0E-02	9.2E-04
YLK--GALGFRLGEE	-2	49	79	2.7E-02	4.4E-03
YLKEAGNGALGFRLGEE	3	48	73	2.2E-02	1.4E-02
YLKERGNGALGFRLGEE	3	38	57	1.3E-02	3.5E-02
Exonuclease loop variants	InDel	R0 Count (n= 32402)	R1 Count (n= 4544)	EC	E-test
TKVE --RVWAYGYMN	-2	2329	406	2.4E+01	1.0E-01
TKVE DCRVWAYGYMN	0	259	111	3.0E+00	1.0E-06
TKVEARDCRVWAYGYMN	2	192	46	5.0E-01	1.6E-01
TKVEPGDCRVWAYGYMN	2	130	30	2.1E-01	3.3E-01
TKVERGDCRVWAYGYMN	2	224	38	1.9E-01	5.9E-01

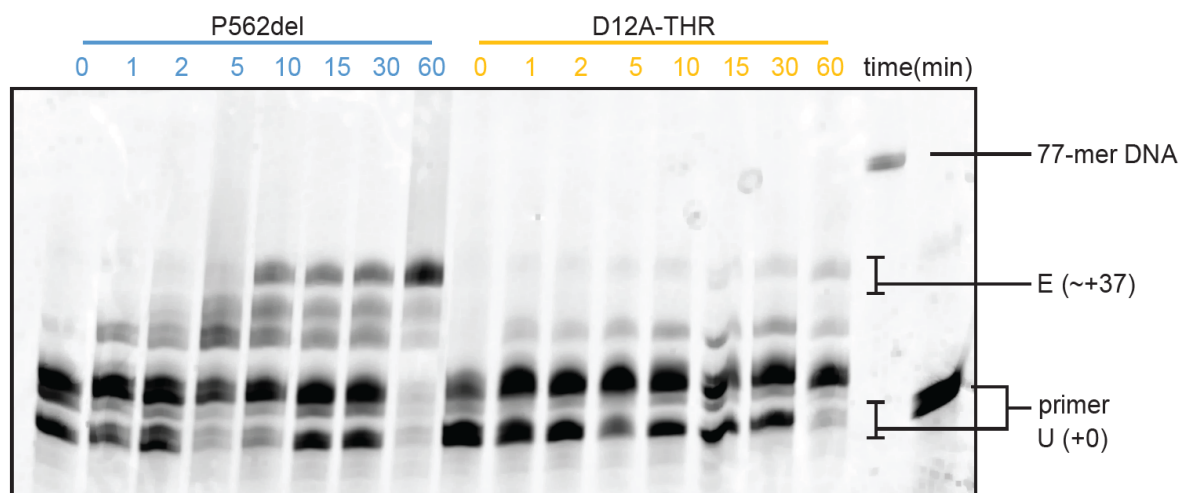
Table 1. Most enriched variants isolated from individual libraries. Complex enrichment (EC) scores and statistical significance for the top 5 variants from each of the phi29 DNAP Exonuclease, TRP2, and thumb InDel loop libraries. The EC is calculated as the difference in frequency between rounds times the average number of counts. Wild-type sequences and their corresponding scores are shown in blue. Sequences with higher EC than the wild-type and significantly enriched (E-test p-value <0.05) are shown in red. Deletions are shown as hyphens and insertions are shown in not in bold.

Out of the identified mutants, P562del (Phi29 DNAP M8R, D12A, V51A, M97T, G197D, E221K, Q497P, K512E, F526L, ΔP562) and PG563del (Phi29 DNAP M8R, D12A, V51A, M97T, G197D, E221K, Q497P, K512E, F526L, ΔP562-G563) variants were notable, having enriched better than the parental enzyme (D12A-THR: Phi29 DNAP M8R, D12A, V51A, M97T, G197D, E221K, Q497P, K512E, F526L) in selection. In view that deletion of P562 alone was sufficient for the increased enrichment, we chose to focus on this variant for further characterisation.

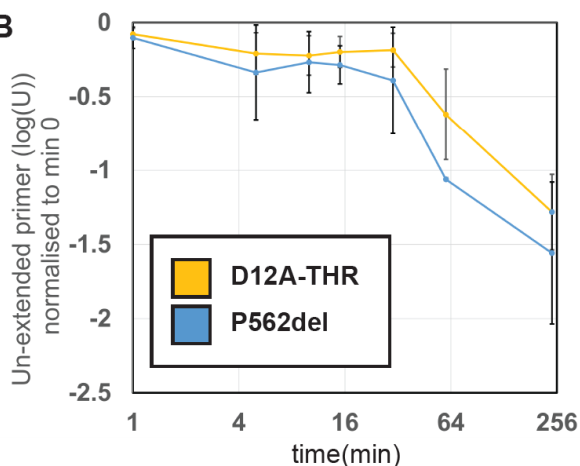
Deletion of P562 reduces processivity and fidelity

Although kinetic parameters of incorporation are often used to describe DNA polymerase activity and fidelity (when carried out with non-cognate substrates), they often do not consider effects that emerge during elongation, which are known to be significant for XNA synthesis (Lutz et al., 1999; Xia et al., 2002). For instance, hexitol nucleotide triphosphate (hNTP) incorporation kinetics is comparable to dNTP kinetics (Vastmans et al., 2000) in the B-family Vent DNA polymerase, yet Vent cannot incorporate more than 7 hNTPs in a row.

A



B



C

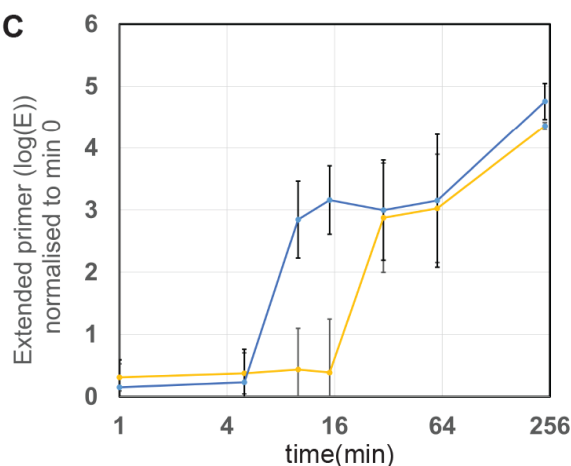


Figure 3. P562 deletion in phi29 DNAP enables faster initiation of HNA synthesis.

Primer extension assays for HNA synthesis were carried out by incubating 10 pmol of

single stranded DNA template pre-annealed to 1 pmol of fluorescently labelled DNA primer with 12 nM D12A-THR or P562del phi29 DNAP over different amounts of time. (A) HNA primer extension products synthesized by D12A-THR and P562 mutants separated by denaturing PAGE. Unextended primer (U) and extension products of ~37 incorporations (E) are highlighted. (B) Average depletion of unextended primer (U) by D12A-THR (orange) and P562del (blue) normalized to the earliest time point (0 min). (C) Average extended products with ~37 incorporations (E) by D12A-THR and P562del normalized to the earliest time point (0 min). P562del shows a faster start to its hNTP incorporations than D12A-THR.

Therefore, initial characterisation focused on confirming the increased activity of P562del for HNA synthesis by primer extension using concentration-normalised purified enzymes – confirming the results of selection and analysis. In short reactions (60 min extension), as shown in **Figure 3**, P562del indeed outperforms the starting D12A-THR polymerase, synthesising longer HNAs (~37 incorporations vs. ~18 from D12A-THR) and more efficiently (less visible unextended primer). Longer incubations (**Figure S2A**), however, allow D12A-THR to match and eventually outperform P562del – differences that cannot be seen with the natural nucleoside triphosphates. Notably, like the previously reported Δ TPR2 mutant (Rodriguez et al., 2005), HNA synthesis by the P562del is distributive while D12A-THR is processive (with few visible accumulated intermediates between primer and full-length extension). Primer extension using alternative templates ruled out possible sequence-specific effects (**Figure S2B**).

Multiple polymerase mechanisms could account for the “early burst” phenotype observed. For instance, changes in binding affinity to the template (which begins as dsDNA but changes into an HNA/DNA heteroduplex as the polymerase extend the primer), changes in processivity (by increasing the dissociation rate of the enzyme from the template), or a drop in fidelity (misincorporations stall the polymerase and can also lead to more frequent sampling of the editing conformations). We therefore focused characterization on assays that could give some insight into those processes, always comparing P562del, to the

parental enzymes, with (Exo+THR: Phi29 DNAP M8R, V51A, M97T, G197D, E221K, Q497P, K512E, F526L) or without (D12A-THR) active exonuclease function.

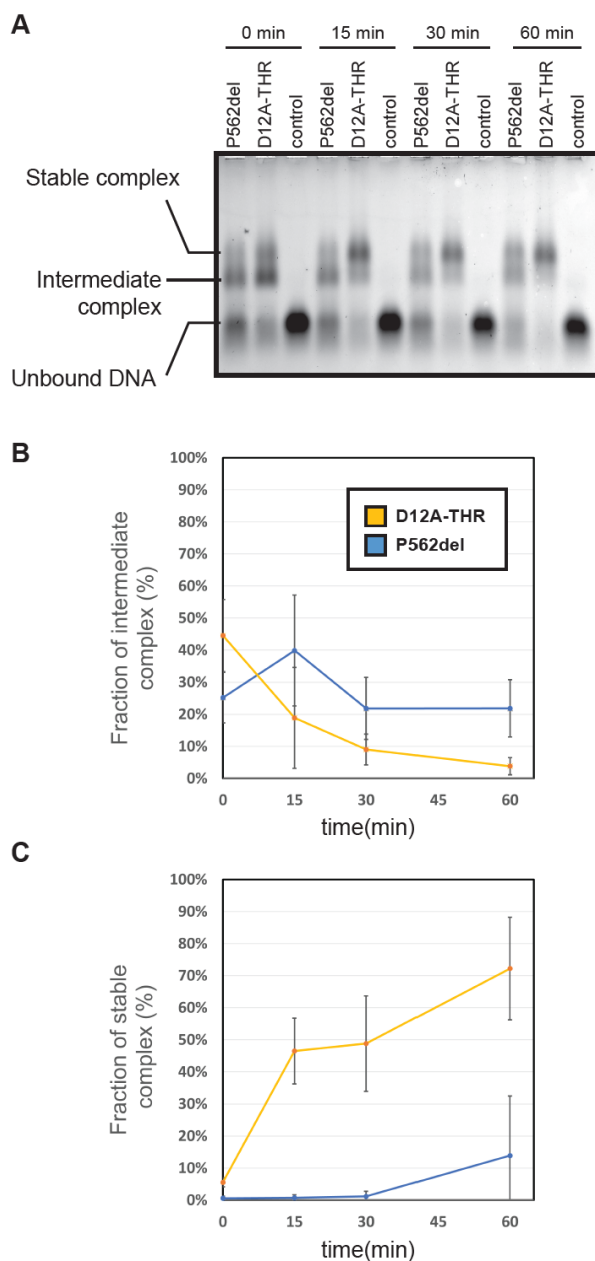


Figure 4. Phi29 DNAP P562del reduced DNA binding capacity. For the EMSA, 60 pM commercial phi29 DNAP (control), D12A-THR or P562del were allowed to bind a fluorescently labelled primer pre-annealed to a ssDNA template (A) Reactions with commercial phi29 DNAP show no binding, probably as a result of storage buffer components interfering with quantification. (B) Fraction of intermediate Pol-DNA complex,

and (C) the fraction of stable Pol-DNA complex by D12A-THR (orange) and P562del (blue) over time.

Electrophoretic mobility shift assays (EMSA) showed a significant difference between P562del and D12A-THR (**Figure 4**). The parental enzyme, as previously reported, shows complexes with two different mobilities: an unstable intermediate that transitions to a stable replication-competent complex (Rodriguez et al., 2005). The deletion of P562 changes that process, with both complexes co-existing even after long incubations, suggesting that the stability of the replication-competent complex is reduced.

Rolling circle amplification (RCA) confirmed the detrimental impact of D12A (Bernad et al., 1989) and showed that deletion of P562 results in an enzyme unable to perform RCA (**Figure 5A**). Average extension lengths appear shorter than D12A-THR and no full-length product could be observed, supporting that P562 may be an important contact between thumb and TPR2 subdomains, and essential for TPR2's role in strand displacement and high processivity. In the context of the previous primer extension and EMSA results, this supports a role for P562 in the dissociation rate of the polymerase from its DNA template.

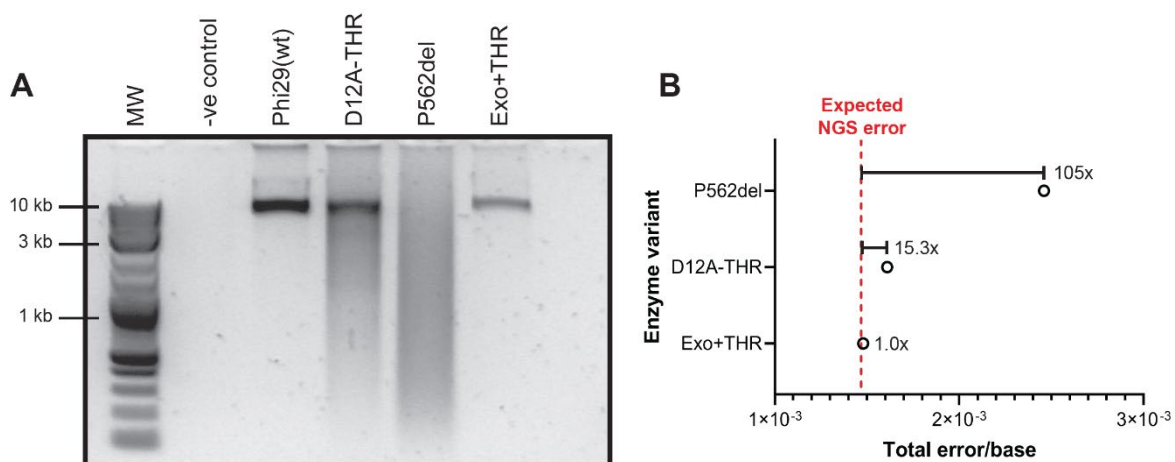


Figure 5. Impaired rolling circle amplification (RCA) activity and reduced fidelity of Phi29 DNAP P562del. (A) The RCA assay used 10 ng of plasmid template and 3 nM of each enzyme in a 3-hour reaction. A sample without enzyme (-ve control) was used to confirm the polymerase-dependent amplification. Restriction of the RCA yields full-length

products. **(B)** Relative fidelity of the D12A-THR and p562del mutants to Exo+THR. The corrected total error rates from **Table 2** of each mutant were divided by that of the Exo+THR, resulting in the relative fidelity scores.

DNA synthesis fidelity was investigated using an adaptation of our previously reported XNA fidelity assay, based on primer extension assays (**Figure S3A**). Given the very low error rates shown by wild-type Phi29 DNA polymerase (in the range between 3×10^{-6} (Nelson et al., 2002) and 9.5×10^{-6} (Paez et al., 2004) errors per base), errors introduced by PCR and deep sequencing steps significantly affect the data obtained. Deep sequencing errors alone are reported around 4.73×10^{-3} for MiSeq Illumina platforms (Stoler & Nekrutenko, 2021).

We opted to use Exo+THR fidelity data to normalize for PCR and deep sequencing errors in the experiment, thus establishing a baseline against which the impact of D12A and P562 could be measured. As expected, disabling the exonuclease activity (D12A-THR) increased the rate of substitutions by approximately an order of magnitude, while showing marginal increases in deletions and insertions. Removal of P562 had a more pronounced effect, with substitution rates approximately 100-fold higher than Exo-THR and significant increases in deletion (7-fold) and insertion (6-fold) frequencies (**Figure 5B** and **Table 2**).

Mutant	Substitution rate ^a	Deletion rate ^a	Insertion rate ^a	Total error rate	R ^b	Corrected total error rate	Total bases
Exo+THR	1.4×10^{-3}	9.0×10^{-5}	4.0×10^{-6}	1.5×10^{-3}	2.6×10^3	9.5×10^{-6}	1,754,723
D12A-THR	1.4×10^{-3}	1.7×10^{-4}	6.1×10^{-6}	1.6×10^{-3}	2.7×10^3	1.5×10^{-4}	1,808,769
P562del	1.8×10^{-3}	6.1×10^{-4}	2.3×10^{-5}	2.5×10^{-3}	2.7×10^3	9.9×10^{-4}	1,834,463

Table 2. Quantification of insertions, deletions, and substitutions rates (errors/bp) of phi29 DNAP mutants. ^aRaw error rates that do not consider PCR and NGS error rates. ^bR: PCR and Error estimate calculated from the number of errors expected to be introduced by the

commercial phi29 DNAP (using the published 9.5×10^{-6} error rate (Paez et al., 2004)) and subtracting this figure from the observed number of errors introduced by Exo+THR. The remainder was used to calculate the expected PCR and NGS error estimate for the total bases sequenced of the D12A-THR and p562del mutants.

Detailed analysis of error identity and position showed context-dependent mutational hot spots for the three tested enzymes but particularly salient for P562del (**Figure S4-S5**). P562del also showed elevated levels of transition substitutions (26.8%) compared to D12A-THR (21.2%) – shown in **Figure S3C**.

The phenotype shown by P562del bear similarities to the previously reported Δ TPR2, in that the variants form a less stable complex with their template, resulting in lower processivity and loss of strand displacement activity. That, however, is not sufficient to explain P562del enrichment in selection and its increased initial rate of HNA synthesis. Given the previous report of P562 interaction with TPR2 (Rodriguez et al., 2005), we postulate that P562 deletion removes a critical contact involved in regulating Phi29 polymerase binding to its template, making the process more dynamic – faster at assembling replication-competent complexes (explaining selection and primer extension), and faster at dissociating from DNA (explaining lower DNA binding stability and poor RCA performance). Dissociation from template becomes even more likely once it becomes an HNA/DNA heteroduplex, as it has been observed for other polymerases (Vastmans et al., 2000), explaining the “early burst” phenotype observed.

Discussion

DNA polymerases are essential for life and, logically, they have been extensively studied in the last 60 years. In addition to biochemical characterisation of natural enzymes, mutants – whether selected or targeted – have also significantly contributed to our current detailed mechanistic understanding of this family of enzymes (Pineiro, 2019).

Nonetheless, that understanding remains incomplete for polymerases. For instance, we still have no mechanistic understanding of how some natural DNA-dependent DNA polymerases display reverse transcriptase activity – such as the DNAP from *Thermus thermophilus* for RNA (Myers & Gelfand, 1991), or the DNAP from *Geobacillus stearothermophilus* for some XNAs (Jackson et al., 2019). We do not know how to explain the different impact of a given modification between different polymerase chassis (Nikoomanzar et al., 2019), and we still compare enzyme activities at standardized conditions, rather than at their individual reaction optimum (Nikoomanzar et al., 2017). Such lack of understanding also explains the current generation of XNA polymerases, which lag severely behind natural DNA polymerases (using natural substrates) on processivity, fidelity and catalytic rates.

Numerous individual residues have been implicated in function, but they cannot act in isolation. How the different residues interact as a system to enable a polymerase to function remains elusive, with only a handful of examples demonstrating that such systems exist (Li et al., 2010).

For less well-characterised enzymes, the prospect of biochemically characterising panels of mutants becomes increasingly difficult, and only accessible by technically challenging high-throughput screening platforms (Gielen et al., 2016). Still, selection platforms have the potential to explore even higher throughputs ($> 10^7$ variants) and at a lower technical entry barrier (e.g. emulsion or *in vivo* platforms).

Selection platforms are at the core of directed evolution, but here the emphasis is changed: selection conditions are chosen not to enhance certain feature of the enzyme but to recover functional enzymes that perform best at the selected conditions. Consequently, there is greater need for robust statistical analysis of selection outputs,

and for a more systematic exploration of how reaction conditions influence selection – areas that increasingly being considered in directed evolution.

We implement an E-test to compare how individual sequences enrich between rounds, approximating individual sequence frequencies to Poisson distribution parameters. It assumes a degree of robustness in selection (so that frequencies can be good population estimates) that is only justifiable for heavily oversampled libraries – here, designed libraries were under 10^4 variants, transformations recovered in excess of 10^6 CFU and CST selections are carried out with over 10^7 cells, fulfilling those criteria. Nonetheless, a natural progression would be to include stochastic measures that assume incomplete sampling, as commonly implemented for aptamer selection (Hoinka et al., 2015), and to extend analysis to include interacting residues (Nikoomanzar et al., 2019; Salinas & Ranganathan, 2018).

Another crucial aspect of this approach is the library design. Substitutions are traditionally used to sample sequence space around wild-type enzymes, but they do not necessarily abolish backbone interactions. The P562 residue is not conserved among homologous Phi29-like enzymes, and a P562C mutation has been shown to have little functional impact in reducing conditions (Rodríguez et al., 2009), suggesting that its interaction with TPR2 and exonuclease domains may be the result of backbone interactions.

Indels, while abundant in nature, are usually less explored as a tool for protein engineering (Tizei et al., 2021; Tóth-Petróczy & Tawfik, 2013), because of their expected impact on protein structure and increased complexity in analysis.

While this work represents a proof-of-principle that selection can be used to gain biological insight into the workings of a well-characterised enzyme, it can be readily implemented to other enzymes and at larger scales, towards reconstructing the functional networks that define a protein function.

MATERIALS AND METHODS

Library design, selection, NGS and analysis leading to the identification of P562del.

The exonuclease, TPR2 and thumb loop InDel libraries were generated through inverse PCR (iPCR) on the pET23-P2-D12A-THR (**Table S1**) with phosphorylated InDel mutagenic primers (Table S2) as previously described (Handal Marquez, 2019) and summarized in **Figure 1B**. Each InDel library was subjected to rounds of selection through CST (Pinheiro et al., 2014) adapted for the mesophilic phi29 DNAP as previously described (Handal Marquez, 2019). Briefly, 1×10^8 cells from each library, post-induction (3 h at 30°C), were resuspended in 100µl of activity reaction mix, composed of 30 pmol of a short biotinylated oligo (CST_04(7)exoR, **Table S2**), 200µM of hNTPs, 1x Phi29 reaction buffer (NEB), 1x Bovine Serum Albumin (BSA, Sigma Aldrich), 1 M betaine (Thermo Scientific), 2µL NotI (NEB), 1 mg/ml lysozyme (Thermo Scientific) and 5µg/ml polymyxin (Sigma-Aldrich) in 100µL molecular grade water. Resuspended cells were emulsified in water-in-oil emulsions and subjected to freezing-thawing cycles for cell lysis and plasmid denaturation. HNA synthesis was then carried out for 10 min at 30°C and products were captured, post-emulsion disruption, via biotin-streptavidin pulldown using paramagnetic beads. Captured products were amplified through PCR using KOD Xtreme Hot Start DNA Polymerase (EMD Millipore), cloned into the pET23-P2-D12A-THR expression backbone, and transformed into fresh electrocompetent *E. coli* T7 Express cells (NEB). The libraries pre- and post-selection were then grown for 3 h at 37°C. Plasmids were then extracted and used as template in ~400 bp NGS amplicon generation PCR reactions using KOD Xtreme Hot Start DNA Polymerase (EMD Millipore). The amplicons were gel-extracted using the Monarch DNA Gel Extraction kits (NEB). NGS-based amplicon sequencing was carried out for all libraries at Genewiz UK Ltd using the Amplicon-EZ service. Sequencing data, which can be found in the NCBI SRA database (BioProject: PRJNA883233), was processed using the *NGS_preprocessing.ga* script (**Supplementary Information S2**) in the Galaxy (Afgan et al., 2016) public server (usegalaxy.org). InDel frequencies before and after selection, enrichment complex scores and the E-test for comparing two Poisson means (Krishnamoorthy & Thomson, 2004) were calculated using the *InDel_Quantification.jl* script (**Supplementary Information**

S3). The calculated scores can be found in the *InDel_EC.xlsx* file (**Supplementary Information S4**).

Generation of phi29 mutants. The P562del and Exo+THR mutants were generated on the pET23-P2-D12A-THR (Table S1) background through inverse Polymerase Chain Reactions (iPCR) using Q5® High-Fidelity DNA Polymerase (NEB) following the manufacturer's instructions in 25 µl reactions with primers p2_thumb_loop_R/DEL1 (Table S2) and iPCR_P2_Exo+_F1/R1 (Table S2), respectively. Amplified products were treated with 0.4U/µl DpnI for 1 h at 37°C prior to the PCR purification with the GeneJET PCR Purification Kit (Thermo Scientific). 100 ng of DNA products were blunt-end ligated in 20 µl reactions with 40U/µl T4 DNA ligase in 1x T4 DNA ligase buffer overnight at room temperature. 5 µl of the ligation products were transformed in NEB® 5-alpha competent *E. coli* cells following the recommended High Efficiency Transformation Protocol (C2987, New England Biolabs) described by the commercial strain provider. Successful cloning was confirmed through Sanger Sequencing in all cases.

Expression, purification and protein quantification of mutants. The pET23 plasmid encoding the phi29 DNA polymerase mutants were transformed into *E. coli* T7 Express cells (NEB). Isolated transformants were grown at 37°C until an OD₆₀₀ of 0.8 was reached. Protein expression was induced by adding IPTG to a final concentration of 1 mM and incubating cultures for 4 h at 30°C. Cells were then pelleted through centrifugation at 5000xg and the supernatant was discarded. Pellets were frozen for an hour at -20°C and resuspended in 4 mL of B-PER Reagent (Thermo Scientific) per gram of cell pellet with 50 mg/mL lysozyme and 250U/uL Benzoase (Sigma Aldrich NV). Cells were subsequently incubated at room temperature for 15 min. Protease inhibitor (1 mM Pefabloc) and reducing agents (1 mM DTT) were added, and cell debris removed through centrifugation at 5000xg. The supernatant was diluted 3-fold in 50 mM Tris-HCl pH 8.0, 1 mM DTT and 1 mM Pefabloc. Expressed protein was purified using the HisPur™ Ni-NTA Resin (Thermo Scientific) using 50 mM Tris-HCl pH 8.0, 1 M NaCl, 1 mM DTT and increasing concentrations of imidazole following the manufacturer's recommendations. Eluted proteins were concentrated, and buffer-exchanged to 10 mM Tris-HCl, 100 mM KCl, 1 mM DTT, 0.1 mM EDTA, pH 7.4 at 25°C using Amicon Ultra-4 Centrifugal 30 kDa

Filter Units (Sigma Aldrich NV). Glycerol (final concentration: 50% v/v), Tween 20 (0.5%), Nonidet P40 (0.5%) were then added for long term storage at -20°C.

To quantify protein concentration, a series of Bovine Serum Albumin (BSA) standards in the 100 µg/ml to 1,200 µg/ml range were run in parallel to the concentrated proteins in an SDS-PAGE gel to establish a BSA standard curve used to determine the concentration of the mutants. For the quantification of the commercial phi29 DNAP, the ML (Modified Lowry) Protein Assay (G-Biosciences) was used following the Microtube ML-Protein Assay Protocol (1.5-2.0ml Assay Tubes).

Primer Extension Assay for HNA synthesis. Reactions were prepared by mixing in 10 µl: 1x NEB 10X Phi29 DNAP buffer, 0.1 mg BSA, 1 M Betaine, 0.2 mM hNTPs, 12 nM phi29 DNAP mutant and 10 pmol of single stranded DNA template, TempN-exoR (**Table S2**) pre-annealed to 1 pmol of fluorescently labelled DNA primer, Tag01F3-exoR (**Table S2**). Pre-annealing was carried out by incubating primer and template at 95°C for 5 min and cooling them down to 4°C at a rate of 0.1°C/sec in a thermocycler. The primer extension assays were incubated at 30°C from 1 min to 16 h (depending on the experiment) but always inactivated by incubating it at 65°C for 20 min. Samples were diluted in an equal volume of 2x loading buffer (98% (v/v) formamide, 10 mM ethylenediaminetetraacetic acid (EDTA), 0.02% (w/v) Orange G) and boiled at 95°C for 5 min. Samples were loaded on 20% TBE-Urea polyacrylamide gels and run in 1X TBE buffer at 25 watts for 2 hours. Gels were imaged using a Typhoon™ FLA 9500 biomolecular imager (GE Healthcare) using recommended filter and detector settings based on the fluorophore being used and analysed with ImageJ (Schneider et al., 2012). Three independent experiments (biological replicates, n = 3) were carried out for the quantification of the primer extension assays.

DNA Binding Assay – Electrophoretic mobility shift assay (EMSA). The EMSA protocol was adapted from (Povilaitis et al., 2016). Oligonucleotide Tag01F_ExoR (**Table S2**) was annealed to TemN-ExoR (**Table S2**) at a 1:1.2 molar ratio in the presence of 0.2 M NaCl and 60 mM Tris-HCl, pH 7.5 by incubating at 95°C for 5 min and cooling down to 4°C at a rate of 0.1°C/sec in a thermocycler. A 50 µl complex-forming reaction was prepared by mixing 33 mM Tris-acetate pH 7.9, 66 mM potassium acetate, 10% glycerol,

0.1 mg/ml BSA with the 2 pmol pre-annealed primer/template substrate and corresponding enzyme at a final 60 pM concentration. Reactions were incubated at 30°C for different time intervals (0, 15, 30 or 60 min). After incubation, samples were mixed with loading dye (10% glycerol and 0.03% orange G) and analysed by electrophoresis in 0.8% (w/v) agarose gel in 0.25x TBE buffer (22.25 mM Tris, 22.25 mM boric acid, and 0.5 mM EDTA pH 8.3). Electrophoresis was performed in 0.25x TBE buffer at room temperature for 2 h at 100 V. Gels were imaged using the Typhoon™ bioimager and quantified using ImageQuant TL 8.2 (Cytiva). Four independent experiments (biological replicates, n = 2) were carried out for the quantification of the EMSA assays.

Rolling Circle Amplification (RCA). 100 pmol of P2_RCA_N8_ExoR (5'-NNNNNN*N*N-3', **Table S2**) with two phosphorothioated DNA bases at the 3' end for exonuclease resistance was pre-annealed to 10 ng of pET23_P2_D12A_THR (**Table S1**) plasmid in 1x NEB 10X Phi29 DNAP buffer, per 100 µl RCA reaction, 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 thermocycler. The pre-annealed template mix was supplemented with 0.2 mM dNTPs, 0.2 mg/mL BSA, 5 ng ET SSB (NEB), 4 nM phi29 DNAP in a 100 µl reaction and incubated for 3 h at 30°C. The RCA products were digested with 10U of NotI in 1x Cutsmart buffer for 1 h at 37°C. The digested RCA products were loaded on an agarose gel for visualisation.

Isothermal polymerase fidelity assay. Single-stranded template was generated by mixing 1.8 pmol of pET23_KOD_DA_Mut (**Table S1**) with 0.7U/µl Nb.BbvCI, 10U/µl of ExoIII and 1x Cutsmart buffer in a 30 µl reaction. 10 pmol of PH_pET23_DA_Biotin3 (**Table S2**) biotinylated oligonucleotide (4.9 µl) was pre-annealed to 0.2 pmol of single stranded template by incubating the mixture at 95°C for 5 min and cooling them down to 4°C at a rate of 0.1°C/sec in a thermocycler. The annealed primer-template was made up to 50 µl with deionized distilled water (ddH₂O). 5 µl of Dyanabeads MyOne C1 (ThermoFisher), per reaction, were washed twice with 100 µl of 2x BWBS-T (20 mM Tris-HCl pH 7.4, 2 M NaCl, 0.2% v/v Tween20, 2 mM EDTA) and blocked in 1 mL 2x BWBS-T on a rotator for 1 h at room temperature. Beads were captured, resuspended in 50 µl 2x BWBS-T, and mixed with the 50 µl reaction prior to being incubated for 3 h on a rotator at room temperature. Beads were washed twice with 200 µl 30 mM NaOH at 37°C, then

with 200 μ l EB-T (10 mM Tris-HCl pH 8.8, 0.1 mM EDTA, 0.01% Tween20). Beads were resuspended in 10 μ l elution buffer (10 mM Tris-HCl, pH 8.5). Amplicons for deep sequencing were generated using KOD Xtreme™ Hot Start DNA Polymerase (EMD Millipore) following the manufacturer's instructions in 50 μ l reactions with 1 μ l of the resuspended beads as template, and *outnest_1* (**Table S2**) and *P2_fidelity_inestR1* (**Table S2**) as primers. The cycling parameters used were 2 min at 95°C, followed by 20 cycles of 15 s at 98°C, 15 s at 65°C and 10 s at 68°C, with a final polishing step of 2 min at 72°C. Primers were degraded by adding 0.8U/ μ l Exonuclease I to each 50 μ l PCR reaction.

Next Generation Sequencing of fidelity assay data and analysis. The amplicons described above were purified with the Monarch PCR and DNA purification kit (New England Biolabs) and sent for NGS EZ Amplicon Sequencing by Genewiz. NGS data was pre-processed in the Galaxy public server (usegalaxy.org) using the *NGS_preprocessing.ga* script (**Supplementary information S2**) and trimmed using part 1 of the Julia (v1.7) *Fidelity_Quantification.jl* script (**Supplementary Information S5**). The reference sequence, 'Fidelity_ref' (**Table S2**), was added to the trimmed reads and the reads were then aligned using the random chain algorithm of MAFFT version 7 (Katoch & Standley, 2013) on the MAFFT server (mafft.cbrc.jp/alignment/server/large.html) with the following parameters: *mafft --thread 8 --threadtb 5 --threadit 0 --inputorder --randomchain input > output*. The number of insertions, deletions, and substitutions as well as the number of individual error types were calculated by comparing in a per-base manner the aligned reads to the reference 'Fidelity_ref' (**Table S2**) sequence within the alignment using part 2 of the *Fidelity_Quantification.jl* script (**Supplementary Information S5**). Raw error rates were obtained by dividing the sum of the errors by the total number of bases sequenced. The PCR and NGS error (R) estimate for Exo+THR was obtained by subtracting the expected number of errors of the commercial phi29 DNAP (using the published error rate 9.5×10^{-6}) from the observed number of errors by Exo+THR as follows:

$$R_{Exo+THR} = (Error\ rate_{Exo+THR} \times Total\ bases_{Exo+THR}) - (Error\ rate_{phi29} \times Total\ bases_{Exo+THR})$$

(1)

The $Err_{Exo+THR}$ score was then used to calculate the PCR and NGS error of the D12A-THR and p562del mutants as follows:

$$R_{mutant} = \frac{R_{Exo+THR}}{Total\ base_{Exo+THR} \times Total\ base_{mutant}} \quad (2)$$

The corrected total error rates were then calculated by dividing the corrected error number over the total bases sequenced from each variant.

Entropy scores were calculated using part 2 of the *Fidelity_Quantification.jl* script (**Supplementary Information S4**) using the following equation:

$$H(l) = -\sum f(b, l) \log_2 f(b, l) \quad (3)$$

where $H(l)$ is the entropy at position l , b is a residue or gap and $f(b, l)$ is the frequency at which residue b is found at position l .

The location (start position, end position and length) and frequency of each Indel across the MSA were obtained using part2 of the *Fidelity_Quantification.jl* script (**Supplementary Information S5**), by comparing each sequence to the 'Fidelity_ref' sequence within the alignment. The frequency values were divided by the total number of insertions or deletions and multiplied by 100 to obtain a relative frequency distribution. Only sequences >5% relative frequency were plotted.

AVAILABILITY

All raw and processed data used in this manuscript are available in our GitHub repository (<https://github.com/PinheiroLab/>).

ACCESSION NUMBERS

Next generation sequencing data has been deposited on NCBI SRA under the following accession number: PRJNA883233.

SUPPLEMENTARY DATA

SI_Phi29_HNA_synthesis.docx (**Supplementary Information**)

NGS_preprocessing.ga script (**Supplementary Information S2**)

InDel_Quantification.jl (**Supplementary Information S3**)

InDel_EC.xlsx (**Supplementary Information S4**)

Fidelity_Quantification.jl (**Supplementary Information S5**)

ACKNOWLEDGEMENTS

VBP, LLT and PHM thank ERASynBio (grant BB/N01023X/1; *invivoXNA*). VBP and PHM thank FWO (grant G0H7618N). PHM thanks FWO (studentship 3M180645).

CONTRIBUTIONS

VBP, LLT and PHM contributed to the project design. PHM performed all experimental work. VBP and PHM carried out NGS analysis.

REFERENCES

Afgan, E., Baker, D., van den Beek, M., Blankenberg, D., Bouvier, D., Čech, M., Chilton, J., Clements, D., Coraor, N., Eberhard, C., Grüning, B., Guerler, A., Hillman-Jackson, J., Kuster, G. von, Rasche, E., Soranzo, N., Turaga, N., Taylor, J., Nekrutenko, A., & Goecks, J. (2016). The Galaxy platform for accessible,

- reproducible and collaborative biomedical analyses: 2016 update. *Nucleic Acids Research*, 44(W1), W3–W10. <https://doi.org/10.1093/nar/gkw343>
- Bebenek, A., Dressman, H. K., Carver, G. T., Ng, S. S., Petrov, V., Yang, G., Konigsberg, W. H., Karam, J. D., & Drake, J. W. (2001). Interacting Fidelity Defects in the Replicative DNA Polymerase of Bacteriophage RB69. *Journal of Biological Chemistry*, 276(13), 10387–10397. <https://doi.org/10.1074/JBC.M007707200>
- Bernad, A., Blanco, L., Lázaro, J. M., Martín, G., & Salas, M. (1989). A conserved 3'→5' exonuclease active site in prokaryotic and eukaryotic DNA polymerases. *Cell*, 59(1), 219–228. [https://doi.org/10.1016/0092-8674\(89\)90883-0](https://doi.org/10.1016/0092-8674(89)90883-0)
- Chica, R. A., Doucet, N., & Pelletier, J. N. (2005). Semi-rational approaches to engineering enzyme activity : combining the benefits of directed evolution and rational design. *Current Opinion in Biotechnology*, 16, 378–384. <https://doi.org/10.1016/j.copbio.2005.06.004>
- Cozens, C., Pinheiro, V. B., Vaisman, A., Woodgate, R., & Holliger, P. (2012). A short adaptive path from DNA to RNA polymerases. *Proceedings of the National Academy of Sciences*, 109(21), 8067–8072. <https://doi.org/10.1073/pnas.1120964109>
- Freund, N., Taylor, A. I., Arangundy-Franklin, S., Subramanian, N., Peak-Chew, S. Y., Whitaker, A. M., Freudenthal, B. D., Abramov, M., Herdewijn, P., & Holliger, P. (2022). A two-residue nascent-strand steric gate controls synthesis of 2'-O-methyl- and 2'-O-(2-methoxyethyl)-RNA. *Nature Chemistry*. <https://doi.org/10.1038/s41557-022-01050-8>
- Frey, M. W., Nossal, N. G., Capson, T. L., & Benkovic, S. J. (1993). Construction and characterization of a bacteriophage T4 DNA polymerase deficient in 3' → 5' exonuclease activity. *Proceedings of the National Academy of Sciences of the United States of America*, 90(7), 2579–2583. <https://doi.org/10.1073/PNAS.90.7.2579>

- Garmendia, C., Bernad, A., Esteban, J. A., Blanco, L., & Salas, M. (1992). The bacteriophage ϕ 29 DNA polymerase, a proofreading enzyme. *Journal of Biological Chemistry*, 267(4), 2594–2599. <https://doi.org/10.13039/100000002>
- Ghadessy, F. J., Ong, J. L., & Holliger, P. (2001). Directed evolution of polymerase function by compartmentalized self-replication. *Proceedings of the National Academy of Sciences*, 98(8), 4552–4557. <https://doi.org/10.1073/pnas.071052198>
- Gielen, F., Hours, R., Emond, S., Fischlechner, M., Schell, U., & Hollfelder, F. (2016). Ultrahigh-throughput-directed enzyme evolution by absorbance-activated droplet sorting (AADS). *Proceedings of the National Academy of Sciences of the United States of America*, 113(47), E7383–E7389. <https://doi.org/10.1073/pnas.1606927113>
- Handal Marquez, P. (2019). Sampling the Functional Sequence Neighbourhood of Phi29 DNA Polymerase for XNA Synthesis. (MPhil) UCL (University College London).
- Hoinka, J., Berezhnoy, A., Dao, P., Sauna, Z. E., Gilboa, E., & Przytycka, T. M. (2015). Large scale analysis of the mutational landscape in HT-SELEX improves aptamer discovery. *Nucleic Acids Research*, 43(12), 5699–5707. <https://doi.org/10.1093/nar/gkv308>
- Houlihan, G., Arangundy-Franklin, S., Porebski, B. T., Subramanian, N., Taylor, A. I., & Holliger, P. (2020). Discovery and evolution of RNA and XNA reverse transcriptase function and fidelity. *Nature Chemistry*, 12(August), 1–8. <https://doi.org/10.1038/s41557-020-0502-8>
- Jackson, L. N., Chim, N., Shi, C., & Chaput, J. C. (2019). Crystal structures of a natural DNA polymerase that functions as an XNA reverse transcriptase. *Nucleic Acids Research*, 47(13), 6973–6983. <https://doi.org/10.1093/nar/gkz513>
- Kamtekar, S., Berman, A. J., Wang, J., Lázaro, J. M., de Vega, M., Blanco, L., Salas, M., & Steitz, T. A. (2004). Insights into strand displacement and processivity from

- the crystal structure of the protein-primed DNA polymerase of bacteriophage ϕ 29. *Molecular Cell*, 16(4), 609–618. <https://doi.org/10.1016/j.molcel.2004.10.019>
- Katoh, K., & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: Improvements in performance and usability. *Molecular Biology and Evolution*, 30(4), 772–780. <https://doi.org/10.1093/molbev/mst010>
- Khare, V., & Eckert, K. A. (2002). The proofreading 3' \rightarrow 5' exonuclease activity of DNA polymerases: A kinetic barrier to translesion DNA synthesis. *Mutation Research - Fundamental and Molecular Mechanisms of Mutagenesis*, 510(1–2), 45–54. [https://doi.org/10.1016/S0027-5107\(02\)00251-8](https://doi.org/10.1016/S0027-5107(02)00251-8)
- Krishnamoorthy, K., & Thomson, J. (2004). A more powerful test for comparing two Poisson means. *Journal of Statistical Planning and Inference*, 119(1), 23–35. [https://doi.org/10.1016/S0378-3758\(02\)00408-1](https://doi.org/10.1016/S0378-3758(02)00408-1)
- Laos, R., Thomson, J. M., & Benner, S. A. (2014). DNA polymerases engineered by directed evolution to incorporate nonstandard nucleotides. *Frontiers in Microbiology*, 5(OCT), 1–14. <https://doi.org/10.3389/fmicb.2014.00565>
- Lázaro, J. M., Blanco, L., & Salas, M. (1995). Purification of Bacteriophage ϕ 29 DNA Polymerase. *Methods in Enzymology*, 262(C), 42–49. [https://doi.org/10.1016/0076-6879\(95\)62007-9](https://doi.org/10.1016/0076-6879(95)62007-9)
- Li, V., Hogg, M., & J, R.-K. L. (2010). Identification of a New Motif in Family B DNA Polymerases by Mutational Analyses of the Bacteriophage T4 DNA Polymerase. *Journal of Molecular Biology*, 400(3), 295–308. <https://doi.org/10.1016/j.jmb.2010.05.030>.Identification
- Lieberman, K. R., Cherf, G. M., Doody, M. J., Olasagasti, F., Kolodji, Y., & Akeson, M. (2010). Processive replication of single DNA molecules in a nanopore catalyzed by phi29 DNA polymerase. *Journal of the American Chemical Society*, 132(50), 17961–17972. <https://doi.org/10.1021/ja1087612>

- Lutz, S. (2011). *Beyond directed evolution - semi-rational protein engineering and design*. 21(6), 734–743. <https://doi.org/10.1016/j.copbio.2010.08.011>. Beyond
- Lutz, S., Burgstaller, P., & Benner, S. A. (1999). An in vitro screening technique for DNA polymerases that can incorporate modified nucleotides. Pseudothymidine as a substrate for thermostable polymerases. *Nucleic Acids Research*, 27(13), 2792–2798. <https://doi.org/10.1093/nar/27.13.2792>
- Myers, T. W., & Gelfand, D. H. (1991). Reverse Transcription and DNA Amplification by a *Thermus thermophilus* DNA Polymerase. *Biochemistry*, 30(31), 7661–7666. <https://doi.org/10.1021/bi00245a001>
- Nelson, J. R., Cai, Y. C., Giesler, T. L., Farchaus, J. W., Sundaram, S. T., Ortiz-Rivera, M., Hosta, L. P., Hewitt, P. L., Mamone, J. A., Palaniappan, C., & Fuller, C. W. (2002). TempliPhi, ϕ 29 DNA polymerase based rolling circle amplification of templates for DNA sequencing. *BioTechniques, Suppl*, 44–47. <https://doi.org/10.2144/jun0208>
- Nikoomanzar, A., Dunn, M. R., & Chaput, J. C. (2017). Evaluating the Rate and Substrate Specificity of Laboratory Evolved XNA Polymerases. *Analytical Chemistry*, 89(23), 12622–12625. <https://doi.org/10.1021/acs.analchem.7b03807>
- Nikoomanzar, A., Vallejo, D., & Chaput, J. C. (2019). Elucidating the Determinants of Polymerase Specificity by Microfluidic-Based Deep Mutational Scanning. *ACS Synthetic Biology*, 8(6), 1421–1429. <https://doi.org/10.1021/acssynbio.9b00104>
- Paez, J. G., Lin, M., Beroukhim, R., Lee, J. C., Zhao, X., Richter, D. J., Gabriel, S., Herman, P., Sasaki, H., Altshuler, D., Li, C., Meyerson, M., & Sellers, W. R. (2004). Genome coverage and sequence fidelity of phi29 polymerase-based multiple strand displacement whole genome amplification. *Nucleic Acids Research*, 32(9), 1–11. <https://doi.org/10.1093/nar/gnh069>
- Patel, P. H., & Loeb, L. A. (2000). Multiple amino acid substitutions allow DNA polymerases to synthesize RNA. *Journal of Biological Chemistry*, 275(51), 40266–40272. <https://doi.org/10.1074/jbc.M005757200>

- Pinheiro, V. B. (2019). Engineering-driven biological insights into DNA polymerase mechanism. *Current Opinion in Biotechnology*, 60, 9–16.
<https://doi.org/10.1016/j.copbio.2018.11.008>
- Pinheiro, V. B., Arangundy-franklin, S., & Holliger, P. (2014). Compartmentalized Self-Tagging for In Vitro-Directed Evolution of XNA Polymerases. *Current Protocols in Nucleic Acid Chemistry*, 57, 9.9.1-9.9.18.
<https://doi.org/10.1002/0471142700.nc0909s57>
- Pinheiro, V. B., Taylor, A. I., Cozens, C., Abramov, M., Renders, M., Zhang, S., Chaput, J. C., Wengel, J., Peak-Chew, S. Y., McLaughlin, S. H., Herdewijn, P., & Holliger, P. (2012). Synthetic genetic polymers capable of heredity and evolution. *Science*, 336(6079), 341–344. <https://doi.org/10.1126/science.1217622>
- Povilaitis, T., Alzbutas, G., Sukackaite, R., Siurkus, J., & Skirgaila, R. (2016). In vitro evolution of phi29 DNA polymerase using isothermal compartmentalized self replication technique. *Protein Engineering, Design and Selection*, 29(12), 617–628.
<https://doi.org/10.1093/protein/gzw052>
- Rodriguez, I., Lazaro, J. M., Blanco, L., Kamtekar, S., Berman, A. J., Wang, J., Steitz, T. A., Salas, M., & Vega, M. de. (2005). A specific subdomain in ϕ 29 DNA polymerase confers both processivity and strand-displacement capacity. *Proceedings of the National Academy of Sciences*, 102(18), 6407–6412.
- Rodríguez, I., Lázaro, J. M., Salas, M., & de Vega, M. (2009). Involvement of the TPR2 subdomain movement in the activities of ϕ 29 DNA polymerase. *Nucleic Acids Research*, 37(1), 193–203. <https://doi.org/10.1093/nar/gkn928>
- Salinas, V. H., & Ranganathan, R. (2018). Coevolution-based inference of amino acid interactions underlying protein function. *ELife*, 7, 1–20.
- Sampling the Functional Sequence Neighbourhood of Phi29 DNA Polymerase for XNA Synthesis - UCL Discovery*. (n.d.). Retrieved June 2, 2022, from <https://discovery.ucl.ac.uk/id/eprint/10072479/>

- Schneider, C. A., Rasband, W. S., & Eliceiri, K. W. (2012). NIH Image to ImageJ: 25 years of image analysis. *Nature Methods*, *9*(7), 671–675.
<https://doi.org/10.1038/nmeth.2089>
- Stoler, N., & Nekrutenko, A. (2021). Sequencing error profiles of Illumina sequencing instruments. *NAR Genomics and Bioinformatics*, *3*(1), 1–9.
<https://doi.org/10.1093/NARGAB/LQAB019>
- Tizei, P. A. G., Harris, E., Withanage, S., Renders, M., & Pinheiro, V. B. (2021). A novel framework for engineering protein loops exploring length and compositional variation. *Scientific Reports*, *11*(1), 1–13. <https://doi.org/10.1038/s41598-021-88708-4>
- Tóth-Petróczy, Á., & Tawfik, D. S. (2013). Protein insertions and deletions enabled by neutral roaming in sequence space. *Molecular Biology and Evolution*, *30*(4), 761–771. <https://doi.org/10.1093/molbev/mst003>
- Tran, N. Q., Lee, S. J., Akabayov, B., Johnson, D. E., & Richardson, C. C. (2012). Thioredoxin, the processivity factor, sequesters an exposed cysteine in the thumb domain of bacteriophage T7 DNA polymerase. *Journal of Biological Chemistry*, *287*(47), 39732–39741. <https://doi.org/10.1074/jbc.M112.409235>
- Truniger, V., Lazaro, J. M., & Salas, M. (2004). Function of the C-terminus of phi29 DNA polymerase in DNA and terminal protein binding. *Nucleic Acid Therapeutics*, *32*(1), 361–370. <https://doi.org/10.1093/nar/gkh184>
- Tsai, Y. C., & Johnson, K. A. (2006). A new paradigm for DNA polymerase specificity. *Biochemistry*, *45*(32), 9675–9687. <https://doi.org/10.1021/bi060993z>
- Vanmeert, M., Razzokov, J., Mirza, M. U., Weeks, S. D., Schepers, G., Bogaerts, A., Rozenski, J., Froeyen, M., Herdewijn, P., Pinheiro, V. B., & Lescrinier, E. (2019). Rational design of an XNA ligase through docking of unbound nucleic acids to toroidal proteins. *Nucleic Acids Research*, *47*(13), 7130–7142.
<https://doi.org/10.1093/nar/gkz551>

- Vastmans, K., Pochet, S., Peys, A., Kerremans, L., van Aerschot, A., Hendrix, C., Marliere, P., & Herdewijn, P. (2000). Enzymatic incorporation in DNA of 1,5-anhydrohexitol nucleotides. *Biochemistry*, 39(42), 12757–12765.
<https://doi.org/10.1021/bi001297g>
- Wang, W., Wu, E. Y., Hellinga, H. W., & Beese, L. S. (2012). Structural factors that determine selectivity of a high fidelity DNA polymerase for deoxy-, dideoxy-, and ribonucleotides. *Journal of Biological Chemistry*, 287(34), 28215–28226.
<https://doi.org/10.1074/jbc.M112.366609>
- Waters, L. S., Minesinger, B. K., Wiltrout, M. E., D'Souza, S., Woodruff, R. v., & Walker, G. C. (2009). Eukaryotic Translesion Polymerases and Their Roles and Regulation in DNA Damage Tolerance. *Microbiology and Molecular Biology Reviews*, 73(1), 134–154. <https://doi.org/10.1128/mnbr.00034-08>
- Wrenbeck, E. E., Faber, M. S., & Whitehead, T. A. (2017). Deep sequencing methods for protein engineering and design. *Current Opinion in Structural Biology*, 45, 36–44. <https://doi.org/10.1016/j.sbi.2016.11.001>
- Xia, G., Chen, L., Sera, T., Fa, M., Schultz, P. G., & Romesberg, F. E. (2002). Directed evolution of novel polymerase activities: Mutation of a DNA polymerase into an efficient RNA polymerase. *Proceedings of the National Academy of Sciences of the United States of America*, 99(10), 6597–6602.
<https://doi.org/10.1073/pnas.102577799>