ACS chemical biology

Articles

pubs.acs.org/acschemicalbiology

Improving Polymerase Activity with Unnatural Substrates by Sampling Mutations in Homologous Protein Architectures

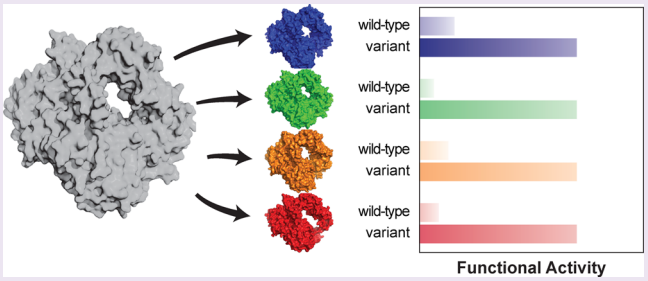
Matthew R. Dunn,[†,§,∥] Carine Otto,[§] Kathryn E. Fenton,[§] and John C. Chaput*[,‡,§,∥,⊥]

[†]School of Life Sciences, [‡]Department of Chemistry and Biochemistry, and [§]The Biodesign Institute at Arizona State University, Tempe, Arizona 85287-5301, United States

[∥]Department of Pharmaceutical Sciences, University of California, Irvine, California 92697-3958, United States

Supporting Information

ABSTRACT: The ability to synthesize and propagate genetic information encoded in the framework of xeno-nucleic acid (XNA) polymers would inform a wide range of topics from the origins of life to synthetic biology. While directed evolution has produced examples of engineered polymerases that can accept XNA substrates, these enzymes function with reduced activity relative to their natural counterparts. Here, we describe a biochemical strategy that enables the discovery of engineered polymerases with improved activity for a given unnatural polymerase function. Our approach involves identifying specificity determining residues (SDRs) that control polymerase activity, screening mutations at SDR positions in a model polymerase scaffold, and assaying key gain-of-function mutations in orthologous protein architectures. By transferring beneficial mutations between homologous protein structures, we show that new polymerases can be identified that function with superior activity relative to their starting donor scaffold. This concept, which we call scaffold sampling, was used to generate engineered DNA polymerases that can faithfully synthesize RNA and TNA (threose nucleic acid), respectively, on a DNA template with high primer-extension efficiency and low template sequence bias. We suggest that the ability to combine phenotypes from different donor and recipient scaffolds provides a new paradigm in polymerase engineering where natural structural diversity can be used to refine the catalytic activity of synthetic enzymes.



The ability to encode and decode genetic information in XNA polymers is a major goal of synthetic biology and an effort that would improve our understanding of why nature chose DNA as the molecular basis of life's genetic material.[1,2] In principle, information transfer could be accomplished by nonenzymatic template-directed polymerization in which XNA building blocks assembled on a DNA template are coupled together using purely chemical methods. Over the past few decades, this approach has been used to "transcribe" natural DNA or RNA templates into a wide range of biopolymer analogues,[3] including systems with modified bases and backbones,[4−8] as well as other systems with diverse chemical functionality.[9] In a dramatic recent advance, Liu and co-workers extended this concept to include the synthesis, selection, and amplification of peptide nucleic acids (PNA) from a library of 10^8 different PNA molecules.[10] This proof-of-principle demonstration, which showed how PNA could be made to evolve *in vitro*, was later expanded to include sequence-defined polymers with chemical compositions that are structurally unrelated to DNA.[11] However, despite significant progress, enzyme-free polymerization methods remain a challenging strategy for XNA synthesis due to the reduced coupling efficiency observed for monomeric and short polymeric units.

Enzyme-mediated strategies that mimic certain biosynthetic pathways found in nature provide an alternative approach to generating synthetic polymers. DNA replication and RNA transcription, for example, use polymerases to synthesize information-carrying polymers that maintain the flow of genetic information in biological systems. Unfortunately, natural polymerases represent an ancient class of enzymes that have evolved over billions of years to maintain cellular integrity by excluding damaged and noncognate substrates from the genome.[12] This high level of substrate specificity provides a formidable barrier to the synthesis of XNA polymers with diverse backbone architectures.

Fortunately, recent advances in polymerase engineering are making it possible to synthesize nucleic acid polymers with modified bases and sugars.[13,14] Several laboratories have identified variants of natural DNA polymerases that expand the genetic alphabet by incorporating unnatural bases alongside the naturally occurring bases of A, C, G, and T (U).[15−17] In the area of modified backbones, our laboratory recently identified a polymerase pair that will convert genetic information back and forth between TNA and DNA.[18] Similar efforts by Holliger and colleagues have led to the discovery of polymerases that can code and decode artificial genetic polymers composed of TNA,

Received: November 17, 2015
Accepted: February 10, 2016
Published: February 10, 2016

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

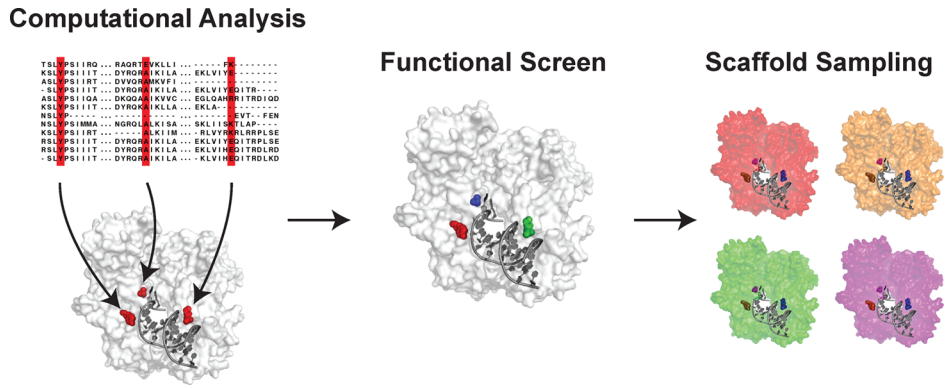


Figure 1. Polymerase engineering strategy. Overview of the process of identifying polymerases with enhanced activity for incorporating non-natural nucleotides. The first step is to identify specificity determining residues (SDRs) within a set of related polymerases. Briefly, a computational analysis is used to identify amino acid positions that are conserved in sequence and structure and located within 10 Å of the DNA primer−template complex. Positions that have been described previously in the literature are given higher priority in functional assays. Functional screens are performed in a model polymerase scaffold to identify gain-of-function mutations at SDR positions. Mutations with the strongest activity are then examined in homologues for properties that are not present in the starting donor scaffold.

hexitol nucleic acid (HNA), locked nucleic acid (LNA), cyclohexyl nucleic acid (CeNA), arabinonucleic acid (ANA), and 2′-fluoro-arabino nucleic acid (FANA).[19] Some of these enzymes have been used to evolve functional XNA molecules with ligand binding and catalytic activity.[19−23]

Although enzyme-mediated examples provide the foundation for storing and accessing information in sequence-defined polymers, all of the XNA polymerases developed to date function with reduced activity relative to their natural counterparts. This constraint limits the types of experiments that can be performed with XNA, as efficient coding and decoding of genetic information is necessary for many synthetic biology projects, like *in vitro* selection and information storage.[1] To overcome this barrier, we have developed a protein design strategy that was used to refine the activity of two different classes of engineered polymerases. Our methodology relies on the natural structural diversity of orthologous protein architectures to improve or fine-tune the activity of a strong gain-of-function mutation or set of mutations. By transferring beneficial mutations between homologous protein scaffolds, we found that it is possible to generate new polymerases that function with superior activity relative to their starting donor scaffold. In particular, we show that scaffold sampling can be used to identify engineered DNA polymerases that can faithfully synthesize RNA and TNA polymers, respectively, with high primer-extension efficiency and low template sequence bias.

RESULTS AND DISCUSSION

Specificity Determining Residues. Previous observations of the low fidelity of TNA synthesis by an engineered DNA polymerase stimulated us to explore the structural factors that govern polymerase activity.[18,24,25] Recognizing that compelling structural and phylogenetic evidence exists to support an adaptive pathway in the evolution of DNA and RNA polymerases,[26] we hypothesized that substrate specificity could be defined by a minimum set of specificity determining residues (SDRs). This hypothesis suggests that engineered polymerases could be developed with greater efficiency by targeting the subset of amino acid positions that control substrate recognition rather than the more typical approach of targeting positions with high phylogenetic variability.[27,28] While a similar approach has been developed for kinases,[29] this strategy has not yet been applied to polymerases or widely used.

Scaffold Sampling. We further postulated that orthologs—genes from different species that evolved from a common ancestor and share the same function—provided a possible strategy for expanding the functional properties of engineered polymerases. Because of subtle differences in protein folding and dynamics, we predicted that gain-of-function mutations discovered in one polymerase would exhibit different phenotypic properties when transferred to a homologous protein scaffold. We call this approach scaffold sampling as it allows newly discovered mutations to sample slightly different regions of protein fold space where unique dynamic environments could lead to beneficial changes in catalytic activity.

Although it is known that mutations can be transferred between orthologs,[30] efforts to harness the natural structural diversity of known protein folds remain limited. Early work by Stemmer *et al.* and others found that DNA shuffling is an effective method for enzyme engineering; however, this strategy requires iterative rounds of directed evolution that are time-consuming and unlikely to distinguish molecules with similar phenotypes.[31] By contrast, scaffold sampling is an optimization strategy that can be used to refine the activity of an engineered polymerase by carefully examining the functional activity of a defined mutation that has been introduced into a set of related polymerase architectures. In this regard, scaffold sampling can be thought of as a fine-grain optimization step that occurs once a gain-of-function mutation has been discovered.

Scaffold Sampling Design Strategy. The combination of both ideas, SDR analysis and scaffold sampling, led us to derive a biochemical strategy that could be used to optimize polymerase functions for a given unnatural substrate (Figure 1). For this study, we chose family B DNA polymerases as a model system because (i) a number of engineered variants have been developed to function with enhanced activity for unnatural substrates, including TNA triphosphates (tNTPs; Supporting Information Table 1);[19,25,32−34] (ii) family B polymerases are found throughout the tree of life, including Archaea, Eubacteria, and Eukarya, and in the viruses that infect these organisms, indicating that substantial sequence diversity exists between related members; and (iii) the three-dimensional

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

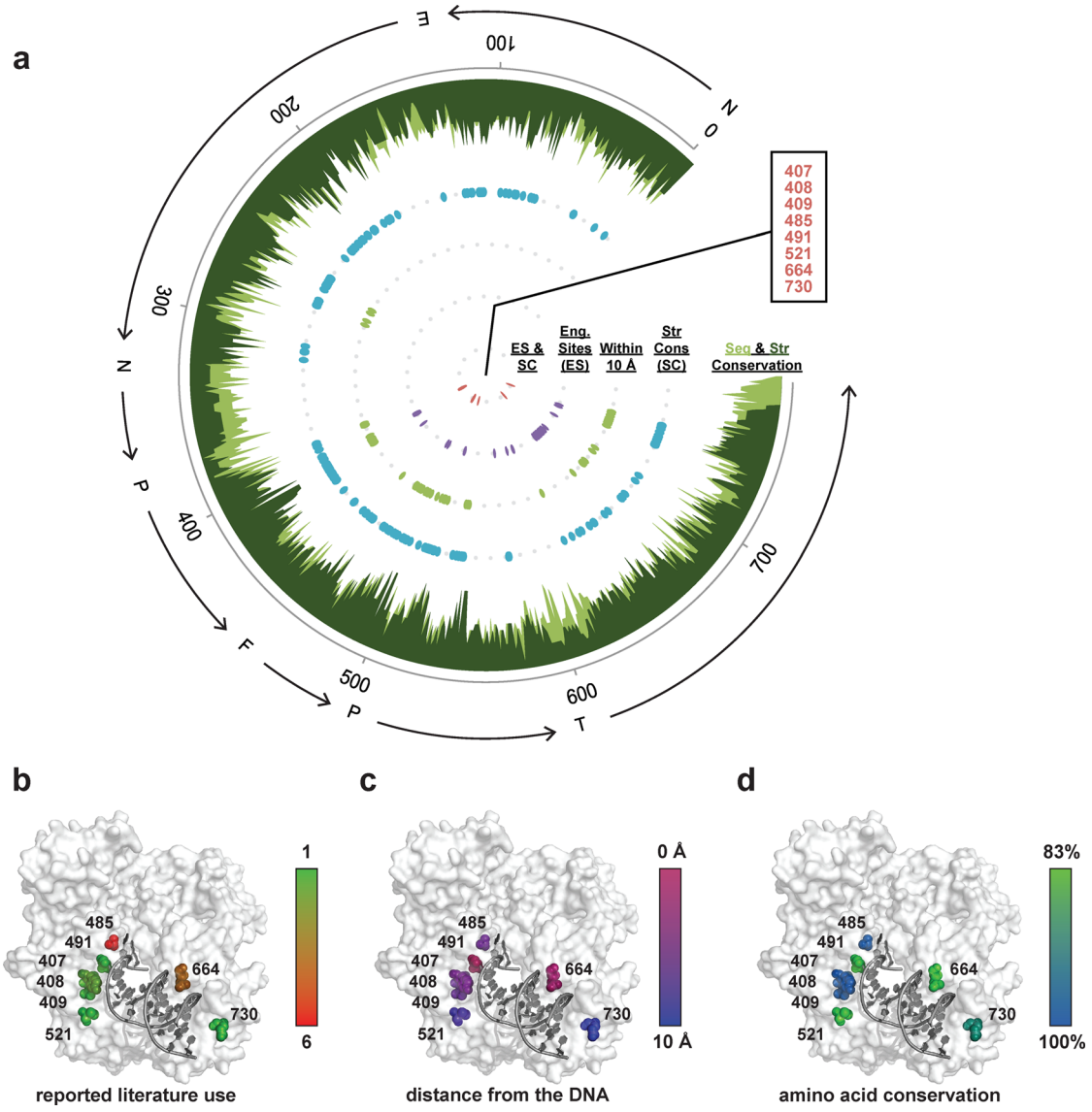


Figure 2. Specificity determining residues in family B DNA polymerases. (a) Circular diagram showing the sequence (light green) and structural (dark green) conservation of representative family B polymerases. Inner rings depict amino acid positions that are conserved in both sequence and structure (blue dots), the subset of conserved sites that lie within 10 Å of the DNA primer−template complex (green dots), and amino acid positions with reported substrate modifying activity (purple dots). Sites where computational and empirical data overlap (red dots) represent the top eight specificity determining residues (SDRs). Domain abbreviations: N, N-terminus; E, exonuclease; P, palm; F, finger; and T, thumb. (b−d) Heat maps showing the eight predicted SDRs mapped onto the surface of 9n DNA polymerase (PDB: 4K8X). Residues are colored by their frequency of reported literature use (b), distance from the primer-template complex (c), and evolutionary conservation (d).

(3D) coordinates of several high resolution crystal structures are available in the protein databank (PDB), which makes it possible to perform structural comparisons between polymerases from different clades.[35]

We used a computational procedure to identify putative SDRs in a representative set of family B polymerases.[29] Comparative phylogenetic and structural data allowed us to identify 144 amino acid positions that are conserved in both sequence and structure, 50 of which lie within 10 Å of the primer-template complex (Figure 2a). We considered these 50 sites to be putative SDRs due to their close proximity to the enzyme active site.[36] A thorough review of the literature revealed that eight of these sites overlapped with positions known to impact the recognition of unnatural substrates and were therefore given a higher priority in functional assays (Figure 2b and Supporting Information Table 1). Three of the eight sites, residues 409, 491, and 664, form primary contacts to the primer−template complex, while other positions are more distal and could form critical second-shell contacts that help define the shape and electrostatic potential of the enzyme active site (Figure 2c). Relative to most other positions in the polymerase family, these sites are highly conserved at the amino acid level (>83% vs 55% sequence identity; Figure 2d, Supporting Information Figure 1).

RNA Synthesis as a Model for Scaffold Sampling. To test the hypothesis that scaffold sampling could be used to identify functional differences between homologous protein scaffolds, we searched the literature for examples of engineered polymerases that contained one or more of the top eight SDRs predicted by our computational analysis. We focused our search on DNA polymerases that have been developed to synthesize RNA, which has long been a model system for polymerase

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

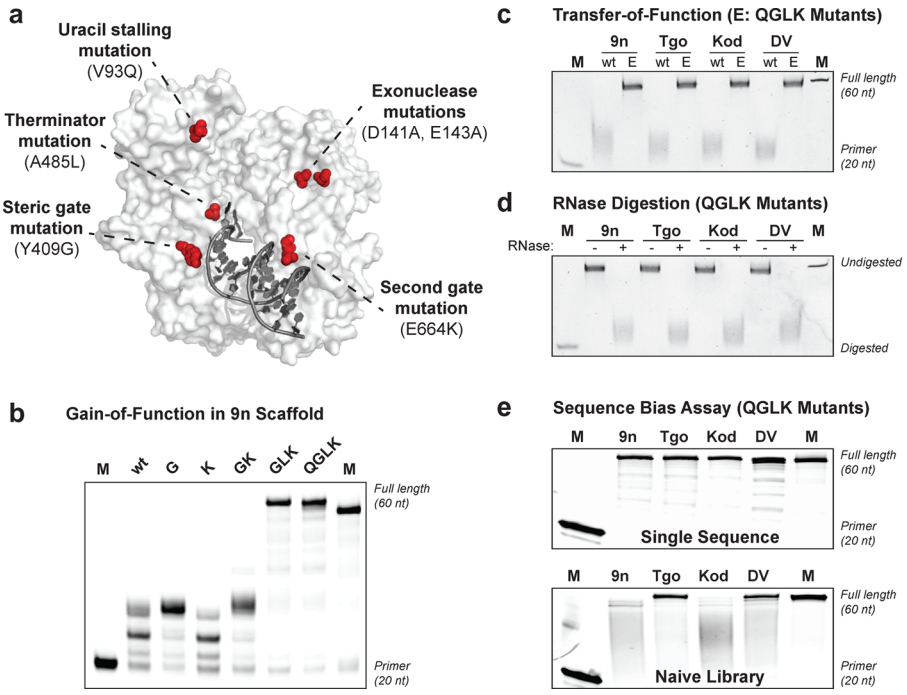


Figure 3. Scaffold sampling reveals functional differences between homologous polymerase scaffolds. (a) Mutations required for RNA polymerase activity in Tgo DNA polymerase mapped onto the structure of 9n DNA polymerase (PDB: 4K8X). Abbreviations: Q, glutamine; G, glycine; L, leucine; K, lysine; and A, alanine. (b) Cumulative RNA polymerase activity of the QGLK mutations in the 9n DNA polymerase scaffold. The full-length marker was generated using dNTPs. (c) RNA polymerase activity for wild-type and engineered 9n, Tgo, Kod, and DV polymerases bearing the QGLK mutations. (d) RNase digestion of primer-extended RNA products generated by DNA-dependent RNA synthesis. (e) RNA polymerase efficiency assay performed on a single sequence and a naïve library using 9n, Tgo, Kod, and DV polymerases engineered with QGLK.

engineering. The enzyme that best satisfied our criteria is a mutant version of an Archaean DNA polymerase isolated from the species *Thermococcus gorgonarius* (Tgo).[37] This polymerase, referred to as Tgo-QGLK, contains mutations at three of the eight predicted SDRs (Y409G, A485L, and E664 K) as well as additional mutations at the 3′,5′-exonuclease (exo-) silencing (D141A and E143A) and uracil-stalling (V93Q) positions (Figure 3a).

We constructed a series of engineered polymerase variants by sequentially introducing the QGLK mutations into a related Archaean DNA polymerase isolated from *Thermococcus* sp. 9°N exo- (9n). Following expression and purification from *E. coli*, we challenged 9n-QGLK and its variants to extend a DNA primer−template complex with ribonucleotide triphosphates (NTPs). We found that the QGLK mutations endow 9n with strong RNA synthesis activity, as evidenced by the ability of 9n-QGLK to efficiently extend a DNA primer with 40 sequential ribonucleotides (Figure 3b). With the exception of 9n-GLK, the remaining polymerases each stalled after 10−15 nucleotide incorporations. The observation that the GLK mutations function with high activity is consistent with the model that these sites occupy SDR positions in the polymerase scaffold. Notably, mutation V93Q, which was not identified in our analysis of SDR positions, exhibits no gain-of-function activity over the GLK mutations in the 9n scaffold (Figure 3b).

Control experiments were performed in the presence and absence of dNTP substrates to ensure that the expressed enzymes were properly folded, functional, and free of dNTP contamination that could lead to a false positive in the primer extension reaction (Supporting Information Figures 2 and 3). In all cases, full-length DNA product was observed when dNTPs were present in the polymerase activity assay, while the primer remained unextended in the absence of dNTPs.

To test the generality of transferring functions between polymerase scaffolds, we performed the same primer-extension assay on a set of related Archaean DNA polymerases isolated from the species *Thermococcus gorgonarius* (Tgo), *Thermococcus kodakarensis* (Kod), and *Pyrococcus* sp. deep vent (DV). Because these enzymes were not readily available, DNA constructs encoding exo- versions of each polymerase along with the QGLK mutations (~2.4 kb) were obtained from DNA2.0. The gene products were cloned into an expression vector, sequence verified, and expressed and purified from *E. coli*. Testing of the recombinant enzymes in a polymerase activity assay revealed that the QGLK variants have strong RNA synthesis activity relative to their wild-type counterparts (Figure 3c,d), indicating that the selected mutations endow all four wild-type scaffolds with the ability to synthesize RNA. While this assay demonstrated that mutations and their encoding phenotypes could be transferred between homologous scaffolds, it did not address other issues of enzyme performance such as template bias and fidelity, which are critical in XNA applications that involve template copying.

It is well-known that many polymerases have intrinsic sequence biases that can limit their activity in functional assays.[38] This problem is overlooked when polymerases are tested on individual templates that by design are unstructured and devoid of nucleotide repeats. To help determine the extent to which a polymerase scaffold can impact the phenotypic expression of one or more beneficial mutations, we evaluated the QGLK polymerases under passive and stringent conditions. In this assay, each polymerase was challenged to synthesize RNA by extending a DNA primer that was annealed to either a

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

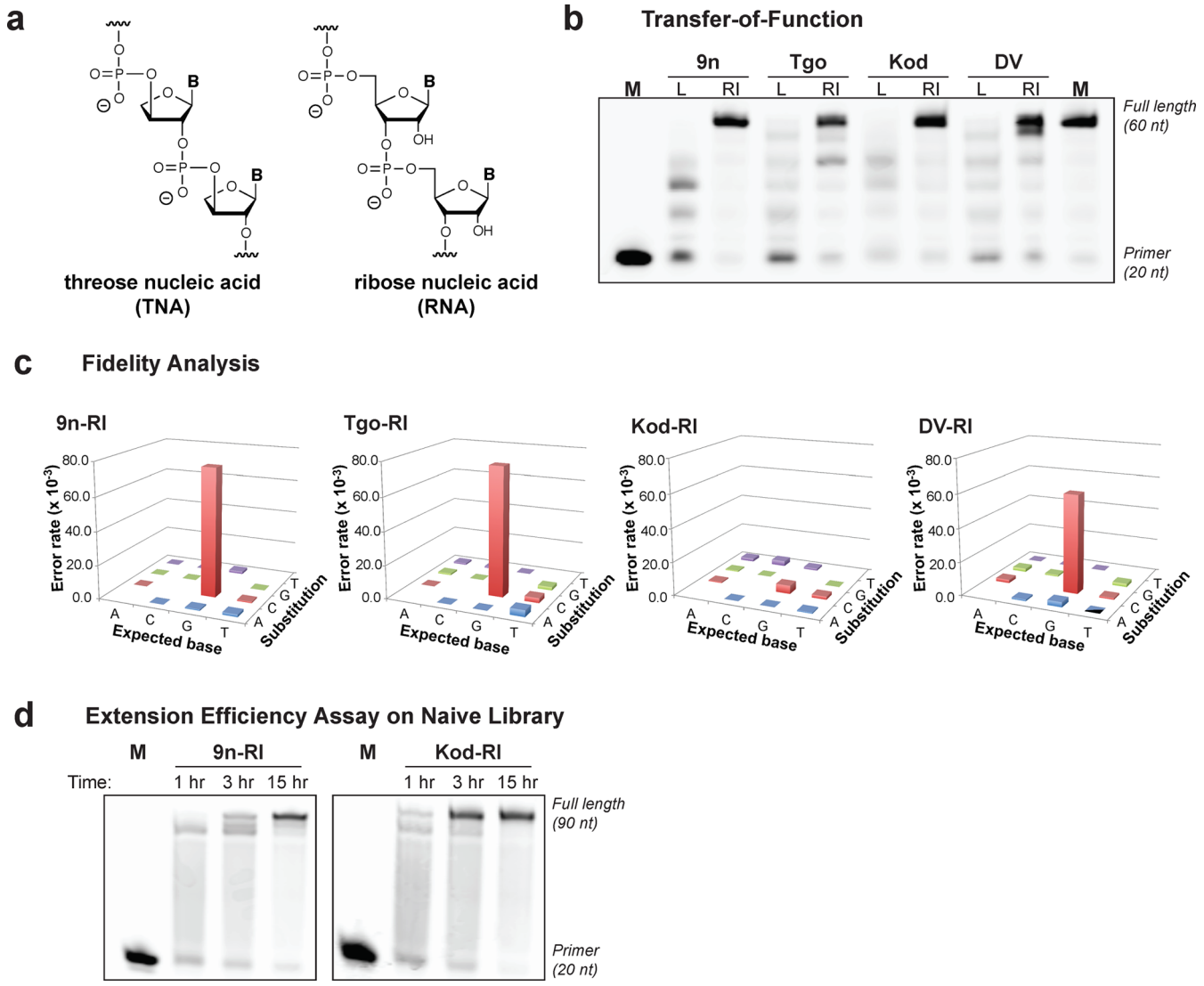


Figure 4. Engineering a DNA-dependent TNA polymerase with improved activity. (a) Constitutional structure for the linearized backbone of threose nucleic acid, TNA (left), and ribose nucleic acid, RNA (right). TNA is an unnatural genetic polymer composed of repeating α-L-threose sugars vicinally connected by 2′,3′-phosphodiester linkages. (b) TNA polymerase activity for engineered 9n, Tgo, Kod, and DV polymerases bearing the L (A485L) and RI (A485R and E664I) mutations. Positions 485 and 664 were identified by mutational analysis of predicted SDRs. (c) Aggregate fidelity profile of TNA replication using 9n-RI, Tgo-RI, Kod-RI, and DV-RI. (d) TNA polymerase efficiency assay performed on a naïve DNA library using 9n-RI and Kod-RI.

single, readily transcribable sequence (passive) or a large population of random sequences (stringent). Consistent with the predictions of scaffold sampling, the four QGLK polymerases exhibit markedly different levels of RNA synthesis activity when challenged with a library of sequences as compared to a single well-defined template (Figure 3e). While all four engineered polymerases (9n, Tgo, Kod, and DV) copy the single template into full-length RNA, only the Tgo and DV scaffolds maintain their activity when challenged with a library of DNA templates. Since no significant differences were observed when the polymerases were assayed in a time course format (Supporting Information Figure 4), we concluded that incubation time was not a critical factor in the reduced activity of the 9n and Kod scaffolds. Instead, our results indicate that the Tgo and DV scaffolds have structural or dynamic properties that allow both polymerases to synthesize RNA with reduced template bias.

To confirm that the increased template-copying efficiency observed for the Tgo and DV scaffolds did not come at a cost of reduced fidelity, we measured the aggregate fidelity of RNA replication for all four QGLK polymerases. This term refers to the fidelity observed after a complete replication cycle (DNA → XNA → DNA), which is operationally different than the more restricted view of fidelity as the accuracy of a single-nucleotide incorporation event (Supporting Information Figure 5). In the context of synthetic biology, aggregate fidelity is an important parameter of polymerase function as it reflects the combined effects of nucleotide misincorporation, insertions and deletions (indel), and the sequence context in which these mistakes occur. By sequencing ~1000 nucleotide positions per enzyme−template pair, we found that the QGLK polymerases misincorporate with a frequency of 1−5 nucleotides per 1000 incorporation events (Supporting Information Table 3), which is low for a polymerase that has been engineered to synthesize noncognate nucleotides. Together, the increased primer

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

extension efficiency (which can also be viewed as reduced template bias) and high fidelity observed for the Tgo and DV scaffolds highlight the utility of scaffold sampling as a strategy for achieving improved polymerase activity.

Expanding Scaffold Sampling to Include Polymerases with XNA Activity. Next, we sought to expand the concept of scaffold sampling to include the development of an engineered polymerase with XNA activity. In particular, we wanted to generate a polymerase that could synthesize TNA—an artificial genetic polymer composed of repeating α-L-threose sugars that are vicinally linked by 2′,3′-phosphodiester bonds (Figure 4a).[39] In addition to serving as a DNA analogue, TNA is an attractive candidate for therapeutic and diagnostic applications due to its stability against nuclease degradation.[18] Unfortunately, the current TNA polymerase (9n-L: sold commercially as Therminator DNA polymerase) suffers from a strong sequence bias that precludes the synthesis of many TNA polymers.[40] This problem is due to a propensity for G−G mispairing in the enzyme active site, which leads to a rise in G to C transversions when TNA is reverse-transcribed back into DNA.[18] While the fidelity of TNA synthesis can be improved with 7-deazaguanine (7dG), a base analogue that suppresses G−G mispairing, 7dG is a costly solution due to the scale at which most TNA transcription reactions are performed.[41]

In an attempt to identify a TNA polymerase variant that functions with improved fidelity, we targeted SDRs at positions 485 and 664, which are known to be strong gatekeepers of substrate specificity.[34,37] Cassette mutagenesis was used to construct variants of 9n that contain random mutations at positions 485 and 664. In total, we generated 12 single-variant polymerases with six mutations at positions 485 and 664, respectively. Polymerase variants were isolated and tested in a polymerase activity assay that involved extending a DNA primer-template complex with chemically synthesized tNTPs.[42,43] This analysis revealed a strong preference for arginine (R) at position 485 and isoluecine (I), glutamine (Q), histidine (H), or lysine (K) at position 664. Subsequent screening of covariants for possible synergistic activity led to the identification of A485R and E664I, termed 9n-RI, as an efficient DNA-dependent TNA polymerase. 9n-RI generates 15 times more TNA than wild-type 9n and can synthesize full-length TNA products on templates that cannot be copied with wild-type 9n or Therminator (data not shown). Surprisingly, we found that 9n-RI has a diminished capacity for DNA synthesis, indicating that the RI mutations alter the wild-type activity of the enzyme (Supporting Information Figure 6).

Next, we examined whether the strong TNA synthesis activity observed for 9n-RI could be extended to other homologues. For this experiment, the RI mutations were inserted into Tgo, Kod, and DV. Following expression and purification from *E. coli* (Supporting Information Figures 2 and 7), the RI mutant polymerases were tested in side-by-side assays against an identical set of polymerases that carried only the Therminator (A485L) mutation (Figure 4b). In this assay, we used a DNA template that we knew was difficult for 9n-L to copy into TNA, so that functional differences could be observed between the four sets of single and double mutant polymerases. Analysis of the resulting primer-extension assay revealed that each of the four RI-polymerases could efficiently extend the DNA primer into a full-length TNA product (Figure 4b). By contrast, the A485L mutant polymerases exhibit noticeably less activity. The fact that the RI mutant polymerases are able to copy a difficult DNA template into TNA demonstrates that the RI-polymerases function with reduced sequence bias relative to their Theminator A485L counterparts.

Encouraged by the enhanced activity of the RI mutant polymerases, we decided to investigate the impact of scaffold sampling on the fidelity and efficiency of TNA synthesis using a more rigorous test of enzyme performance than simple primer extension on a single well-defined template. We began by measuring the aggregate fidelity of TNA replication using an assay that analyzes the DNA product isolated when a DNA template is copied into TNA, purified, and reverse copied back into DNA (Supporting Information Figure 5). For the reverse-transcription step, we used an engineered version of MMLV reverse transcriptase that was previously identified as an efficient TNA-dependent DNA polymerase.[18] By comparing the fidelity profiles produced by the four different polymerase architectures, we found that Kod-RI functions with markedly higher replication fidelity (~8-fold) than the other three RI-polymerases or the intermediate polymerases 9n-L and 9n-R (Figure 4c and Supporting Information Table 3). Furthermore, the enhanced fidelity of Kod-RI was accompanied by a meaningful gain in template synthesis efficiency as evidenced by the ability of Kod-RI to synthesize a full-length TNA library in 3 h as compared to the 15 h required for 9n-RI (Figure 4d). Taken together, the dramatic improvements in TNA synthesis fidelity and efficiency provide strong evidence that scaffold sampling can be used to discover new phenotypic properties that are not present in the starting donor scaffold.

Structural Considerations. Rationalizing the superior activity of Kod-RI requires understanding the individual roles of the RI mutations. Some insights can be gained from the crystal structure of wild-type Kod, which has been solved in both the apo and binary form with DNA (PDB: 1WNS and 4K8Z, respectively).[35,44] Comparison of the two structures indicates that a pronounced structural movement occurs upon DNA binding wherein the thumb domain rotates inward to bind the DNA in a groove formed by the thumb and palm domains. The DNA itself adopts a standard B-form helix with the sugar residues favoring the 2′ endo conformation. Although Kod forms numerous direct contacts to the phosphate backbone, very few interactions are observed to the nucleobases or sugar oxygen atoms. The A485 residue is located on the backside of the finger domain, suggesting that the increased bulk of the arginine residue favors rotation of the finger domain toward the DNA helix, possibly altering the geometry of the enzyme active site. E664 contacts the DNA by interacting with coordinated water molecules in the minor groove of the DNA helix. Substitution of this residue for a hydrophobic side chain could increase TNA synthesis efficiency by weakening contacts to primer−template complex; however, further experiments are need to support this claim.

Significance of Scaffold Sampling. DNA and RNA polymerases are highly selective enzymes that have very little tolerance for modified substrates. This stringent level of substrate specificity makes it difficult to identify polymerases that can recognize unnatural substrates, either as nucleoside triphosphates or as templates, and synthesize long strands of modified nucleotides. Although directed evolution and high throughput library screening methods have been used successfully to broaden the substrate specificity profile of many natural polymerases,[13,14] such approaches are difficult to implement and can be overly expensive when using XNA substrates that are not commercially available.

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

In this study, we developed a polymerase engineering strategy that uses structural and phylogenetic information to identify conserved residues that control substrate specificity. These highly conserved sites, which we refer to as SDRs or specificity determining residues are then targeted by cassette mutagenesis to identify mutations that function with enhanced activity in a polymerase activity assay. Mutations that exhibit the highest gain-of-function activity are then further explored for additional gains in activity using a scaffold sampling process that examines the phenotypic properties of key beneficial mutations in a series of homologous protein scaffolds.

Our method of polymerase engineering was developed to help reduce the time and cost associated with polymerase engineering efforts aimed at developing polymerases for XNA substrates that are not commercially available. This methodology builds on earlier concepts in directed evolution where "smart libraries" have been developed to limit the search through sequence space to regions of an enzyme that have a higher likelihood of yielding functional variants.[27] Rationale design or rationale design aided by phylogenetic information have been used successfully to isolate catalysts from focused libraries that target amino acid residues in close structural proximity to the substrate.[28,45−47] However, close analysis of the selection results indicate that gain-of-function mutations are often biased toward amino acid positions with high phylogenetic variability, while loss-of-function mutations often occur at highly conserved positions. By contrast, our results demonstrate that highly conserved sites are critical for generating strong gain-of-function mutations that alter substrate specificity. A second point of distinction between our work and previous studies is the difference between using orthologs as starting points for enzyme evolution and the concept of using orthologs as an engineering tool to fine-tune the functional properties of a particular mutation or set of mutations once a selection is complete. We suggest that the use of scaffold sampling as a fine-grain optimization tool is analogous to Monte Carlo simulations. However, just as with computational simulations, the best solutions will emerge only when a mutation is close to its optimal solution.

Conclusion. Scaffold sampling provides a new paradigm for optimizing the functional properties of strong gain-of-function mutations. We suggest that it many be possible to apply scaffold sampling to other protein classifications where existing enzymes suffer from problems related to low catalytic activity or poor solubility. This line of research provides an exciting opportunity to develop new enzymes that impact broad areas of synthetic biology, molecular medicine, and biotechnology.

METHODS

General Information. DNA oligonucleotides were purchased from Integrated DNA Technologies, purified by denaturing polyacrylamide gel electrophoresis, electroeluted, concentrated by ethanol precipitation, and quantified by UV absorbance. NTPs and dNTPs were purchased from Sigma. TNA triphosphates (tNTPs) were obtained by chemical synthesis as previously described.[23,24] AccuPrime DNA polymerase and SuperScript II reverse transcriptase were purchased from Invitrogen. A CloneJET PCR cloning kit was purchased from Fermentas. RNase A and hen egg lysozyme were purchased from Sigma. The 9n-L gene was kindly provided by Andreas Marx in a pGDR11 expression vector. DNA encoding the genes for DV^{exo-}, Kod^{exo-}, and Tgo^{exo-} were purchased from DNA2.0. Ni-NTA affinity resin was purchased from Qiagen. DpnI restriction enzyme was purchased from NEB. Ultracell YM-30 concentrators were purchased from Millipore. Costar Spin-X protein concentrating columns were purchased from Sigma.

Sequence and Structural Analysis of Family B DNA Polymerases. A list of family B DNA polymerases with structural coordinates available in the protein databank (PDB) was compiled by searching the PDB for family B DNA polymerases. In cases where multiple structures were present for the same enzyme, the polymerase with the most complete 3D data set was chosen as the representative member. In total, we identified 12 unique DNA polymerases that derive from diverse evolutionary clades including Archaea, Eubacteria, and Eukarya, and the viruses that infect these organisms. The selected polymerases are *Desulfurococcus* sp. Tok (1QQC), *Escherichia coli* DNA Polymerase II (3MAQ), Herpes Simplex virus type 1 (2GV9), Phi 29 (2PZS), *Pyrococcus abyssi* (4FM2), *Pyrococcus furiosus* (2JGU), Enterobacteria Phage RB69 (1CLQ), *Sulfolobus solfataricus* (1S5J), *Thermococcus gorgonarius* (1TGO), *Thermococcus kodakarensis* (4K8Z), *Thermococcus* sp. 9°N-7 (4K8X), and Yeast pol delta (3IAY). For these polymerases, we calculated the sequence and structural conservation at each amino acid position. Sequence conservation was calculated using MUSCLE. Structural conservation was calculated with PyMol by separating the proteins into their individual domains (N-terminal, exonuclease, palm, finger, and thumb) and then performing a multiple structure alignment on the individual domains. The sequence and structural analyses produced two separate histograms that were overlaid with the sequence of our model DNA polymerase (*Thermococcus* 9°N). The list of conserved sites was then narrowed to the subset of amino acid positions that are located within 10 Å of the DNA primer−template complex.

Site-directed Mutagenesis. The polymerase genes were cloned into the pGDR11 vector, and amino acid substitutions were sequentially introduced by site-directed mutagenesis.[32] Briefly, 30 ng of plasmid was combined with 5 pmol of primer (Supporting Information) and 2 units of Accuprime DNA polymerase in 1× AccuPrime buffer in 20 μL reaction volume. The solution was thermocycled as follows: 2 min at 95 °C followed by 16 cycles of 30 s at 95 °C, 60 s at 55 °C, and 7.5 min at 72 °C. After cycling, DpnI (5 units) was added, and the solution was incubated for 1 h at 37 °C to digest the starting DNA plasmid. The undigested DNA was transformed into XL1-blue chemically competent *E. coli* and cloned, and the correct mutations were verified by sequencing the entire gene (ASU Core Facility). A list of all polymerases mutants can be found in Supporting Information Table 2.

Polymerase Expression and Purification. *E. coli* strain XL1-Blue carrying the pGDR11 vector encoding the polymerase of interest was streaked onto solid media, and a starter culture was established by inoculating LB-ampicillin liquid media with a single colony. The starter culture was used to inoculate an expression culture at OD_{600} of 0.1 in LB-ampicillin. The culture was grown at 37 °C with shaking until reaching an OD_{600} of 0.8 at which time protein expression was induced with 1 mM IPTG. After an additional 4 h of growth at 37 °C, the cells were pelleted, resuspended in nickel purification buffer [50 mM phosphate, 250 mM sodium chloride, 10% (v/v) glycerol, pH 8], sonicated with 0.1 mg mL^{-1} hen egg lysozyme, and heat treated for 45 min at 75 °C. The lysate was clarified by centrifugation for 15 min at 4000 rpm. Polymerases were purified from the clarified lysate by binding to a nickel affinity

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

resin. The column was washed with nickel purification buffer and eluted with nickel purification buffer supplemented with 75 mM imidazole. Eluted protein was exchanged into storage buffer [10 mM Tris-HCl, 100 mM KCl, 1 mM DTT, 0.1 mM EDTA, pH 7.4] and concentrated using an Amicon (10 NMWL) protein concentrator. Polymerase samples were normalized by comparing their A_{280} absorption value and assaying protein concentrations by SDS PAGE with coumassie staining. Some polymerases were further diluted so that each protein had the same approximate band density.

Polymerase Activity Assay. Polymerase activity assays were performed in a 10 μL reaction volume containing 5 pmol of DNA primer−template complex. The RNA polymerase reactions (and corresponding DNA control reactions) used the 4nt.3ga template for a single sequence and a random library (N_{50}) to represent a complex mixture of templates. TNA polymerase reactions (and corresponding DNA control reactions) used the 4nt.1 g-2g-3g template for a single sequence and the random library (N_{50}) to represent a complex mixture of templates. In all cases, the PBS2 primer was labeled at the 5′-end with an IR800 dye. The primer−template complex was annealed in 1× ThermoPol buffer [20 mM Tris-HCl, 10 mM $(NH_4)_2SO_4$, 10 mM KCl, 2 mM $MgSO_4$, 0.1% Triton X-100, pH 8.8] by heating for 5 min at 90 °C and cooling for 10 min at 4 °C. For assays performed with tNTPs, the polymerase (1 μL volume of 0.5−1.0 mg mL^{-1} solution) was pretreated with 1 mM $MnCl_2$ then added to the reaction mixture. We have previously shown that manganese ions reduce the substrate specificity of the polymerase.[24] The slight variation in TNA polymerase amounts is based on differences in activity observed each time the enzyme is expressed and purified. For reactions with dNTPs, NTPs, or no added triphosphates, the polymerase was added directly to the reaction mixture. The reactions were initiated with the addition of nucleotide triphosphates (100 μM), and the solutions were incubated at 55 °C. The extension time for RNA synthesis was 3 h, while the extension time for TNA and DNA synthesis was 30 min. Extension times on DNA libraries varied according to each time course reaction and are noted in the figures. The reactions were quenched in stop buffer [1× Tris-boric acid buffer, 20 mM EDTA, 7 M urea, pH 8] and analyzed by denaturing PAGE and visualized using a LI-COR Oddysey CLx imager.

Polymerase Controls for Activity and Contamination. To ensure that the recombinant polymerases were functional and free of dNTP or NTP contamination left over from the cell lysate, the enzymes were challenged to extend a DNA primer−template complex in the presence and absence of dNTPs using conditions described in the polymerase activity assay. Polymerase activity assays were analyzed by denaturing PAGE and imaged using a LI-COR Odyssey CLx.

RNase Digestion. To verify that the QGLK polymerases generated full-length RNA products, the extended material was digested with 2.5 μg of RNase A for 1 h at 37 °C. The reactions were quenched in stop buffer (see above) and denatured by incubating for 5 min at 70 °C before cooling to RT. Reaction products were analyzed by denaturing PAGE and imaged using a LI-COR Odyssey CLx.

Aggregate Fidelity Analysis. Aggregate fidelity reactions were performed by extending a DNA primer-template complex in a 100 μL reaction volume containing 100 pmol of fidelity.temp and 100 pmol of PBS2.mismatch primer. The primer and template were annealed in 1× ThermoPol buffer by heating for 5 min at 95 °C and cooling for 10 min at 4 °C. The polymerase (5 ug) was added to the reaction mixture. For TNA extensions, the polymerase was pretreated with 1 mM $MnCl_2$. The reactions were initiated by addition of the nucleotide triphosphates (100 μM). Following a 4 h incubation at 55 °C, the reactions were quenched in stop buffer and denatured at 90 °C for 5 min. Elongated primers were purified by denaturing PAGE, electroeluted, and concentrated using a YM-30 concentrator device.

The purified transcripts (TNA or RNA) were reverse transcribed in a final volume of 100 μL. PBS1 primer (100 pmol) was annealed to the template in 1× First Strand Buffer [50 mM Tris-HCl, 75 mM KCl, 3 mM $MgCl_2$, pH 8.3] by heating for 5 min at 90 °C and cooling for 10 min at 4 °C. Next, 500 μM dNTPs and 10 mM DTT were added, and the reaction was allowed to incubate for 2 min at 42 °C. Finally, 3 mM $MgCl_2$ and 10 U μL^{-1} SuperScript II reverse transcriptase were added, and the reaction was allowed to incubate for 1 h at 42 °C. For reactions performed on TNA templates, 1.5 mM $MnCl_2$ was included in the reaction mixture.

After reverse transcription, the cDNA strand was PCR amplified with PBS1 and extra.primer primers and ligated into a pJET vector following manufacturer's protocol. The ligated product was transformed into XL1-blue *E. coli*, cloned, and sequenced (ASU Core Facility). Sequencing results were analyzed using CLC Main Workbench. Sequences lacking the T to A watermark were discarded as they were generated from the starting DNA template rather than replicated material.

■ ASSOCIATED CONTENT

Supporting Information

The Supporting Information is available free of charge on the ACS Publications website at DOI: 10.1021/acschembio.5b00949.

Sequences for the expression plasmids for the four wild-type polymerases have been deposited in Genbank under accession codes KP682506 (9n), KP682507 (Tgo), KP682508 (Kod), and KP682509 (DV) (PDF)

■ AUTHOR INFORMATION

Corresponding Author

*Tel.: 949-824-8149. E-mail: jchaput@uci.edu.

Present Address

[⊥]Department of Pharmaceutical Sciences, University of California, Irvine, CA 92697

Author Contributions

M.R.D. performed conservation analysis. M.R.D. and C.O. generated polymerase variants and performed activity assays. K.E.F. expressed and purified the polymerases. M.R.D. and J.C.C. designed the study and wrote the paper. All authors discussed results and commented on the manuscript.

Notes

The authors declare no competing financial interest.

■ ACKNOWLEDGMENTS

We would like to thank J. Szostak for helpful discussions and critical reading of the manuscript and A. Marx for providing the pGDR11 expression plasmid. This work was supported by the DARPA Folded Non-Natural Polymers with Biological Function (Fold F(x)) Program under award number

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

N66001-14-2-4054 and by a grant from the National Science Foundation 1304583.

REFERENCES

(1) Chaput, J. C., Yu, H., and Zhang, S. (2012) The emerging world of synthetic genetics. *Chem. Biol. 19*, 1360−1371.
(2) Anosova, I., Kowal, E. A., Dunn, M. R., Chaput, J. C., Van Horn, W. D., and Egli, M. (2016) The structural diversity of artificial genetic polymers. *Nucleic Acids Res.*, gkv1472.
(3) Leitzel, J. C., and Lynn, D. G. (2001) Template-directed ligation: from DNA towards different versatile templates. *Chem. Rec. 1*, 53−62.
(4) Chaput, J. C., and Switzer, C. (2000) Non-enzymatic transcription of an isoG-isoC base pair. *J. Am. Chem. Soc. 122*, 12866−12867.
(5) Chaput, J. C., Sinha, S., and Switzer, C. (2002) 5-Propynyluracil-diaminopurine: an efficient base pair for non-enzymatic transcription of DNA. *Chem. Commun.*, 1568−1569.
(6) Rosenbaum, D. M., and Liu, D. R. (2003) Efficient and sequence-specific DNA-templated polymerization of peptide nucleic acid aldehydes. *J. Am. Chem. Soc. 125*, 13924−13925.
(7) Heuberger, B. D., and Switzer, C. (2006) Nonenzymatic synthesis of RNA by TNA templates. *Org. Lett. 8*, 5809−5811.
(8) Koppitz, M., Nielsen, P. E., and Orgel, L. E. (1998) Formation of oligonucleotide-PNA-chimeras by template-directed ligation. *J. Am. Chem. Soc. 120*, 4563−4569.
(9) Ura, Y., Beierle, J. M., Leman, L. J., Orgel, L. E., and Ghadiri, M. R. (2009) Self-assembling sequence-adaptive peptide nucleic acids. *Science 325*, 73−77.
(10) Brudno, Y., Birnbaum, M. E., Kleiner, R. E., and Liu, D. R. (2010) An in vitro translation, selection and amplification system for peptide nucleic acids. *Nat. Chem. Biol. 6*, 148−155.
(11) Niu, J., Hili, R., and Liu, D. R. (2013) Enzyme-free translation of DNA into sequence-defined synthetic polymers structurally unrelated to nucleic acids. *Nat. Chem. 5*, 282−292.
(12) Alba, M. M. (2001) Replicative DNA polymerases. *Genome Biol. 2*, reviews3002.1.
(13) Loakes, D., and Holliger, P. (2009) Polymerase engineering: towards the encoded synthesis of unnatural polymers. *Chem. Commun.*, 4619−4631.
(14) Chen, T., and Romesberg, F. E. (2014) Directed polymerase evolution. *FEBS Lett. 588*, 219−229.
(15) Yang, Z. Y., Chen, F., Alvarado, J. B., and Benner, S. A. (2011) Amplification, Mutation, and Sequencing of a Six-Letter Synthetic Genetic System. *J. Am. Chem. Soc. 133*, 15105−15112.
(16) Malyshev, D. A., Dhami, K., Quach, H. T., Lavergne, T., Ordoukhanian, P., Torkamani, A., and Romesberg, F. E. (2012) Efficient and sequence-independent replication of DNA containing a third base pair establishes a functional six-letter genetic alphabet. *Proc. Natl. Acad. Sci. U. S. A. 109*, 12005−12010.
(17) Moran, S., Ren, R. X.-F., Rumney, S., and Kool, E. T. (1997) Difluorotoluene, a nonpolar isostere for thymine, codes specifically and efficiently for adenine in DNA replication. *J. Am. Chem. Soc. 119*, 2056−2057.
(18) Yu, H., Zhang, S., Dunn, M., and Chaput, J. C. (2013) An efficient and faithful in vitro replication system for threose nucleic acid. *J. Am. Chem. Soc. 135*, 3583−3591.
(19) 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., and Holliger, P. (2012) Synthetic genetic polymers capable of heredity and evolution. *Science 336*, 341−344.
(20) Yu, H., Zhang, S., and Chaput, J. C. (2012) Darwinian evolution of an alternative genetic system provides support for TNA as an RNA progenitor. *Nat. Chem. 4*, 183−187.
(21) Taylor, A. I., Pinheiro, V. B., Smola, M. J., Morgunov, A. S., Peak-Chew, S., Cozens, C., Weeks, K. M., Herdewijn, P., and Holliger, P. (2015) Catalysts from synthetic genetic polymers. *Nature 518*, 427−430.
(22) Kimoto, M., Yamashige, R., Matsunaga, K.-I., Yokoyama, S., and Hirao, I. (2013) Generation of high affinity DNA aptamers using an expanded genetic alphabet. *Nat. Biotechnol. 31*, 453−457.
(23) Sefah, K., Yang, Z., Bradley, K. M., Hoshika, S., Jimenez, E., Zhang, L., Zhu, G., Shanker, S., Yu, F., Turek, D., Tan, W., and Benner, S. A. (2014) In vitro selection with artificial expanded genetic information systems. *Proc. Natl. Acad. Sci. U. S. A. 111*, 1449−1454.
(24) Chaput, J. C., and Szostak, J. W. (2003) TNA synthesis by DNA polymerases. *J. Am. Chem. Soc. 125*, 9274−9275.
(25) Horhota, A., Zou, K., Ichida, J. K., Yu, B., McLaughlin, L. W., Szostak, J. W., and Chaput, J. C. (2005) Kinetic analysis of an efficient DNA-dependent TNA polymerase. *J. Am. Chem. Soc. 127*, 7427−7434.
(26) Sousa, R. (1996) Structural and mechanistic relationships between nucleic acid polymerases. *Trends Biochem. Sci. 21*, 186−190.
(27) Jochens, H., and Bornscheuer, U. T. (2010) Natural diversity to guide focused directed evolution. *ChemBioChem 11*, 1861−1866.
(28) Hibbert, E., Senussi, T., Costelloe, S. J., Lei, W., Smith, M. E. B., Ward, J. M., Hailes, H. C., and Dalby, P. A. (2007) Directed evolution of transketolase activity on non-phosphorylated substrates. *J. Biotechnol. 131*, 425−432.
(29) Li, L., Shakhnovich, E. I., and Mirny, L. A. (2003) Amino acids determining enzyme-substrate specificity in prokaryotic and eukaryotic protein kinases. *Proc. Natl. Acad. Sci. U. S. A. 100*, 4463−4468.
(30) Chen, R., and Jeong, S.-S. (2000) Functional predictions: identification of protein orthologs and paralogs. *Protein Sci. 9*, 2344−2353.
(31) Stemmer, W. P. C. (1994) Rapid evolution of a protein in vitro by DNA shuffling. *Nature 370*, 389−391.
(32) Gardner, A. F., and Jack, W. E. (1999) Determinants of nucleotide sugar recognition in an archaeon DNA polymerase. *Nucleic Acids Res. 27*, 2545−2553.
(33) Staiger, N., and Marx, A. (2010) A DNA polymerase with increased activity for ribonucleotides and C5-modified deoxyribonucleotides. *ChemBioChem 11*, 1963−1966.
(34) McCullum, E. O., and Chaput, J. C. (2009) Transcription of an RNA aptamer by a DNA polymerase. *Chem. Commun.*, 2938−2940.
(35) Bergen, K., Betz, K., Welte, W., Diederichs, K., and Marx, A. (2013) Structures of KOD and 9°N DNA polymerases complexed with primer template duplex. *ChemBioChem 14*, 1058−1062.
(36) Morley, K. L., and Kazlauskas, R. J. (2005) Improving enzyme properties: when are closer mutations better? *Trends Biotechnol. 23*, 231−237.
(37) Cozens, C., Pinheiro, V. B., Vaisman, A., Woodgate, R., and Holliger, P. (2012) A short adaptive path from DNA to RNA polymerases. *Proc. Natl. Acad. Sci. U. S. A. 109*, 8067−8072.
(38) Ross, M. G., Russ, C., Costello, M., Hollinger, A., Lennon, N. J., Hegarty, R., Nusbaum, C., and Jaffe, D. B. (2013) Characterizing and measuring bias in sequencing data. *Genome Biol. 14*, R51.
(39) Schoning, K. U., Scholz, P., Guntha, S., Wu, X., Krishnamurthy, R., and Eschenmoser, A. (2000) Chemical etiology of nucleic acid structure: the alpha-threofuranosyl-(3′− > 2′) oligonucleotide system. *Science 290*, 1347−1351.
(40) Ichida, J. K., Horhota, A., Zou, K., McLaughlin, L. W., and Szostak, J. W. (2005) High fidelity TNA synthesis by therminator polymerase. *Nucleic Acids Res. 33*, 5219−5225.
(41) Dunn, M. R., Larsen, A. C., Fahmi, N., Zahurancik, W. J., Meyers, M., Suo, Z., and Chaput, J. C. (2015) Therminator-mediated synthesis of unbiased TNA polymers requires 7-deazaguanine to suppress G-G mispairing during TNA transcription. *J. Am. Chem. Soc. 137*, 4014−4017.
(42) Zhang, S., and Chaput, J. C. (2012) Synthesis of threose nucleic acid (TNA) phosphoramidite monomers and oligonucleotide polymers. *Cur. Protoc. Nucleic Acid Chem. 4*, 4.51.1.
(43) Zhang, S., Yu, H., and Chaput, J. C. (2013) Synthesis of threose nucleic acid (TNA) triphosphates and oligonucleotides by polymerase-mediated primer extension. *Cur. Protoc. Nucleic Acid Chem. 52*, 4.54.51−54.54.57.
(44) Hashimoto, H., Nishioka, M., Fujiwara, S., Takagi, M., Imanaka, T., Inoue, T., and Kai, Y. (2001) Crystal structure of DNA polymerase

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219

from hyperthermophilic archaeon Pyrococcus Kodakaraensis KOD1. *J. Mol. Biol. 306*, 469−477.
(45) Hohne, M., Schatzle, S., Jochens, H., Robins, K., and Bornscheuer, U. T. (2010) Rational assignment of key motifs for function guides in silico enzyme identification. *Nat. Chem. Biol. 6*, 807−813.
(46) Cochran, J. R., Kim, Y.-S., Lippow, S. M., Rao, B., and Wittrup, K. D. (2006) Improved mutants from directed evolution are biased to orthologous substitutions. *Protein Eng., Des. Sel. 19*, 245−253.
(47) Khanal, A., McLoughlin, S. Y., Kershner, J. P., and Copley, S. D. (2015) Differential effects of a mutation on the normal and promiscuous activities of orthologs: implications for natural and directed evolution. *Mol. Biol. Evol. 32*, 100−108.

DOI: 10.1021/acschembio.5b00949
ACS Chem. Biol. 2016, 11, 1210−1219