Building better polymerases: Engineering the replication of expanded genetic alphabets

Zahra Ouaray1, Steven A. Benner2, Millie M. Georgiadis3,*, and Nigel G. J. Richards1,2,*,†

From the 1School of Chemistry, Cardiff University, Park Place, Cardiff, United Kingdom, the 2Foundation for Applied Molecular Evolution, Alachua, Florida, USA, and the 3Department of Biochemistry and Molecular Biology, Indiana University School of Medicine, Indianapolis, Indiana, USA

Edited by John M. Denu

DNA polymerases are today used throughout scientific research, biotechnology, and medicine, in part for their ability to interact with unnatural forms of DNA created by synthetic biologists. Here especially, natural DNA polymerases often do not have the “performance specifications” needed for transformative technologies. This creates a need for science-guided rational (or semi-rational) engineering to identify variants that replicate unnatural base pairs (UBPs), unnatural backbones, tags, or other evolutionarily novel features of unnatural DNA. In this review, we provide a brief overview of the chemistry and properties of replicative DNA polymerases and their evolved variants, focusing on the Klenow fragment of Taq DNA polymerase (Klentaq). We describe comparative structural, enzymatic, and molecular dynamics studies of WT and Klentaq variants, complexed with natural or noncanonical substrates. Combining these methods provides insight into how specific amino acid substitutions distant from the active site in a Klentaq DNA polymerase variant (ZP Klentaq) contribute to its ability to replicate UBPs with improved efficiency compared with Klentaq. This approach can therefore serve to guide any future rational engineering of replicative DNA polymerases.

DNA polymerases (1), RNA polymerases (2), and reverse transcriptases (3) are enzymes central to terran biology and permit life to propagate and exploit its genetic inheritance. DNA polymerases are also workhorses in biotechnology (4, 5), where they often must manage DNA molecules having unusual structural features. For example, some WT bacterial polymerases can replicate DNA-containing UBPs that interact by steric complementarity without any interbase hydrogen bonding (6) (Fig. 1). This finding has been exploited to obtain engineered strains of Escherichia coli that can express proteins built from noncanonical amino acids with improved efficiency (11). This capability was a driver of the recent $2.5 billion acquisition of Synthorx by Sanofi-Aventis.

DNA polymerase variants are also playing key roles at other scientific frontiers. For example, NextGen sequencing (17) has transformed medicine (18), but only because it was enabled by the use of DNA polymerase variants capable of replicating “Watson–Crick” (WC) deoxynucleoside triphosphates (7) “tagged” with fluorescent markers and carrying blocking groups appended to their 3′-positions (19).

Expanded genetic alphabets and engineered DNA polymerases

These prior studies, however, only “scratch the surface” of the possible modified nucleotides that have scientific, biotechnological, and medical value. For example, DNA can have many more independently replicable groups than A, T, G, or C (Fig. 1). Scientifically, these groups can be varied to learn whether WC DNA represents the only molecular solution for the long-term storage and replication of biological information (11, 20–22). Alexander Rich was likely the first to recognize the importance of this question, suggesting in 1962 that the isocytidine–isoguanine (S:B) pair (Fig. 1) might be used as an additional information storing unit in DNA (23).

Further, many have asked whether the sugar and phosphate components of the DNA backbone might be modified. Synthetic biologists have therefore synthesized numerous such modified DNA molecules to delineate molecular changes that can (and cannot) be tolerated by duplex DNA (24–26). The extent to which the core concept of WC pairing can be changed has also been explored by synthetic biologists. For example, the laboratories of Eric Kool (8), Floyd Romesberg (9, 10), and Ichiro Hirao (13, 14) have attempted to dispense with interbase hydrogen bonding (Fig. 1).

Developing a different approach, Steven Benner and coworkers (11, 12) synthesized UBPs that exploit orthogonal patterns of hydrogen bonding. This led to multiple discoveries about the molecular changes that can (and cannot) be tolerated in replicating units, especially how the fidelity of replication can be influenced by protonation, deprotonation, and multiple tautomeric forms of the nucleobases (27). These efforts led to the preparation of “hachimoji” DNA (28), an artificially expanded genetic system (AEGIS) in which A, G, T, and C are augmented by two unnatural pyrimidine analogs, pseudocytidine (or isocytidine) (S) and 6-aminopurine (2′-deoxyribofuranosyl)-5-nitro-1H-pyridin-2-one (Z), and their size and hydrogen bond complementary partners, the purine analogs, isoguanine (B), and 2-amino-8-(1β-d-2′-deoxyribofuranosyl)imidazo[1,2-a]1,3,5-triazin-[8H]-4-one- (P) (27, 28) (Fig. 1). The additional P:Z and B:S base pairs form three interbase hydrogen bonds, as does the natural G: C. Work in this area has also been driven by commercial
factors; the use of AEGIS UBPs as components of diagnostic tests has generated lifetime sales in excess of $1 billion (29).

Future disruptive technologies exploiting these and other expanded genetic alphabets will almost certainly require engineered DNA polymerase variants having expanded ranges of biophysical and catalytic properties. Unfortunately, and despite extensive research into the structure and mechanism of DNA polymerases from a wide range of organisms (30–36), rationally predicting how specific amino acid replacements in these enzymes will impact the incorporation of any particular kind of UBP into duplex DNA remains almost impossible. This reflects a lack of fundamental understanding about how multiple factors act synergistically to control UBP incorporation efficiency and fidelity (30, 37, 38). These factors include nucleobase pair complementarity, nucleobase tautomerism, variations in hydrogen bond free energies, and the conformational dynamics of the polymerase itself as incorporation proceeds through the catalytic cycle. Directed evolution methods (39) are therefore viewed as the best approach to identifying DNA polymerase variants that can incorporate UBPs with high efficiency and fidelity (40–42).

These facts make it timely to review how X-ray crystallography, computational modeling, and directed evolution strategies can be combined to obtain DNA polymerases capable of replicating artificial DNA containing UBPs. Given the recent publication of a comprehensive discussion about the replication of UBP-containing DNA by unengineered WT polymerases (43), we draw on work from our laboratories, the only studies (to our knowledge) that employ a combined experimental and computational approach.

Generally speaking, enzymes that replicate DNA must manage three contradictory demands as they catalyze their critical reactions. First, this class of polymerases must be highly specific, making not much more than one mistake every billion turnovers (44). Mistakes mean somatic mutations potentially leading to cancer, germ line mutations that create inherited genetic disease, and, ultimately, an error catastrophe that leads to the death of the organism. Typical replicative polymerases achieve specificity at both the elongation step and a proofreading step, the latter in the form of 3’–5’ exonuclease domains (see below). Other replicative DNA polymerases, such as reverse transcriptases (45), lack

---

**Figure 1. Chemical structures of nucleobase pairs.** R indicates the point of covalent attachment to either deoxyribose or ribose in DNA or RNA, respectively. A, Watson–Crick nucleobase pairs (7). B, isosteric UBPs developed by Kool and co-workers (8). C, hydrophobic UBPs developed by Romesberg and co-workers (9, 10). D, hydrogen-bonding UBPs used in hachimoji DNA developed by Benner and co-workers (11, 12). E, hydrophobic UBPs developed by Hirao and co-workers (13, 14), where R₁ can be a variety of substituents.
proofreading and are therefore more error-prone. DNA polymerases that function primarily in DNA repair, including those important for genome stability and trans-lesion synthesis, also often have lower fidelity (46, 47).

Second, DNA polymerases must accept four different substrates: (i) template dG and dCTP, (ii) template dC and dGTP, (iii) template dA and dTTP, and (iv) template dT and dATP. Most other enzymes accept only one.

Finally, DNA polymerases must work rapidly. An E. coli cell replicates its entire genome by copying 4000 nucleotides each second. Even though many replication forks are used, this rate is as fast as that of many enzymes using only one substrate in pathways able to tolerate mistakes.

Managing these demands is especially difficult because the four substrates have few molecular features in common to present to the polymerase (Fig. 1). For example, the four nucleobases present different functionalities to the major groove, a methyl group for T, a hydrogen for C, two hydrogen bond acceptors for G, and one donor and one acceptor for A.

One feature that all four substrates have in common is their relative size, leading to the rule that large purines must pair with small pyrimidines. Further, the functional groups located on the sides of all four bases in the minor groove of the DNA duplex share a common feature: electron density presented by the exocyclic C=O groups of T and C and the N-3 nitrogen in the purine rings of A and G. Thus, the “minor groove scanning hypothesis” proposes that polymerases donate hydrogen bonds to this electron density in the template, primer, and incoming triphosphate as a way of enforcing Watson–Crick geometry, and hence fidelity, on base pair recognition (48).

These hydrogen bond contacts and size constraints are well-represented in over two dozen crystal structures solved to date (32, 43). The electronic constraint in the minor groove also appears to be important for nonstandard base pairing, even pairing that, as in the Romesberg pair (Fig. 1), involves no inter-base hydrogen bonding (22). Moreover, the easiest AEGIS UBP to incorporate into DNA by polymerase-catalyzed reactions has been Z: P (Fig. 1), which is the only pair of those with shuffled hydrogen-bonding patterns for which both components present electron density to the minor groove (49).

Although the coupling reaction catalyzed by DNA polymerases seems superficially straightforward, this simplicity hides many details that remain controversial (30, 34, 50–52). DNA synthesis proceeds in the 5’ to 3’ direction in a primer-dependent reaction requiring Mg2+ as a cofactor. In the first step, the enzyme forms an initial binary complex with a DNA duplex composed of a template strand and a shorter, complementary primer strand possessing a 3’-hydroxyl group. dNTPs then bind to the binary complex to give a ternary (or “pre-incorporation”) complex (Fig. 2). Correctly matched dNTPs reside longer in the active site, allowing the enzyme to adopt a “closed” conformation in which the 3’-hydroxyl group in the primer can attack the α-phosphate of the bound dNTP (Fig. 2A). This reaction requires the presence of two (53, 56), or possibly three (30, 54, 55), Mg2+ ions within the active site of human polymerase β (Fig. 2C) or KlenTaq (Fig. 2D). Subsequent release of PPi generates a binary (or “post-incorporation”) complex in which the primer has been elongated by one nucleobase, which then serves as the substrate for the next templated addition of dNTP. These steps are repeated until no unpaired nucleobases remain in the template strand.

Replicative DNA polymerases, especially those for which DNA binding is the rate-limiting step, exhibit a property referred to as processivity (57). Processivity is a measure of the number of dNTPs incorporated into the growing strand before the template-primer duplex dissociates from the enzyme (58). Processivity, of course, permits these DNA polymerases to achieve high rates of nucleotide incorporation. This process requires translocation of the enzyme along the template-primer or the template-primer along the enzyme (depending on your point of view) during polymerization. We discuss the conformational changes that occur during dNTP incorporation below.

When mismatched dNTPs are present in the ternary complex, kinetic evidence shows that the active site remains in an “open” conformation so that dissociation of the incoming nucleotide can take place prior to formation of the new P–O bond (37, 38) (Fig. 2B). Alternatively, should a mismatched dNTP be incorporated, most replicative polymerases include a 3’–5’ exonuclease domain that can remove the mismatch. In this case, the 3’-hydroxyl end of the primer shifts from the polymerase to the exonuclease active site for removal of the mismatched nucleotide (59, 60). This proofreading step is essential for ensuring fidelity during replication by DNA polymerases, which typically make errors every 10⁴ to 10⁵ nucleobases. The 3’–5’ exonucleolytic activity improves this error rate by 2–3 orders of magnitude (61).

A variety of assays exist to assess the kinetics of nucleobase incorporation and processivity. These have been extensively reviewed (34, 62), but here we give two examples used in our work to determine either incorporation efficiency (63) or fidelity (64) when DNA polymerases and their variants are challenged with AEGIS UBPs (Fig. 3). For example, incorporation efficiency can be measured by nested PCR using primers tagged at the 5’-end with multiple consecutive UBPs. In this assay, ampiclons are formed in the reaction mixture only if the polymerase is capable of replicating the UBPs (Fig. 3A).

Assaying fidelity requires the DNA polymerase to replicate a sequence in which UBPs are present (64) (Fig. 3B). In the event that the UBP is lost during PCR cycles, two possible unique restriction sites are created in the ampiclons, depending on whether the UBP is replaced by T:A or C:G. Incubating the PCR products with two restriction endonucleases therefore gives cleaved products for the ampiclons from which the UBP has been lost in both possible transition mutations (Fig. 3C). This assay thereby reports on the extent to which polymerase fidelity deviates from 100%. In addition, performing the PCR with different concentrations of the UBP components permits a quantitative assessment of fidelity.

Obtaining DNA polymerases engineered for altered substrate specificity

Some, albeit limited, success has been reported in modifying the substrate preferences of DNA polymerases by structure-guided mutagenesis studies (66, 67). The complexity of the conformational changes undergone by the polymerase during each catalytic cycle seems to require that variants contain multiple
residue replacements. Most successful efforts to obtain DNA polymerases with altered catalytic and/or biophysical properties have therefore selected suitable variants from large libraries (40, 68–72). Examples include several Taq polymerase variants evolved for specific functions (Table 1). The fact that several of the amino acid substitutions are found in more than one of these Taq variants may ultimately inform rationale design of variants with new functional properties in the future. To date, however, only one of these published variants (ZP Klentaq) has been subjected to structural and dynamic characterizations to determine the impact of each of the various substitutions on the properties of the variant polymerase (see below).

Of the plethora of directed evolution strategies, compartmentalized self-replication (CSR) (76, 77) or one of its modifications (40, 68) is often used to obtain engineered DNA polymerases. In CSR, water droplets containing appropriate primers and dNTPs are formed in a water/oil emulsion. Droplets that also contain a single E. coli cell carrying a plasmid encoding a single variant of a thermophilic DNA polymerase, such as Klentaq, can then be used to select enzymes with desired properties. In our work (65), we employ the nested PCR strategy described above (Fig. 3A) to identify DNA polymerase variants that can replicate UBPs orthogonal to WC nucleobases, such as Z and P, because only these will replicate copies of their gene in the plasmid (Fig. 4). Those variants that cannot incorporate UBPs efficiently will not replicate their genes. Thus, genes encoding DNA polymerases with the desired substrate specificity become represented more in the mixture of DNA amplicons obtained when the emulsion is broken at the end of several rounds of PCR (Fig. 4). The resulting DNA amplicons are inserted into a new set of plasmids that are used to transform E. coli for use in a subsequent selection round. After a defined number of rounds, genes encoding DNA polymerase variants with the ability to incorporate UBPs become highly enriched in the library. The set of enriched genes can be sequenced to identify the molecular changes that permit the polymerase variant to replicate the UBP. Standard methods can then be used to express and purify individual variants prior to any engineering.

---

**Figure 2.** A, kinetic scheme for the incorporation of matched dNTPs by T7 DNA polymerase (38, 53). When the incoming dNTP is correct, the enzyme undergoes a rapid conformational change from the open to closed form of the pre-incorporation complex. B, kinetic scheme for the incorporation of mis-matched dNTPs by T7 DNA polymerase (38, 53). When the incoming dNTP is incorrect, the rate constant for the chemical step becomes smaller than that for the conformational change back to the open form of the pre-incorporation complex. C, chemical mechanism of primer elongation for the 2-Mg (54) and 3-Mg models (55). The location of the third Mg$^{2+}$ ion together with its water ligands is shown in red. Residue numbering is for human DNA polymerase b. D, chemical mechanism of primer elongation in the active site of T. aquaticus DNA polymerase, which lacks one of the carboxylate ligands present in human DNA polymerase β.
Figure 3. A, nested PCR-based incorporation assay for DNA containing Z:P nucleobase pairs using a target DNA sequence (cyan-red-green) composed of only WC base pairs. Replication is initiated by chimeric primers that can bind to the original gene but have UBPs at the 5' -end. These chimeric primers are present at low concentration (Step 1). After consumption of the chimeric primers (Step 2), amplification is continued by external 5'-PPPP-OH primers (containing only UBPs), which are present at high concentration. As a result, full-length amplicons are produced only if the DNA polymerase can efficiently incorporate UBPs (based on Fig. 2 of Ref. 65). B, PCR-based replication fidelity assay for DNA containing Z:P nucleobase pairs. If the UBP is lost to a T:A base pair, then a Dra restriction site (red) is generated in the amplicons. Similarly, if the UBP is lost to a C:G base pair, then an Alu restriction site (blue) is produced. C, gel showing misincorporation rates in the PCR amplification of template DNA in the presence of dZTP and/or dPTP components using WT Klentaq and WT Phusion DNA polymerases at different pH values (64). Four parallel PCRs were performed followed by digestion with PspOMI for 16 h. The ratio between the amount of full-length product (FLP) and all oligonucleotides gives the misincorporation rates shown in the figure. Lane 1, negative control PCR amplification of template DNA in the presence of WC dNTPs (200 mM each), followed by digestion with PspOMI. Lane 2, five-letter PCR amplification of template DNA in the presence of WC dNTPs (200 mM each) and dZTP (25 mM), followed by digestion with PspOMI. Lane 3, five-letter PCR amplification of template DNA in the presence of WC dNTPs (200 mM each) and dPTP (25 mM), followed by digestion with PspOMI. Lane 4, six-letter PCR amplification of template DNA in the presence of WC dNTPs (200 mM each), dZTP (25 mM), and dPTP (25 mM), followed by digestion with PspOMI. Taken from Ref. 64 and used with permission.

Table 1

Examples of evolved Taq DNA polymerase variants

Residues in boldface/italic type are found in two or more of the reported Taq DNA polymerase variants. The second column indicates whether variants were generated for specific substrates, such as C2’ modified nucleic acid, or for a specific property, such as thermostability.

| Taq variant | Substrate/Property | Amino acid substitutions | Reference |
|-------------|--------------------|--------------------------|-----------|
| SFM4-3      | C2’ modification   | I614E/E615G/N583S/D656N/E681K/E742Q/M747R | 70        |
| AA40        | RNA, DNA           | E602V/A608V/I614M/E615G | 73        |
| M1          | 3’ mismatches      | G84A/D144G/K314R/E520G/F598I/A608V/E742G | 74        |
| M4          | 3’ mismatches      | D58G, R74P, A109T, L245R, R343G, G370D, E520G, N583S, E694K, A743I | 74        |
| SD4         | Hydrophobic bases  | V62I, Y78H, T88S, P114Q, P264S, E303V, G389V, E424G, E432G, E602G, A608V, E742G, M761T, M775T | 75        |
| H15         | Thermostability    | K225E/E303V/K314R/D370G/M375K | 76        |
| RT-KTQ      | m6-A sequencing    | L459M/S515R/I638Y/M747K | 69        |
| ZP          | Z:P substrates     | M444V/P527A/E551E/E832V | 65        |
A detailed evaluation of their fidelity and incorporation efficiencies (63, 64).

The original CSR strategy, however, requires the replication of the entire polymerase gene, which contains more than 2000 nucleobases. Thus, only the most active variants are recovered. Many technical aspects, including the appropriate choice of oil or surfactant containing the water droplets (79, 80), must be optimized if reproducible results are to be obtained (81).
Improved variations of this selection method have therefore been reported by Philip Holliger and co-workers (78, 82), including compartmentalized self-tagging (CST) (Fig. 4B). In CST, the positive feedback loop depends on the polymerase “tagging” a plasmid containing its encoding gene by extension of a biotinylated oligonucleotide. As a result, plasmids containing genes that encode DNA polymerases with altered substrate selectivity can be enriched by selective capture onto an affinity column. Clearly, the UBP(s) must be located within a segment of plasmid DNA that is replicated with high efficiency and fidelity by the DNA polymerase variant. In addition, the sensitivity of the assay is higher because CST does not require multiple self-replications of the complete gene encoding the polymerase variant.

Insights from structural studies of WT and evolved Klentaq DNA polymerase complexes

In contrast to some other critical cellular enzymes and machines, DNA polymerases have been reinvented throughout evolution (32). As a consequence, several distinct structural classes of polymerases provide multiple opportunities to identify an enzyme that will efficiently and faithfully replicate an unnatural base pair. Replicative DNA polymerases have been classified as families A–D, with D being the most recently identified DNA polymerase family from Archaea (32).

To date, naturally occurring family A and B DNA polymerases have been used to successfully replicate single UBPs. Family A polymerases are found primarily in bacteria; their peptide fold is described as including fingers, palm, and thumb domains associated with polymerase activity (Fig. 5A). They also have 3′→5′ exonuclease and 5′→3′ exonuclease domains, the second removing the RNA primers required for lagging strand synthesis (32). In Klentaq, the Thermus aquaticus DNA polymerase (83), corresponding to the Klenow or large fragment of the Thermus aquaticus DNA polymerase, has been replicated with selectivities exceeding 99% using DeepVent DNA polymerase (family B) (85) or a combination of DeepVent DNA polymerase and Taq DNA polymerase (family A) (14), respectively. It has proven much more difficult, however, to replicate templates containing consecutive hydrophobic nucleobases. The problem appears to be caused, in part, by disruptions to the DNA double helix resulting from the presence of hydrophobic UBPs (86). For example, the dNaM:d5SICS pair adopts an intercalative stacking mode in duplex DNA, as seen in both NMR structures and binary template-primer Klentaq complexes (87, 88). At least 6 natural base pairs must separate ds-Px pairs if replication is to be successful (89); as yet there is no report of DNA containing consecutive NaM:PTP3 (or structurally related hydrophobic UBPs) being replicated. Of course, the necessity of replicating more than one UBP, either consecutive or separated by standard WC nucleobases, depends on the biotechnological application (15).

A truly artificial genome would presumably include consecutive UBPs just as natural genomes include runs of sequential A:T or G:C pairs; cells possessing such a genome would require a DNA polymerase that can replicate through these regions of sequence. We have shown that up to 4 consecutive Z:P pairs can be incorporated by WT Klentaq polymerase (63); these UBP tetradis do not adversely impact the ability of duplex DNA to adopt A- and B-conformations (90, 91). Unfortunately, the efficiency of Z:P incorporation by WT Klentaq is only 99.8% per theoretical PCR cycle (63), meaning that the UBP is easily lost during PCR-based applications. An engineered Klentaq variant (ZP Klentaq), exhibiting improved fidelity and Z:P incorporation efficiency was, however, obtained using CSR (Fig. 4A) (65). Remarkably, ZP Klentaq contains only four amino acid substitutions, M444V, P527A, D551E, and E832V, all of which are located distal from the active site (Fig. 5A). In family A DNA polymerases, a large conformational change in the fingers domain occurs upon binding of the complementary dNTP, resulting in the formation of a closed complex that facilitates incorporation (Fig. 5). Of the structurally characterized DNA polymerases related to Klentaq at the sequence level (Table 2), only three have been captured in both pre- and post-incorporation complexes: ZP Klentaq (92), WT Klentaq (6), and WT Geobacillus kaustophilus DNA polymerase (93). Incorporation of nucleotides by the Geobacillus enzyme exhibits a fingers domain closure angle of 37°. In contrast to the Geobacillus enzyme, incorporation of natural dNTPs by WT Klentaq involves a much larger conformational change in which the fingers domain rotates by ∼59° (Fig. 5, B and C). In the closed conformation of the polymerase, correctly paired nucleotides are selected by hydrogen bonding and size complementarity. Despite retaining these features, Z:P pairs are more efficiently and faithfully incorporated by ZP Klentaq, in which the fingers close down by ∼64° (Table 3). Protein-nucleic acid interactions in both pre- and post-incorporations for ZP Klentaq are similar to those observed in the analogous WT complexes, suggesting that the evolved polymerase closely mimics the WT enzyme in this regard (Fig. 5, D and E).

The post-incorporation ZP Klentaq complex (PDB entry 5W6Q) (92) also exhibits increased thumb domain motion (∼4°) relative to the analogous WT Klentaq complex (PDB entry 3SZ2) (6). Although Z:P pairs are readily accommodated in both B- and A-form DNA duplexes (90, 91) and retain standard features (groove widths, base stacking parameters, etc.) consistent with these helical forms, modeling of a post-incorporation complex of P:Z bound to the WT Klentaq suggested that some structural alterations within the active site would be required to accommodate the UBP (92). The two base pairs closest to the site of incorporation of the next dNTP exhibit A-form, which is significantly wider than B-form. This feature of the active site may allow WT Klentaq to accommodate a single hydrophobic UBP, which is wider than hydrogen-bonded pairs (6, 94).

Of the amino acid substitutions in ZP Klentaq, M444V was considered to be the best candidate for providing access to increased relative domain motion as compared with the WT
Figure 5. Structural properties of Klentaq DNA polymerase. A, the open form of ZP Klentaq (from PDB entry 5W6Q), shown in a cartoon rendering, includes fingers (blue), palm (green), thumb (yellow), and 3′–5′ exonuclease (red) domains. This architecture is conserved in all family A polymerases. Variant amino acid substitutions are shown as sphere models (orange, carbon; blue, nitrogen; red, oxygen), M444V, P527A, and D551E. (The fourth amino acid substitution, E832V, is disordered in the structures). α-Helices G, H, and I contribute to dynamic motion in ZP Klentaq. B, a similar rendering of ZP Klentaq in the binary, post-incorporation complex (PDB entry 5W6Q) is shown with template-primer, including P:Z in the active site (magenta cartoon rendering). C, the pre-incorporation complex of ZP Klentaq (PDB entry 5W6K) is shown superimposed on the post-incorporation complex from B. The black arrow indicates the most significant conformational change associated with the fingers domain (light blue for the initial and post-incorporation and blue for the pre-incorporation complex). The O helix closes down, forming the active site in the ternary complex with bound dZTP (magenta stick rendering). The thumb domains show a slight shift in position (light orange for post-incorporation and yellow for pre-incorporation complex). The palm and exonuclease domains show no significant differences in overall position in the two complexes. Close-ups of hydrogen-bonding interactions in the active site for the WT Klentaq bound to ddCTP-dG and dZTP-dP are shown in D and E, respectively. In each case, the dNTP is positioned through a water-mediated hydrogen-bonding interaction involving O2 (minor groove edge of the nucleobase) with Gln-754. Similarly, the penultimate primer nucleobase presents either N3 or O2 for hydrogen-bonding to Arg-573 (again a minor groove interaction). In the case of the ZP Klentaq ternary complex (E), Arg-660 hydrogen-bonds to the NO2 group present on the major groove edge of Z. F, the active site including template-primer and dZTP (magenta stick rendering), Mg2+ ions (orange spheres), and coordinating residues Asp-610 and Asp-785 (green stick rendering) are shown with a ribbon rendering of the enzyme, palm in green and fingers in blue.

Table 2
Structure-based alignments of Taq-related DNA polymerases

| PDB  | Type | Source                  | No. of aligned Ca atoms | RMSD  | Sequence identity |
|------|------|-------------------------|-------------------------|-------|------------------|
| 5DKU | Apo  | Plasmodium              | 418/521                 | 2.8   | 42.8             |
| 4DQS | Binary | G. kaustophilus        | 486/521                 | 1.8   | 43.4             |
| 1KFD | Apo  | E. coli                 | 429/521                 | 1.9   | 42.7             |
| 2BPD | Binary | Geobacillus stearothermophilus | 488/521        | 1.8   | 43.4             |
| 4XV1 | Binary | Homo sapiens (Pol α) | 439/521                 | 2.5   | 31.9             |
| 6VDE | Apo  | Mycobacterium smegmatis | 487/521                 | 2.1   | 40.0             |
| 4 × 0Q | Ternary | H. sapiens (Pol θ) | 443/522                 | 2.4   | 29.6             |
enzyme. Residue 444 resides within the hydrophobic core of the palm domain, which serves as the command center for the enzyme. Both the fingers and thumb domains are connected to the palm domain. In addition, the essential catalytic residues Asp-610 and Asp-785, which coordinate Mg$^{2+}$ and the incoming dNTP, are in the active site. Residues 444 and 551, which are in the hydrophobic core of the palm domain, are replaced by valine and glutamate, respectively. Substitution of a Val for Met at this position creates open space within the core that could potentially translate into increased motion of the fingers and thumb domains.

### Augmenting crystallography with molecular dynamics simulations of natural and evolved DNA polymerases

Molecular dynamics (MD) simulations are a well-validated method of studying protein dynamics (95) and have yielded important information about conformational changes undergone by DNA polymerases (96, 97). Our recent work also demonstrates that MD simulations can be used to understand how substitutions far from the active site might alter substrate selectivity (98). For example, microsecond trajectories have shown how the dynamic motions of ZP Klentaq differ from those of WT Klenow in the binary complex containing template-primer duplexes (Fig. 5f). They can also highlight correlated motions in networks of residues that are critical to altered domain motions in the variant DNA polymerase (Fig. 6B) (98), which permit the engineered enzyme to bind WC and Z:P-containing template-primer duplexes in an equivalent fashion, thereby increasing the efficiency of UBP incorporation. This seems to be a consequence of replacing Met-444 and Asp-551 at the base of the thumb domain by valine and glutamate, respectively, which allows the fingers, palm, and thumb domains in ZP Klentaq to move into position about the AEGIS template-primer duplex. Thus, replacing Met-444 by valine in the hydrophobic core changes the populated interactions with adjacent residues (Phe-564, Met-779, and Leu-780) in the hydrophobic core (Fig. 6C) with the consequence that the G helix becomes more flexible in the variant compared with the WT polymerase (98).

As well as modeling how amino acid substitutions might impact long-time-scale motions, such as domain reorientations, MD simulations can illuminate specific molecular changes that have the greatest impact on the altered catalytic biophysical and catalytic properties of the DNA polymerase variant. In the case of ZP Klentaq, the role of each of the substitutions (M444V, P527A, D551E, or E832V) was determined using MD simulations of the four single-point Klentaq variants (98). These calculations showed that replacing Met-444 in the hydrophobic core of the palm domain with valine makes a significant contribution to altering dynamic motions in ZP Klentaq. In addition, changing Asp-551 to glutamate alters the interactions of the H and I helices at the base of the thumb domain, which affects G helix flexibility (Fig. 6B) (98). Care must be taken to perform MD simulations on the correct form of the enzyme. For example, introducing valine in place of Glu-832, which breaks a salt bridge with Arg-596, has little impact on the dynamics of the ZP Klentaq/DNA binary complex. However, it does alter protein/DNA interactions. This substitution may therefore exert its effects in the pre-incorporation (ternary) complex during dNTP binding and incorporation.

MD simulations have also provided interesting insights into the “closing mechanism” of DNA polymerase I prior to the incorporation of WC nucleobases (102) and the temperature-dependent activity of a Klentaq variant in which Ile-707, located 20 Å away from the active site, is replaced by leucine (103). This I707L Klentaq variant exhibits low activity at 37 °C but efficiently replicates DNA at 68 °C, permitting improved PCR amplification of difficult sequences. MD simulations show that the I707L Klentaq variant is almost immobile at 37 °C but exhibits increased flexibility at 68 °C. In contrast, the WT enzyme exhibits less pronounced dynamic changes in simulations performed at 37 and 68 °C. Thus, it is likely that the reduced activity of the I707L Klentaq variant is associated with the enzyme remaining in its closed conformation at lower temperatures (i.e. dNTP binding is blocked) and that the pre-incorporation complex cannot form. As observed for ZP Klentaq (98), the molecular origins of this behavior could be traced to changes in the hydrophobic core; replacing isoleucine by leucine results in an alternate conformation of an adjacent phenylalanine (Phe-749), thereby repositioning the O and O1 helices. As a consequence, the active site becomes filled with nucleobases in the template overhang, thereby blocking entry of the incoming dNTP.

### Outlook

Realizing the full potential of UBPs in the creation of new research, diagnostic, and therapeutic tools will ultimately depend on our ability to identify DNA polymerase variants that can faithfully and efficiently replicate those UBPs. Many of
these applications will necessarily require enzymes that are active at the high temperatures used in PCR, limiting the pool of candidate DNA polymerases primarily to family A and family B members. For other applications, such as using dNaM-dTPT3 (Fig. 1) in semi-synthetic organisms that introduce non-canonical amino acids into specific proteins or the isothermal amplification of AEGIS UBPs (104), DNA polymerases that function at 37 °C, such as those in *E. coli*, will be adequate as starting points for structure-based engineering or library construction.

To date, progress in engineering DNA polymerases has been encouraging (105). In our work aimed at improving the ability of these enzyme to replicate UBPs (65, 92, 98), ZP Klentaq takes advantage of increased domain flexibility to improve the replication efficiency of Z:P pairs, even though both WT Klentaq and ZP Klentaq grip the template-primer duplex in a similar manner prior to and following incorporation of the AEGIS UBP.

Combining X-ray crystallography with MD simulations yields two important conclusions. First, not all substituted residues impact the properties of the evolved enzyme equally. Second, substitution of a key residue, such as Met-444, can result in increased relative domain motion, even in an enzyme that exhibits a very large conformational change during catalysis. It is this altered motion that allows ZP Klentaq to incorporate UBPs more readily than the WT enzyme.

CSR, CST, and related strategies for the directed evolution of DNA polymerases, which typically produce variants containing several amino acid substitutions, might therefore be improved by targeting fewer residues outside the active site for variation. Coupling structure determinations, MD simulations, and enzymatic characterization will provide guiding principles for CSR.
library design and facilitate rational approaches to optimize family A DNA polymerases to replicate multiple UBPs. Such work can be guided by understanding amino acid substitutions already identified in Klentaq variants (Table 1); for example, Asn-583, Ile-614, and Met-747 are substituted in three of eight variant enzymes, whereas Glu-520, Glu-602, Ala-609, Glu-615, and Glu-742 are substituted in two of eight variants. None of these residues are involved in direct hydrogen-bonding interactions with the substrate template-primer or dNTP in the active site (Fig. 7). The past evolutionary history of DNA polymerases may also prove useful for selecting specific sites for variation (42, 106).

We also note that whereas most engineering efforts to date have focused exclusively on improving the ability of DNA polymerases to incorporate UBPs with little consideration for exonuclease activity, a Thermococcus gordonarius DNA polymerase variant selected to replicate xeno-nucleic acids contains two amino acid substitutions that inactivate the 3′−5′ exonuclease site (107). In theory, incorporation of UBPs that slow down polymerization would allow the substrate to be positioned within the exonuclease active site for proofreading. Whether the exonuclease would efficiently process the newly incorporated unnatural nucleobase or not has yet to be fully investigated. It is possible that engineered DNA polymerases may have to possess optimized exonuclease function as well as improved efficiencies for UBP incorporation to maintain artificial genomes.

In summary, the next generation of useful DNA polymerases must (i) replicate more than one UBP with rates and fidelities similar to those observed for WT DNA polymerases when replicating WC dNTPs and (ii) exhibit no significant pausing following UBP incorporation to avoid invoking proofreading mechanisms. These properties will permit the fidelity of these engineered enzymes to approach 99.99%, allowing them to generate full-length replication products in applications involving multiple rounds of PCR. We suggest a strategy for obtaining these DNA polymerase variants that builds on studies of ZP Klentaq, in which a combination of X-ray crystallography and computer simulations showed the importance of a few key residues for conferring increased dynamic motion in the thumb and fingers domains, thereby improving UBP incorporation efficiency. In this approach, altering the properties of an evolved DNA polymerase variant by introducing additional amino acid substitutions will take advantage of a fundamental, structure-based understanding of how dynamic changes in residue “networks” impact kinetic properties (108, 109).

Funding and additional information—This work was supported by Biotechnology and Biological Sciences Research Council, UK, Grant P/018017/1 (to N. G. J. R.). This review is also based on work supported by the National Science Foundation under Grant MCB-1939086 (to M. M. G. and S. A. B.). Research reported in this publication was also supported by the National Institutes of Health under Director’s Award R01GM128186 (to S. A. B.). The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

Conflict of interest—S. A. B. owns intellectual property associated with AEGIS nucleobases. Some of the compounds mentioned in this article are sold by Firebird Biomolecular Sciences, LLC, which is owned by S. A. B.

Abbreviations—The abbreviations used are: UBP, unnatural base pair; Taq, Thermus aquaticus DNA polymerase; Klentaq, the Klentaq fragment of Taq DNA polymerase; ZP Klentaq, an evolved Klentaq DNA polymerase variant; WC, Watson-Crick; AEGIS, artificially expanded genetic information system; CSR, compartmentalized self-replication; CST, compartmentalized self-tagging; MD, molecular dynamics; PDB Protein Data Bank.

References
1. Goodman, M., and Tippin, B. (2000) The expanding polymerase universe. Nat. Rev. Mol. Cell Biol. 1, 101–109 CrossRef Medline
2. Young, R. A. (1991) RNA polymerase II. Annu. Rev. Biochem. 60, 689–715 CrossRef Medline
3. Autexier, C., and Lue, N. F. (2006) The structure and function of telomerase reverse transcriptase. Annu. Rev. Biochem. 75, 493–517 CrossRef Medline
4. Aschenbrenner, J., and Marx, A. (2017) DNA polymerases and biotechnological applications. Curr. Opin. Biotechnol. 48, 187–195 CrossRef Medline
5. Gardner, A. F., Jackson, K. M., Boyle, M. M., Buss, J. A., Potapov, V., Gehring, A. M., Zatopek, K. M., Corrêa, I. R., Jr., Ong, J. L., and Jack, W. E. (2019) Therminator DNA polymerase: modified nucleotides and unnatural substrates. Front. Mol. Biol. 6, 28 CrossRef Medline
6. Betz, K., Malyshev, D. A., Lavergne, T., Welte, W., Diederichs, K., Dwyer, T. J., Ordoukhian, P., Romesberg, F. E., and Marx, A. (2012) Klentaq polymerase replicates unnatural base pairs by inducing a Watson-Crick geometry. Nat. Chem. Biol. 8, 612–614 CrossRef Medline
7. Watson, J. D., and Crick, F. H. C. (1953) Molecular structure of nucleic acids. A structure for deoxyribonucleic acid. Nature 171, 737–738 CrossRef Medline
8. Kool, E. T. (2002) Replacing the nucleobases in DNA with designer molecules. Acc. Chem. Res. 35, 936–943 CrossRef Medline
discriminates Zika, chikungunya and dengue viruses in mosquitoes. BMC Infect. Dis. 19, 418 CrossRef Medline

30. Raper, A. T., Reed, A. J., and Sano, Z. (2018) Kinetic mechanism of DNA polymerases: contributions of conformational dynamics and a third metal ion. Chem. Rev. 118, 6000–6025 CrossRef Medline

31. Wu, W.-J., Yang, W., and Tsai, M.-D. (2017) How DNA polymerases catalyse replication and repair with contrasting fidelity. Nat. Rev. Chem. 1, 68 CrossRef

32. Raia, P., Delarue, M., and Sauquet, L. (2019) An updated structural classification of replicative DNA polymerases. Biochem. Soc. Trans. 47, 239–249 CrossRef Medline

33. Steitz, T. A. (1999) DNA polymerases: structural diversity and common mechanisms. J. Biol. Chem. 274, 17395–17398 CrossRef Medline

34. Johnson, K. A. (2010) The kinetic and chemical mechanism of high-fidelity DNA polymerases. Biochim. Biophys. Acta 1804, 1041–1048 CrossRef Medline

35. Kieber, J. R., Mao, C., Braman, J. C., and Beece, L. S. (1998) Visualizing DNA replication in a catalytically active Bacillus DNA polymerase crystal. Nature 391, 304–307 CrossRef Medline

36. Kropp, H. M., Dürr, S. L., Peter, C., Diederichs, K., and Marx, A. (2018) Snapshots of a modified nucleotide moving through the confines of a DNA polymerase. Proc. Acad. Natl. Acad. Sci. U. S. A. 115, 9992–9997 CrossRef Medline

37. Joyce, C. M., and Benkovic, S. J. (2004) DNA polymerase fidelity: kinetics, structure, and checkpoints. Biochemistry 43, 14317–14324 CrossRef Medline

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

39. Chen, T., and Romesberg, F. E. (2014) Directed polymerase evolution. FEBS Lett. 588, 219–229 CrossRef Medline

40. Laos, R., Thomson, J. M., and Benner, S. A. (2014) DNA polymerases engineered by directed evolution to incorporate non-standard nucleotides. Front. Microbiol. 5, 565 CrossRef Medline

41. Betz, K., and Marx, A. (2020) The structural basis for processing of unnatural base pairs by DNA polymerases. Chem. Eur. J. 26, 3446–3463 CrossRef Medline

42. Fowler, R. G., Degnen, G. E., and Cox, E. C. (1974) Mutational specificity of a conditional Escherichia coli mutator, mutDS. Mol. Gen. Genet. 133, 179–191 CrossRef Medline

43. Yu, H., and Goodman, M. F. (1992) Comparison of HIV-1 and avian myeloblastosis virus reverse transcriptase fidelity on RNA and DNA templates. J. Biol. Chem. 267, 10888–10896 Medline

44. Brambati, A., Barry, R. M., and Sfeir, A. (2020) DNA polymerase theta (Polθ)—an error-prone polymerase necessary for genome stability. Curr. Opin. Genet. Dev. 60, 119–126 CrossRef Medline

45. Du, H., Wang, P., Wu, J., He, X., and Wang, Y. (2020) The roles of polymerases μ and θ in replicative bypass of O^6- and N^2-alkyl-2'-deoxyguanosine lesions in human cells. J. Biol. Chem. 295, 4556–4562 CrossRef Medline

46. Hendrickson, C., Devine, K., and Benner, S. A. (2004) Probing minor groove recognition contacts by DNA polymerases and reverse transcriptases using 3-deaza-2'-deoxycytidine. Nucleic Acids Res. 32, 2241–2250 CrossRef Medline

47. Benner, S. A. (2004) Understanding nucleic acids using synthetic chemistry. Acc. Chem. Res. 37, 784–797 CrossRef Medline

48. Joyce, C. M., and Steitz, T. A. (1994) Function and structure relationships in DNA polymerases. Annu. Rev. Biochem. 63, 777–822 CrossRef Medline

49. Kuchta, R. D., Benkovic, P., and Benkovic, S. J. (1988) Kinetic mechanism whereby DNA polymerase-1 (Klenow) replicates DNA with high fidelity. Biochemistry 27, 6716–6725 CrossRef Medline
deoxy-, dideoxy-, and ribonucleotides. J. Biol. Chem. 287, 28215–28226

94. Betz, K., Kimoto, M., Diederichs, K., Hirao, I., and Marx, A. (2017) Structural basis for expansion of the genetic alphabet with an artificial nucleobase pair. Angew. Chem. Int. Ed. Engl. 56, 12000–12003

95. Dror, R. O., Dirks, R. M., Grossman, J. P., Xu, H., and Shaw, D. E. (2012) Biomolecular simulation: a computational microscope for molecular biology. Annu. Rev. Biophys. 41, 429–452

96. Walker, A. R., and Cisneros, G. A. (2017) Computational simulations of DNA polymerases: detailed insights on structure/function/mechanism from native proteins to cancer variants. Chem. Rev. Toxicol. 30, 1922–1935

97. Meli, M., Sustarsic, M., Craggs, T. D., Kapanidis, A. N., and Colombo, G. (2016) DNA polymerase conformational dynamics and the role of fidelity-conferring residues: insights from computational simulations. Front. Mol. Biosci. 3, 20

98. Ouaray, Z., Singh, I., Georgiadis, M. M., and Richards, N. G. J. (2020) Building better enzymes: molecular basis of improved non-natural nucleobase incorporation by an evolved DNA polymerase. Protein Sci. 29, 455–468

99. Skjaerven, L., Martinez, A., and Reuter, N. (2011) Principal component and normal mode analysis of proteins; a quantitative comparison using the GroEL subunit. Proteins 79, 232–243

100. Taylor, D., Cawley, G., and Hayward, S. (2014) Quantitative method for the assignment of hinge and shear mechanism in protein domain movements. Bioinformatics 30, 3189–3196

101. Grant, B. J., Rodrigues, A. P. C., El Sawy, K. M., McCammon, J. A., and Caves, L. S. D. (2006) Bio3d: an R package for the comparative analysis of protein structures. Bioinformatics 22, 2695–2696

102. Miller, B. R., 3rd, Beece, L. S., Parish, C. A., and Wu, E. Y. (2015) The closing mechanism of DNA polymerase I at atomic resolution. Structure 23, 1609–1620

103. Modesto, E., Mawby, L., Miller, B., III, Wu, E., and Parish, C. A. (2019) A molecular dynamics investigation of the thermostability of cold-sensitive 1707L KlenTaq1 DNA polymerase and its wild-type counterpart. J. Chem. Inf. Model. 59, 2423–2431

104. Yang, Z., Kim, H.-J., Le, J. T., McLendon, C., Bradley, K. M., Kim, M.-S., Hutter, D., Hoshika, S., Yaren, O., and Benner, S. A. (2018) Nucleoside analogs to manage sequence divergence in nucleic acid amplification and SNP detection. Nucleic Acids Res. 46, 5902–5910

105. Nikoomanzar, A., Chim, N., Yik, E. J., and Chaput, J. C. (2020) Engineering polymerases for applications in synthetic biology. Q. Rev. Biophys. 53, e8

106. Chen, F., Gaucher, E. A., Leal, N. A., Hutter, D., Havemann, S. A., Govindarajan, S., Ortlund, E. A., and Benner, S. A. (2010) Reconstructed evolutionary adaptive paths give polymerases accepting reversible terminators for sequencing and SNP detection. Proc. Natl. Acad. Sci. U. S. A. 107, 1948–1953

107. Houlihan, G., Arangundy-Franklin, S., Porebski, B. T., Subramanian, N., Taylor, A. I., and Holliger, P. (2020) Discovery and evolution of RNA and Xna reverse transcriptase function and fidelity. Nat. Chem. 12, 683–690

108. Hu, S. S., Offenbacher, A. R., Lu, E. D., and Klinman, J. P. (2019) Comparative kinetic isotope effects on first- and second-order rate constants of soybean lipoxygenase variants uncover a substrate-binding network. J. Biol. Chem. 294, 18069–18076

109. Hu, S. S., Offenbacher, A. R., Thompson, E. M., Gee, C. L., Wilcoxen, J., Carr, C. A. M., Prigozhin, D. M., Yang, V., Alber, T., Brit, R. D., Fraser, J. S., and Klinman, J. P. (2019) Biophysical characterization of a disabled double mutant of soybean lipoygenase: the “undoing” of precise substrate positioning relative to metal cofactor and an identified dynamical network. J. Am. Chem. Soc 141, 1555–1567
