Published in final edited form as:

J Am Chem Soc. 2009 October 21; 131(41): 14827–14837. doi:10.1021/ja9039696.

Evolving a polymerase for hydrophobic base analogues

David Loakes¹, José Gallego^{1,3}, Vitor B. Pinheiro¹, Eric T. Kool², and Philipp Holliger^{1,*}

¹ MRC Laboratory of Molecular Biology, Hills Road, Cambridge CB2 0QH, UK

² Stanford University, Department of Chemistry, Stanford, CA 94305, USA

Abstract

Hydrophobic base analogues (HBAs) have shown great promise for the expansion of the chemical and coding potential of nucleic acids but are generally poor polymerase substrates. While extensive synthetic efforts have yielded examples of HBAs with favourable substrate properties, their discovery has remained challenging. Here we describe a complementary strategy for improving HBA substrate properties by directed evolution of a dedicated polymerase using compartmentalized self-replication (CSR) with the archetypal HBA 5-nitroindole (d5NI) and its derivative 5-nitroindole-3-carboxamide (d5NIC) as selection substrates. Starting from a repertoire of chimeric polymerases generated by molecular breeding of DNA polymerase genes from the genus *Thermus*, we isolated a polymerase (5D4) with a generically-enhanced ability to utilize HBAs after five rounds of CSR selection. 5D4 was able to form and extend d5NI and d5NIC (d5NI(C)) self-pairs as well as d5NI(C) heteropairs with all four bases with efficiencies approaching, or exceeding, those of the cognate Watson-Crick pairs, despite significant distortions caused by the intercalation of the d5NI(C) heterocycles into the opposing strand base stack, as shown by nuclear magnetic resonance spectroscopy (NMR). Unlike Taq polymerase, the selected polymerase 5D4 was also to extend HBA pairs such as Pyrene: ϕ (abasic site), d5NI: ϕ , and isocarbostyryl (ICS): 7-azaindole (7AI), allowed bypass of a chemically diverse spectrum of HBAs and enabled PCR amplification with primers comprising multiple d5NI (C)-substitutions, while maintaining high levels of catalytic activity and fidelity. The selected polymerase 5D4 promises to expand the range of nucleobase analogues amenable to replication and should find numerous applications, including the synthesis and replication of nucleic acid polymers with expanded chemical and functional diversity.

Keywords

Hydrophobic base analogue; polymerase; directed evolution; 5-nitroindole; universal base

DNA has unique properties beyond its ability to encode genetic information, which make it an attractive supramolecular scaffold for chemistry, biotechnology and nanotechnology.¹ Despite its polyanionic backbone it can fold into compact molecular structures forming specific receptors (aptamers) and catalysts^{2,3}, it can be assembled into complex nanostructures according to the well-understood rules of Watson-Crick base-pairing² and polymer strands of precisely defined length and sequence can be synthesized, replicated and evolved using DNA polymerases. However, the physicochemical properties of the canonical four bases span only a narrow range. Expanding the chemistry of nucleic acid polymers amenable to synthesis,

ph1@mrc-lmb.cam.ac.uk.

³present address: Centro de Investigación Príncipe Felipe, Avda. Autopista del Saler 16, 46012 Valencia, Spain

SUPPORTING INFORMATION PARAGRAPH Protein sequences of selected polymerases (including 5D4), additional data on kinetic constants, substrate specificity, primer extension, NMR spectroscopy, reversion mutant activity, SCA and FoldX analysis can be found in supporting information.

replication and evolution would greatly enhance the phenotypic diversity and widen their biotechnological and clinical potential.

Hydrophobic base analogues (HBAs) potentially could provide a variety of attributes not present in the canonical bases including photoactivated chemistry, fluorescence and the potential to form novel nucleic acid structures through non-canonical stacking and hydrophobic interactions. HBAs have already found a wide range of applications in nucleic acid manipulation and hybridization and as steric or fluorescent probes of enzyme dynamics and function.¹

In the context of nucleic acid replication, HBAs were originally studied as universal base analogues^{5–7} but have been found to have a number of other intriguing properties. For example, hydrophobic isosteres of the natural bases, were found to display specific pairing with the natural bases clarifying the importance of steric complementarity for replication fidelity.¹ Other HBAs were found to form specific self- and hetero-pairs with the potential to form an orthogonal third base-pair. For example, Romesberg and colleagues have systematically explored a large range of chemical space including substituted phenyl,^{3, 4} pyridyl,^{5, 6} and isocarbostyryl,^{7–9} as well as pyridones,¹⁰ azaindoles¹¹ and other heterocycles¹² systematically probing the requirements for polymerase recognition, and for formation of specific HBA self- and hetero-pairs. Hirao and Yokoyama have also designed a number of HBA base pairing systems based on shape complementarity and steric exclusion and some have been shown to be sufficiently orthogonal to specifically direct the incorporation of fluorophores,^{13, 14} biotin,¹⁵ iodine useful for photocrosslinking to proteins¹⁶ into RNA transcripts and to direct the specific incorporation of the unnatural amino acid 3-chlorotyrosine in a *E. coli* cell free translation system.^{17, 18} Finally, large HBAs such as pyrene¹⁹ or other derivatives of indole²⁰ have been found to be able to “detect” DNA damage by forming highly specific “base-pairs” with abasic sites. HBAs have thus shown clear potential to expand the chemical, functional and coding capabilities of nucleic acids for therapeutic, diagnostic or nanotechnological applications. However, their general application has been restricted by generally poor replication by natural polymerases.

Extensive synthetic efforts have been made to improve the properties of HBAs as polymerase substrates (see above) by optimizing steric fit^{1, 21}, inclusion of minor groove H-bond acceptors^{22, 23} and systematic derivatization with heteroatoms and alkyl substituents²⁴. While significant progress has been made²¹, the identification of HBAs compatible with efficient enzymatic replication has remained challenging. An alternative strategy would be the engineering of the polymerase active site for improved HBA replication. Polymerases have been engineered by design^{25, 26}, screening^{27, 28} and selection^{29–31}. These studies have uncovered significant plasticity in the polymerase active site for the acceptance of non-cognate chemistries^{32–34}. Indeed, Romesberg *et al* have successfully applied their phage-based selection system to the evolution of a variant of the Stoffel fragment of *Taq* polymerase with 30-fold improved extension of the HBA PICS:PICS self-pair.³⁴

We have developed an alternative strategy for the evolution of polymerases, called “compartmentalized self-replication” (CSR).³¹ CSR is based on a simple feedback loop, in which a polymerase replicates only its own encoding gene with compartmentalization into the aqueous compartments of a water-in-oil (w/o) emulsion³⁵ serving to isolate individual self-replication reactions from each other. Thus each polymerase replicates only its own encoding gene to the exclusion of those in other compartments (i.e. self-replicates). In such a system adaptive gains directly (and proportionally) translate into genetic amplification of the encoding gene.

Here we describe the application of CSR to the directed evolution of polymerases for HBA replication. Using flanking primers modified with the archetypal HBA 5-nitroindole (d5NI)³⁶ and its 3-carboxamide derivative (d5NIC), we performed CSR selections and isolated a polymerase (5D4) with a generically improved ability to incorporate, extend and bypass a variety of HBAs. These include a wide variety of non-cognate substrates including HBA self-pairs as well as hetero-pairs with natural bases, abasic sites or other HBAs. Remarkably, the selected polymerase 5D4 is able to process these analogues regardless of a lack of minor groove H-bonding and despite significant distortions to the primer-template duplex structure caused by intercalation into the opposing strand base stack, as shown here for d5NI and d5NIC.

Results

1. Substrate design and selection

As the HBA substrate for polymerase evolution we chose the universal base analogue 5-nitroindole (d5NI). d5NI has found many uses in hybridization applications.³⁷ However, typically for HBAs, d5NI lacks H-bond donors or acceptors and is a very poor substrate for enzymatic replication³⁸. Indeed, initial selection experiments were unsuccessful (not shown). In order to enable directed evolution, we synthesized several d5NI derivatives (not shown) to improve substrate characteristics. Among these, 5-nitroindole-3-carboxamide (d5NIC) (Fig. 1a) showed promise: while virtually indistinguishable from d5NI in its effects on oligonucleotide melting temperature,³⁹ d5NIC proved a superior polymerase substrate, allowing some bypass, where d5NI stalled synthesis by *Taq* polymerase (Fig. 1b) and allowing selections to proceed (see below).

We initiated polymerase selection by compartmentalized self-replication (CSR)³¹ starting from the previously described polymerase library (3T)⁴⁰ prepared by molecular breeding of the *polA* genes from three members of the genus *Thermus* (*Taq* (*T. aquaticus*), *Tth* (*T. thermophilus*), *Tfl* (*T. flavus*)).⁴⁰ To enhance and focus polymerase activity to the chosen substrates (HBAs), we modified CSR selection to include flanking primers comprising HBAs (d5NIC) both at their 3'-ends and (for later rounds of selection) internally (Fig. 2a), thus requiring not only HBA extension but also bypass of template HBAs for efficient self-replication.

By round two the selected polymerase population had adapted sufficiently to d5NIC that selection with 3'-d5NI primers, too, yielded a positive selection signal (not shown). For rounds three to five we therefore carried out selections using both 3'-d5NI- respectively d5NIC-modified primers and further increased selection stringency by using primers bearing internal as well as 3'-d5NI(C) substitutions. After five rounds, the selected population comprised a diverse set of chimeric polymerases. Many of these had a strikingly improved ability to utilize d5NIC- and to a lesser extent d5NI. In particular polymerase 5D4 was able to efficiently bypass template d5NI and d5NIC (Fig. 3a) as well as perform PCR with d5NIC (and to a lesser extent d5NI)-modified primers (Fig. 3b).

Most of the selected polymerases are *Tth/Taq* chimeras and share an arrangement of gene segments, whereby the N-terminal region (comprising part of the 5'-3' exonuclease domain) derived from *Tth*, whereas the protein core derives mainly from *Taq* as observed previously⁴⁰ (supporting information Fig. 1). We chose a round five polymerase 5D4, for detailed investigation. 5D4 is a *Tth/Taq* chimera with 14 additional mutations from the *Taq/Tth* consensus ((V62I, Y78H, T88S, P114Q, P264S, E303V, G389V, E424G, E432G, E602G, A608V, I614M, M761T, M775T) some of which are shared among other selected polymerases ((Fig. 4, supporting information Table 1).

2. Polymerase kinetics and specificity

We determined single nucleotide incorporation kinetics (k_{cat}/K_m) and relative efficiencies $f_{5D4/Taq}$ for wtTaq and 5D4 using a gel-based assay⁴² (Fig. 4a). While Taq and 5D4 displayed very similar kinetic efficiencies for the incorporation of dTTP opposite dA or dGTP opposite dC ($f_{5D4/Taq} = 2.6$ and 1.4 respectively), 5D4 showed up to 680-fold improved incorporation kinetics of d5NICTP opposite the natural bases or incorporation of dNTPs opposite template d5NIC, approaching, or in some cases exceeding, the efficiencies of formation of the canonical base-pairs. Formation of the [d5NI]₂ or [d5NIC]₂ self-pairs by 5D4 was improved by 35 and 310-fold respectively, the latter exceeding the efficiency of formation of a dT-dA pair by the same polymerase. Indeed, all d5NI-dN/d5NIC-dN heteropairs and the [d5NI(C)]₂ self-pairs were formed with similar efficiencies to the formation of the canonical base-pairs (Fig. 5). However, formation of reverse dN-d5NI/dN-d5NIC heteropairs (incorporation of dNTPs opposite template d5NI/d5NIC) only reached maximal 5% of the efficiency of the formation of canonical base-pairs, despite an up to 110-fold improvement by 5D4 compared to wtTaq. Contributions to the improved catalytic efficiency of 5D4 arise from changes in both k_{cat} and K_m , with reductions in K_m for binding d5NITPs/d5NICTPs particularly striking (supporting information Table 2).

Once formed, basepairs involving HBAs need to be extended. While Taq could not extend either of the d5NI(C) self-pairs nor any of the d5NI(C)-dN heteropairs to any detectable extent, 5D4 could efficiently extend both the d5NIC self-pair (Fig. 6a) as well as the d5NI(C)-dN heteropairs (Fig. 6b). Extension of the d5NI selfpair, however, was very inefficient even for 5D4 (supporting information Fig. 2a). The d5NI(C) self-pairs as well as d5NI(C)-purine heteropairs are reminiscent of 3' transversion mismatches, which are very poorly extended by Taq⁴³. However, we found that 5D4 could efficiently extend a A G transversion mismatch (supporting information Fig. 2b).

Extension of d5NI(C)-dN heteropairs by 5D4 depended not only on the identity of the template strand base (dN) ("paired" with d5NI(C)) but also (to a lesser extent) its 5'-neighbour (dN+1). To comprehensively determine extension efficiencies, we carried out extension reactions using 3'-d5NI(C) primers in all 16 possible sequence contexts. It was found that d5NI is extended best when it is paired with dC and to a lesser extent dA, whilst d5NIC extends best when paired with dT followed by dC (Fig. 6b). Little extension was observed, when either analogue was paired with dG. Less striking preferences were observed for N+1, with d5NI generally preferring dT and d5NIC, dA or dC. As mentioned above extension of both 3'-d5NI and 3'-d5NIC (3'-d5NI(C)) either as self-pair or as d5NI(C)-dN heteropair by the wtTaq polymerase was so inefficient that no reliable kinetic constants could be deduced, while extension by 5D4 proceeded with between 1–10% the efficiency of extension of natural base-pairs pair (supporting information Table 3).

Together, the remarkable improvements in catalytic efficiency of 5D4 with the incorporation, extension and bypass of d5NI and d5NIC allowed PCR amplifications using primers comprising d5NI or d5NIC substitutions both at their 3'-ends as well as d5NIC internally (see, Fig. 3b). Cloning and sequencing of PCR products allowed the determination of the coding potential of d5NI and d5NIC when replicated by 5D4 in different sequence contexts. Template d5NI predominantly directed the incorporation of dA (87%), while template d5NIC mainly templated the incorporation dT (75%) followed by dA, dG (22%) and dC (3%) (data not shown).

3. Generic bypass of hydrophobic base analogues

We tested the ability of 5D4 to bypass a range HBAs including 3-nitropyrrole (NP), pyrrole dicarboxamide (PDC), difluorotoluene (DFT), indole (IN), benzimidazole (BI). These include

HBAs that are structurally similar to the d5NI(C) selection bait (IN, BI) as well as structurally unrelated HBAs (NP, PDC, DFT). Furthermore, we investigated the ability of 5D4 to bypass consecutive template d5NIs or d5NICs using a gel-based assay (Fig. 7). While some analogues, notably PDC, NP, DFT and BI were bypassed by *Taq* (although poorly for BI (< 10%)), in all cases bypass by 5D4 was significantly more efficient. Indeed, neither template IN nor tandem d5NI-d5NI or d5NIC-d5NIC showed detectable bypass by *Taq*, while all three could be bypassed by 5D4. Incorporation specificities of dNTPs opposite HBAs show distinct preferences for *Taq* (predominantly following the A-rule^{44–46}, while 5D4 displays much reduced bias (supporting information, Figure 3).

4. Generic extension of specific basepairs involving HBAs

A number of HBAs can be incorporated efficiently and specifically by natural polymerases opposite their cognate partners but the nascent basepair then cannot be further extended and acts as a terminator. Cases in point are Pyrene (as its C-nucleoside triphosphate, dPyTP),^{24, 50,51} the triphosphate of d5NI (d5NITP) as well as a number of other indole derivatives^{25,52,53}. They have been shown to be incorporated with remarkable efficiency and specificity opposite a tetrahydrofuran abasic site analogue (ϕ). In both cases, incorporation is more efficient than “default” incorporation of dATP according to the A-rule^{44–46}, presumably due to good steric complementarity to the missing base pair,⁴⁷ and strong π -stacking with the 5'-nucleotide. However, in both cases further extension of the dPy: ϕ or d5NI: ϕ pairs by naturally occurring polymerases is absent or very inefficient. Indeed, we find that while dPyTP and d5NITP are incorporated efficiently opposite ϕ by *Taq* polymerase, it is then unable to extend beyond the dPy: ϕ or d5NI: ϕ pairs (3'-base: template base) (Fig 8). In contrast, 5D4 could both form and extend the dPy: ϕ or d5NI: ϕ pairs with good efficiency. 5D4 also could bypass a template ϕ by incorporating dATP opposite the abasic site and extending the dA: ϕ pair (not shown), but, surprisingly, extension of a dPy- ϕ pair was superior than a dA: ϕ pair (Fig. 8).

We also examined 7-azaindole (7AI) and isocarbostryl (ICS), which had been reported to form specific self- and heteropairs^{48, 49}, which were poorly extended by natural polymerases. Indeed, we found extension of ICS:ICS, ICS:7AI, 7AI:ICS and 7AI:7AI pairs by *Taq* to be close to undetectable (< 3%). In contrast, 5D4 could extend all pairs to some degree, with extension improved significantly by (1mM) Mn^{2+} . Despite the structural similarity of 7AI and 5NI, extension 3' 7AI was generally weak and extension of 3'-ICS clearly favoured. The ICS:7AI pair proved a particularly good substrate with extension reaching completion within 5 min with little or no stalled intermediates (Fig. 9). It should be noted that much superior HBA heteropairs have been described since by the Romesberg lab²¹ but these were not accessible to us. Our use of ICS and 7AI was to illustrate the potential of 5D4 for improving the utilization of HBAs that are inherently suboptimal polymerase substrates.

Structural analysis of DNA primer-template duplexes containing d5NI and d5NIC

In order to better understand the structural framework for d5NI(C) incorporation and extension, we determined the solution structure of two primer-template duplexes (*tni* and *tnic*) with d5NI or d5NIC poised for extension at the 3'-end of the primer strand by NMR spectroscopy (Fig. 10a).

The most striking deviation from canonical DNA structure is the intercalation of the d5NI and d5NIC heterocycles into the template strand base-stack: in both *tni* and *tnic*, the nitroindole rings are unpaired and stack on the neighboring C14:G5 pair, intercalating between template nucleotides A4 and G5 (Fig. 10b). This conformation is clearly indicated by an extensive

network of sugar-aromatic and aromatic-aromatic NOEs between d5NI(C)15 and the C14, A4 and G5 nucleotides, and by the absence of the expected A4 H2-G5 H1' interaction (Fig. 11a, supporting information Fig. 4).

As observed previously,⁵⁰ both d5NI and d5NIC adopt standard *anti* conformations (Fig. 11a, b), placing the H2 and H3 protons (purine numbering) in the minor groove of the helix and H6, H8 and H7 (d5NI) or the carboxamide group (d5NIC) in the major groove. The d5NIC15 carboxamide is oriented such that the amide-NH₂ group points towards the primer strand backbone and establishes a hydrogen-bonding interaction with the d5NIC phosphate (Fig. 10b). This contact is supported by the observation of two sharp downfield- and upfield-shifted carboxamide proton resonances giving rise to stronger NOE interactions with d5NIC H8 relative to d5NIC H6, and to NOEs with the sugar H3', H2' and H2'' and base H5 and H6 protons of C14, with the downfield-shifted proton generating stronger NOEs with d5NIC H8 and C14 H3' (Fig. 11a). The interaction between the downfield-shifted carboxamide proton and the d5NIC15 phosphate covalently bonded to C14 O3' is further supported by the adoption of an unusual C3'-*endo* conformation by the *tnic* (but not *tni*) C14 sugar, which facilitates the formation of this intra-strand contact. The absence of any unusual resonance broadening in *tni* and *tnic* contrasts with the dynamic effects previously observed in DNA duplexes containing d5NI and d5NIC in internal positions, where the unpaired nitroindole bases were found to exchange between two alternative intercalated conformations⁵⁰; in *tni* and *tnic* no spectral broadening is observed because intercalation between A4 and G5 is the only stable stacked conformation available for the terminal nitroindole base (Fig. 10b).

Discussion

We have used CSR selection³¹ for the directed evolution of a polymerase (5D4) with a generically enhanced ability for the synthesis nucleic acids comprising HBAs, including large HBA analogues that display poor geometric fit and lack minor groove H-bonding capacity, which were previously refractive to enzymatic incorporation, extension and/or bypass. Examples for such HBAs are d5NI and d5NIC, which were used here as the selection “bait”. d5NI is a “archetypal” HBA in that it lacks any H-bonding potential and, consequently, interacts with the opposing base by stacking (rather than pairing with it), giving rise to its universal base properties in hybridization applications³⁶. Typically for many larger HBAs, this stacking interaction causes conformational distortions of the primer-template duplex by intercalation into the opposing strand base-stack as we have shown here by NMR (Fig. 10b). While d5NI and d5NIC display virtually identically hybridization properties, d5NIC is a superior polymerase substrate. Examination of the NMR data suggests a possible mechanism for the enhanced rate of d5NIC incorporation and extension. We find the carboxamide group of d5NIC projecting into the major groove allowing us to exclude a potential participation in minor groove H-bonding, which crucially affect polymerase extension efficiency²². Location in the major groove may be preferred due to its hydrophilic nature, which permits the solvation of the carboxamide group. Indeed, Klewer *et al* studied the NMR structure of duplexes containing another carboxamide substituted HBA (1,2,4-triazole-3-carboxamide) and found that the carboxamide group also resided in the major groove.⁵¹ However, while unable to participate in minor groove interactions with the polymerase, we find the carboxamide group perfectly placed for a hydrogen bonding interaction with its own 5' phosphate backbone group (Fig. 10b). In addition, we find that 5-nitroindole-3-methylcarboxamide, an analogue with similar stacking but in which the ability of the carboxamide group to form hydrogen bonds is hindered, is as poor a polymerase substrate as d5NI (not shown). This suggests an important contribution from this hydrogen bond interaction towards d5NIC's favourable substrate properties, presumably by restriction of lateral movement and improved positioning of the 3'-OH primer terminus for catalysis. Design of proximal H-bonding groups into HBAs may be worth

exploring as a general strategy to improve properties of HBAs or indeed other base analogues for enzymatic replication.

Another potential interaction is observed when d5NIC is “paired” with dT in the opposite strand (as opposed to dA as in Fig. 10). In that case, in addition to the intra-strand hydrogen-bond to its own phosphate described here, the carboxamide group can form an out-of-plane inter-strand hydrogen-bond with O4 of dT.⁵⁰ The latter may provide an explanation for the different sequence bias of d5NIC extension (Fig. 4B) and templating (favouring dT incorporation) compared to d5NI (favouring dA). However, this interaction does not appear to play any role during the incorporation step as neither the incorporation of d5NICTP opposite template dT nor the incorporation of dTTP opposite template d5NIC is especially favoured (Fig. 5).

Despite these non-canonical interstrand and intrastrand interactions by d5NIC and the distortions caused by the intercalation of both d5NI and d5NIC into the template strand base-stack, both the [d5NI(C)]₂ self-pairs as well as some of the d5NI(C)-dA, dG, dC, dT heteropairs are synthesized by 5D4 with kinetic efficiencies approaching or exceeding those of the canonical base-pairs (Fig. 5). Once formed d5NI(C)-dN heteropairs and d5NIC self-pairs (but not d5NI self-pairs) are efficiently extended by 5D4, while neither is extended by Taq (Fig. 6, supporting information Fig. 2). 5D4 also greatly outperforms other polymerases on HBA pairs that do not distort DNA conformation, notably the dPy and d5NI heteropairs with an abasic site (ϕ). In these, the HBA ring occupies the space left by the missing template base and completes the opposing strand base-stack without distorting DNA conformation^{47, 52}. While formed efficiently and specifically by natural polymerases, they act as terminators^{19, 53}. 5D4, in contrast, is not only able to extend the unnatural d5NI- ϕ and dPy- ϕ “base-pairs” efficiently but even extends dPy- ϕ in preference to a “natural” dA- ϕ pair (Fig. 8). Specific formation and efficient extension of dPy- ϕ and d5NI- ϕ heteropairs by 5D4 raises the potential of the synthesis of long DNA polymers with d5NI, dPy (and potentially other large HBAs) inserted at defined positions as determined by the positioning of ϕ groups in the synthetic template.

The ability of 5D4 to efficiently replicate a wide variety of HBAs allowed their potential for coding to be readily explored. We examined the coding preferences of 3-nitropyrrole (NP), pyrrole dicarboxamide (PDC), difluorotoluene (DFT), indole (IN), benzimidazole (BI), two consecutive template d5NIs or d5NICs as well as the tetrahydrofuran abasic site analogue (ϕ) with both wtTaq and 5D4. We find that, when Taq is able to bypass an HBA it predominantly follows the A-rule^{44–46} due to dA’s favourable stacking properties. Although generally favouring dATP and dTTP incorporation, 5D4 displays a much more even incorporation profile, approaching near universal base behavior for NP and PDC both as templating bases and as deoxynucleotide triphosphates (supporting information Fig. 3, 5).

5D4 is a chimeric polymerase incorporating segments from both Taq and Tth polymerases, as well as a very short segment from Tfl at the N-terminus. No crystal structures exist for either Tth or Tfl but with on average 80% sequence homology between the three polymerases, the available structures of Taq polymerase provide a close structural analogue for those regions deriving from Tth and Tfl. Due to its chimeric nature, 5D4 differs by a total of 41 mutation from the Taq consensus. The bulk of these mutations is concentrated in the 5′-3′ exonuclease domain, which largely derives from Tth, while the main polymerase domain of 5D4 largely derives from Taq except for two short Tth segments around residues 710–730 and at the very C-terminus (Supporting information Fig. 1). In addition the mutations deriving from Tth, 5D4 comprises 14 point mutations not present in the parental genes. Some of these are unique to 5D4, while others are shared with a group of polymerases isolated from CSR rounds 4 (4C11) and 5 (5B1, 5B4, 5D3) (Supporting information Figure 1), which display a very similar (if slightly weaker) phenotype to 5D4. This identifies E602G, A608V, I614M, M762T and M775T as mutations within the main polymerase domain that are likely to be associated with the

phenotype as they are present in all or most of the polymerases from this group (supporting information Table 1). However, from a simple inspection it is far from clear how these mutations contribute to the phenotype as, with the exception of I614M, they are distant from the active site. We attempted to rationalize the relative contributions of these mutations as well as others towards the 5D4 phenotype using computational analysis as well as reverse genetics.

By their iterative nature, directed evolution experiments frequently yield mutations that incrementally contribute to the new phenotype. Such mutations are often located distal to the active site as direct perturbation of the active site would be most likely to cause substantial losses in catalytic activity and be selected against. Various strands of evidence suggest that such mutations can affect the functional properties of active sites by correlated motion propagated through networks of amino acids⁵⁴, which connect distal regions of the protein structure with each other. An effective approach to identify such networks is based on the assumption that covarying residues in protein families reveal functionally interacting amino acids independent of their location in the three-dimensional protein structure^{55, 56} and may thus be a useful tool to rationalize the outcome of evolution experiments as suggested by Lockless & Muir⁵⁷. We have applied this approach, utilizing Statistical coupling analysis (SCA) to the main polymerase domain using an alignment of 994 members of the polA family of bacterial DNA polymerases. Hierarchical clustering analysis using a high stringency cut-off to map only the strongest correlations, identified 40 residues (10% of the aligned sequence) that together with the conserved polymerase core (> 97% conservation across all sequences), form a network of spatially contiguous amino acids and display multiple contact points to the primer-template duplex (supporting information Fig. 6). SCA analysis revealed M761T and M775T as part of the network suggesting a conduit by which they could modify polymerase function. The uncovered network connects these mutations to F667 within the polymerase active site in direct contact with the incoming deoxynucleotide triphosphate (Fig. 12). Numerous studies have identified F667 (or its equivalent residue F762 in *E. coli* DNA polymerase I) as a key factor in polymerase substrate selection. Mutation of F667 has been shown to have dramatic effects e.g. in the discrimination against ddNTPs as well as on nucleotide incorporation fidelity^{58, 59}. Its function has been proposed to be to constrain the incoming dNTP molecule as part of geometric substrate selection and correct positioning for attack by the primer 3' OH.

It is important to note that mutations may be selected for entirely different reasons than a direct contribution to HBA utilization. During directed evolution experiments, other traits are also under adaptive pressure. These include expression, folding and in particular protein stability. As the majority of mutations are likely to be destabilizing, there is a limit to the number of mutations that can be sustained before a critical stability threshold is reached and the protein is no longer functional under the selection conditions⁶⁰. This is of especially acute importance in selection regimes such as CSR, where the protein is subjected to high temperature thermocycling conditions. We had previously observed that chimeric polymerases comprising the Tth 5'3' exonuclease domain and the Taq polymerase domain were substantially more thermostable than Taq polymerase⁴⁰. The chimeric nature of 5D4 and the other selected polymerases, comprising a large segments of the Tth 5'-3' exonuclease domain may therefore have been selected for due to its stabilizing effect on overall polymerase structure, thereby promoting evolvability through increased tolerance of destabilizing mutations. Indeed, using FoldX analysis to predict the change in free energy of folding ($\Delta\Delta G$) upon introduction of 5D4 specific mutations into the Taq framework, we found that most mutations in the main polymerase domain are destabilizing (supporting information Fig. 7, Table 5). However, FoldX⁶¹ analysis also identified A608V and surprisingly I614M as key stabilizing mutations shared by all selected polymerases. A608V was previously observed in a mutant Taq polymerase (T8) selected for increased thermostability³¹. We therefore conclude that the bulk of mutations in the 5'3'exonuclease domain together with A608V are likely to have been

selected to increase overall polymerase stability and offset the destabilizing influence of other mutations of adaptive value.

We reverted the E602G, I614M, M762T and M775T mutations in the main polymerase domain that were conserved among all the polymerases displaying the “HBA phenotype” back to wild-type (G602E, M614I, T762M, T775M) and analyzed their properties using the d5NIC PCR assay (Fig. 2c) as it simultaneously determines d5NIC extension and bypass ability and most closely resembles the CSR process by which the mutations were selected. Only one backmutation, 5D4: M614I, displayed a significant reversion phenotype, in that its activity in d5NIC PCR was markedly reduced (supporting information Fig. 8). All the other back mutations showed only marginal reductions in d5NIC PCR activity and may therefore contribute only incrementally towards the phenotype or through interaction with I614M. I614 is located in the A-motif within the polymerase active site and is directly involved in binding the incoming dNTP substrate. The change from Ile to Met, results in decreased steric constraints within the active site by removal of a CH₃ group projecting into the active site and towards the incoming dNTP (Fig. 12). Indeed, mutation of I614 (I614K⁶², I614M⁶³ or I614T³⁰) has been found to decrease discrimination against non-cognate substrates such as NTPs within the polymerase active site either alone or in conjunction with mutation of the juxtaposed E615 steric gate residue. Within the same group of selected polymerases the proximal residue E602 has also been found to be mutated (E602V⁶³) in conjunction with I614M.

Taken together, these findings clearly implicate I614 as a critical residue in the steric control of substrate selection and suggest that the 5D4 phenotype may arise to a substantial extent from a simple relaxation of steric control within the active site through the I614M mutation as well as through the propagation of the effects of the M761T and M775T (and E734N) mutations through the SCA network to polymerase active site.

One prediction arising from such a model would be that Taq and 5D4 should perform approximately equally well on small HBA substrates, while 5D4 should outperform Taq on large HBAs. This is exactly what we observe. For example, while bypass of small template HBAs like DPC and NP proceeds with comparable efficiency for Taq and 5D4 (Fig. 6), only 5D4 is able to bypass larger HBAs such as IN, BI and d5NI(C). Likewise, incorporation of small HBA triphosphates like NP-TP and PMC-TP proceeds with comparable efficiency for Taq and 5D4 (supporting information Fig. 5), while incorporation of larger HBA-TPs like d5NI(C)-TP by 5D4 is up to 300-fold more efficient than Taq (Fig. 5). Incorporation of dyATP,⁶⁵ a large expanded dA analogue with Watson-Crick H-bonding ability but poor geometric fit and extension of a distorting A G transversion mismatch as well as were also enhanced by 5D4 (supporting information Fig. 2b, 9). The relaxation of geometric substrate selection also permits increased tolerance for non-cognate template strand conformations. As we have shown here by NMR (Fig. 10, 11), 3' d5NI(C) causes distortion of the template strand conformation by intercalation. Nevertheless, 3' d5NI(C)-dN heteropairs (and d5NIC-d5NIC self-pairs) are efficiently extended by 5D4 while they stall extension by Taq polymerase (Fig. 6).

Relaxing steric control in the polymerase active site might also be expected to give rise to low fidelity, poor catalytic efficiency and reduced processivity. However, we find dNTP incorporation and extension kinetics (Fig. 4a) as well as efficiency in standard PCR to be comparable for Taq and 5D4 (Fig. 2b). Similarly, we find that, although the overall rate of nucleotide misincorporation (3.1×10^{-4}) by 5D4 is increased ca. 5-fold compared to wtTaq polymerase (M. Arana, PH, T. Kunkel, manuscript in preparation), it is comparable to other polymerases (such as Klenow exo⁻⁶⁶) widely used in molecular biology applications.

In conclusion, CSR selection using oligonucleotide primers comprising the HBA d5NI and its carboxamide derivative d5NIC as substrates has yielded 5D4, a polymerase with a generic

ability to synthesize nucleic acids comprising HBAs while maintaining robust catalytic activity and fidelity. Particularly striking is the capacity of 5D4 to form and extend a diverse collection of unnatural base pairs involving HBAs. The ability of 5D4 to efficiently process large analogues that lack minor groove H-bonding and distort cognate DNA geometry should relax HBA design constraints and expedite the synthesis of DNA fragments comprising diverse HBAs. The properties of 5D4 bode well for its application in unlocking the coding potential of HBAs and other unnatural nucleotide analogues, previously incompatible with enzymatic replication.

Materials and Methods

Nucleotides and Oligonucleotides

5-Nitroindole and difluorotoluene phosphoramidites were supplied by Glen Research. 5-Nitroindole-3-carboxamide phosphoramidite³⁹, 1-(2-deoxy- β -D-ribofuranosyl)-5-nitroindole 5'-triphosphate,³⁸ pyrrole dicarboxamide,⁶⁷ indole, benzimidazole,⁶⁸ pyrene¹⁹ and dyATP⁶⁵ were prepared as previously described. 5-Nitroindole-3-carboxamide attached to controlled pore glass support was prepared according to the method of Pon.⁶⁹ Synthesis of 1-(2-deoxy- β -D-ribofuranosyl)-5-nitroindole-3-carboxamide 5'-triphosphate: To an ice-cold solution of methyl-1-(2-deoxy- β -D-ribofuranosyl)-5-nitroindole-3-carboxylate³⁹ (100mg, 0.3mmol) and proton sponge (96mg, 0.45mmol) in trimethyl phosphate (3cm³) was added phosphoryl chloride (35 μ l, 0.38mmol) and the solution stirred at 0°C for 5h. To this was added simultaneously tributylamine (0.5cm³) and tetrabutylammonium pyrophosphate solution (0.5M in DMF, 2cm³), and the solution stirred for a further 30min. The reaction was then quenched by the addition of 0.5M TEAB buffer (10cm³), and stored at 4°C overnight. The solution was evaporated to dryness and re-dissolved in water (20cm³) and applied to a Sephadex A25 column in 0.05M TEAB buffer. The column was eluted with a linear gradient of 0.05–1.0M TEAB. Appropriate fractions were pooled and evaporated to dryness to give a yellow solid of methyl-1-(2-deoxy- β -D-ribofuranosyl)-5-nitroindole-3-carboxylate 5'-triphosphate. Yield 110mg. HPLC (Phenomenex Luna 10 μ C-18 reverse phase column, buffer A, 0.1M TEAB; buffer B, 0.1M TEAB, 25% MeCN. 25% to 100% buffer B over 45min at 8 ml/min.) showed the product to be pure. δ_P (D₂O) –9.35 (d, γ -P), –10.15 (d, α -P), –22.05 (t, β -P). A solution of methyl-1-(2-deoxy- β -D-ribofuranosyl)-5-nitroindole-3-carboxylate-5'-triphosphate (70mg) in 0.880 ammonia (10cm³) was stirred at room temperature overnight. HPLC showed complete conversion. The solution was evaporated to a yellow solid, and the product purified by HPLC. δ_P (D₂O) –5.15 (d, γ -P), –10.10 (d, α -P), –21.25 (t, β -P). The title compound was converted into its sodium salt by passage through a Dowex 50WX4-200 resin (Na⁺ form). Yield 418.8 OD. δ_P (D₂O) –6.92 (d, γ -P), –9. (d, α -P), –21.05 (t, β -P).

Selection, screening, protein expression and characterization

For selection we used the previously described library 3T (1×10^9 cfu, 70% active clones).⁴⁰ Emulsification and CSR selection were performed as described^{31, 70} using primers 1, 2 for rounds 1, 2, primers 1–4 for round 3 and 3–7 for rounds 4–5, cycled 20x (94°C 30sec, 50°C 30sec, 72°C 5 min), reamplified with gene specific primers 8–13 for rounds 1, 2, with primers 8–13 and out-nested primers 14, 15 or combinations thereof for rounds 3–5 and recloned Xba I/Sal I into pASK75 as described.³¹ After selection rounds one and two, clones were screened by d5NIC PCR with primers 1,2 or 3,4 and by polymerase ELISA as described⁶³ using hairpins 16–19 or 20–23. Promising clones from rounds 3 and 4 were StEP shuffled⁷¹ and backcrossed with parent polymerase genes. Clones analyzed in more detail in this report derive from selection rounds 4 and 5. Selected mutations were reverted back to wild-type Taq sequence using Quickchange Mutagenesis using Pfu Turbo (Stratagene) and primers 31–40. Expression of polymerases for characterization was as described⁴⁰ using a 16/10 Hi-Prep Heparin FF Column (Amersham Pharmacia Biotech). Polymerase fractions eluted around 0.3M NaCl and

were concentrated and dia-filtered into 50mM Tris pH 7.4, 1mM DTT, 50% glycerol and stored at -20°C . Mutation rates were determined using a well-established *in vivo* gap filling assay⁶⁹ (M. Arana, PH, T. Kunkel, unpublished results). 5D4 PCR products with primers 1–7 and pASK75 as template, were reamplified using 5D4 with primers 14, 15, TOPO cloned (Invitrogen) and sequenced.

Primer extension & PCR assays

Extension reactions with purified polymerases were carried out by addition of 4 μl of 2.5mM dNTP mix (final concentration 50 μM each dATP, dTTP, dCTP, dGTP) to 46 μl of a reaction mixture of final concentration containing 1 x Taq buffer, 50pmol ^{32}P -labelled primers 20, 21 or 28, 100pmol templates 22–26, 29, 30 and 1 μl of polymerase (wtTaq (1.5 μg) or 5D4 (16 μg), activity normalized) at 60°C . 8 μl aliquots were removed and added to 8 μl stop solution (8 M urea, 50 mM EDTA, $\sim 0.1\%$ xylene cyanol F) at 0, 10, 20, 30 and 40 minutes, and the products electrophoretically separated on 20% polyacrylamide gels. Kinetic primer extension reactions were carried out by mixing of equal volumes of primers 20 or 21/templates 22–25 (100 μl stock solution containing 1 x Taq buffer, 80pmol ^{32}P -labelled primer and 200pmol template) and polymerase (wtTaq or 5D4)/dXTP mix (100 μl stock solution containing 1xTaq buffer, dXTP to final concentration between 1–160 μM and polymerase, X=dA, dT, dC, dA, d5NI, d5NIC). Reaction mixtures were mixed at 60°C and quenched after various time intervals by the addition of an equal volume of stop solution (8M urea, 50mM EDTA, $\sim 0.1\%$ xylene cyanol F) before electrophoretic separation on 20% polyacrylamide gel. Kinetic reactions were all performed in triplicate. kcat/Km values are in $\% \mu\text{M}^{-1} \text{min}^{-1}$. Polyacrylamide gels were dried and exposed to a phosphorimager screen (Amersham Biosciences or Molecular Dynamics) and scanned on a Typhoon 8610 (Molecular Dynamics). Data was initially analyzed using Geltrak^{72, 73} and processed using Kaleidagraph (Synergy Software) or Excel (Microsoft). PCR assays were performed using primers 1–7, pASK75 template and PCR conditions 20x (94°C 30sec, 50°C 30sec, 72°C 30 sec) for d5NIC primers and 50x (94°C 30sec, 50°C 30sec, 72°C 5 min) for d5NI primers on a MJ TETRAD thermocycler.

NMR spectroscopy and NMR model calculation

NMR spectra of *tni* and *tnic* (Fig. 8) were acquired on Bruker DRX-500 and DMX-600 spectrometers, processed using NMRPIPE,⁷⁴ and analyzed using Sparky 3.106.⁷⁵ Two-dimensional NMR spectra recorded in D_2O included ^1H - ^{31}P HetCOR and uninterrupted series of dqf-COSY, TOCSY, ROESY and NOESY (with 60, 120 and 250 ms mixing times) experiments. NOESY spectra were also acquired in H_2O at 9°C with a mixing time of 120 ms. For structure determination, distance restraints were estimated from NOESY build-ups using NOE interactions corresponding to covalently constrained inter-proton distances as a reference; base-pair hydrogen-bonding restraints were introduced based on chemical shifts and interactions observed in H_2O -NOESY experiments; and sugar-phosphate backbone dihedral restraints were deduced from the cross-peak patterns observed in dqf-COSY, ^1H - ^{31}P HetCOR and NOESY spectra.⁵⁰ Based on strong NMR evidence (Fig. SNMR), the C6-G13 double-helical stems (Fig. 8A) of *tni* and *tnic* were constrained to a B-DNA conformation, and NMR models of *tni* and *tnic* were calculated by restrained molecular dynamics, using extensive distance and dihedral NMR restraint sets for the T1-C6 and G13-d5NI(C)15 nucleotides (Supporting Table NMR), a distance-dependent dielectric constant, and the MMFF94 force field⁷⁶ of SYBYL 6.9 (Tripos Inc.)

Statistical coupling analysis

Starting from the Conserved Domain Database⁷⁷ entry on the polymerase A family (cd06444), 3987 protein sequences were identified in the NCBI database as putative PolA by the position-specific score matrix of the family. Of those, 3484 were of bacterial origin and were selected

for further processing. Clustal W⁷⁸ and manual curation were used to identify duplicate entries and to reduce sampling bias due to single species overrepresentation. The resulting dataset contained 1057 sequences, from which sequence alignment, using MUSCLE⁷⁹, was carried out iteratively to remove sequences of dubious quality and to trim the alignment to the polymerase domain. The resulting sequence set, used in the statistical coupling analysis, contained 994 polymerase domain sequences aligned to the *Thermus aquaticus* DNA polymerase A domain (E432 – E832 in PDB structure 3KTQ⁸⁰) (available on request).

SCA⁵⁵ was performed using the SCA toolbox v3.0, kindly provided by R. Ranganathan (Dallas), in MATLAB (The Mathworks, Inc.) using the 994-polymerase alignment. SCA was carried out using *Thermus aquaticus*, *Geobacillus stearothermophilus* and *Escherichia coli* as query sequences and structures. The SCA output correlation matrix (available on request) was further analyzed using Excel. An arbitrary identity cut-off of 97% was used for analysis. Aligned residues conserved above the cut-off were not considered in the analysis as subalignments used for SCA calculations for those residues would contain fewer than 30 sequences. The log-normal fit of the SCA correlations suggested 0.85 kT^* (mean plus 3 SDs) as the significance threshold. We initially set an arbitrary 1.80 kT^* cut-off to identify the most relevant couplings. A more detailed analysis of the 5D4 mutations was also carried out using the 0.85 kT^* cut-off and to identify couplings from 5D4 mutations with greater sensitivity.

Protein stability analysis using FoldX

FoldX⁶¹ (version 3.0beta3, <http://foldx.crg.es/foldx.jsp>) was used to predict the effect of mutations found in 5D4 on protein stability by comparing the free energy of folding between mutants and wild-type Taq. Comparisons were carried out using 1CMW⁸¹ (apo structure of full-length Taq), 1KTQ⁸² (apo structure of Klentaq fragment), 2KTQ⁸⁰ (Klentaq fragment in an open ternary complex) and 3KTQ⁸⁰ (Klentaq fragment in a closed ternary complex) and the effect of insertion of individual 5D4 mutations into the Taq framework was computed.

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

J.G. is grateful for the support of an EU Marie Curie Fellowship (QLK2-CT-1999-51436). The authors wish to thank Prof. R. Ranganathan (Green Centre for Systems Biology, Dallas, TX, USA) for providing SCA toolbox vs 3.0.

References

1. Kool ET. Acc Chem Res 2002;35:936–943. [PubMed: 12437318]
2. Seeman NC. Trends Biochem Sci 2005;30:119–125. [PubMed: 15752983]
3. Hwang GT, Romesberg FE. Nucleic Acids Res 2006;34:2037–2045. [PubMed: 16617144]
4. Matsuda S, Henry AA, Romesberg FE. J Am Chem Soc 2006;128:6369–6375. [PubMed: 16683801]
5. Kim Y, Leconte AM, Hari Y, Romesberg FE. Angew Chem Int Ed 2006;45:7809–7812.
6. Hari Y, Hwang GT, Leconte AM, Joubert N, Hocek M, Romesberg FE. ChemBioChem 2008;9:2796–2799. [PubMed: 19012285]
7. Berger M, Wu Y, Ogawa AK, McMinn DL, Schultz PG, Romesberg FE. Nucleic Acids Res 2000;28:2911–2914. [PubMed: 10908353]
8. Matsuda S, Romesberg FE. J Am Chem Soc 2004;126:14419–14427. [PubMed: 15521761]
9. Leconte AM, Hwang GT, Matsuda S, Capek P, Hari Y, Romesberg FE. J Am Chem Soc 2008;130:2336–2343. [PubMed: 18217762]
10. Leconte AM, Matsuda S, Hwang GT, Romesberg FE. Angew Chem Int Ed 2006;45:4326–4329.

11. Tae EL, Wu Y, Xia G, Schultz PG, Romesberg FE. *J Am Chem Soc* 2001;123:7439–7440. [PubMed: 11472182]
12. Leconte AM, Matsuda S, Romesberg FE. *J Am Chem Soc* 2006;128:6780–6781. [PubMed: 16719445]
13. Kawai R, Kimoto M, Ikeda S, Mitsui T, Endo M, Yokoyama S, Hirao I. *J Am Chem Soc* 2005;127:17286–17295. [PubMed: 16332078]
14. Kimoto M, Mitsui T, Harada Y, Sato A, Yokoyama S, Hirao I. *Nucleic Acids Res* 2007;35:5360–5369. [PubMed: 17693436]
15. Moriyama K, Kimoto M, Mitsui T, Yokoyama S, Hirao I. *Nucleic Acids Res* 2005;33:e129. [PubMed: 16113238]
16. Kimoto M, Endo M, Mitsui T, Okuni T, Hirao I, Yokoyama S. *Chem Biol* 2004;11:47–55. [PubMed: 15112994]
17. Hirao I, Ohtsuki T, Fujiwara T, Mitsui T, Yokogawa T, Okuni T, Nakayama H, Takio K, Yabuki T, Kigawa T, Kodama K, Yokogawa T, Nishikawa K, Yokoyama S. *Nat Biotech* 2002;20:177–182.
18. Hirao I, Kimoto M, Mitsui T, Fujiwara T, Kawai R, Sato A, Harada Y, Yokoyama S. *Nature Methods* 2006;3:729. [PubMed: 16929319]
19. Matray TJ, Kool ET. *Nature* 1999;399:704–708. [PubMed: 10385125]
20. Zhang X, Donnelly A, Lee I, Berdis AJ. *Biochemistry* 2006;45:13293–13303. [PubMed: 17073450]
21. Seo YJ, Hwang GT, Ordoukhanian P, Romesberg FE. *J Am Chem Soc* 2009;131:3246–3252. [PubMed: 19256568]
22. Morales JC, Kool ET. *Biochemistry* 2000;39:12979–12988. [PubMed: 11041863]
23. Matsuda S, Leconte AM, Romesberg FE. *J Am Chem Soc* 2007;129:5551–5557. [PubMed: 17411040]
24. Leconte AM, Hwang GT, Matsuda S, Capek P, Hari Y, Romesberg FE. *J Am Chem Soc* 2008;130:2336–2343. [PubMed: 18217762]
25. Gardner AF, Jack WE. *Nucleic Acids Res* 1999;27:2545–2553. [PubMed: 10352184]
26. Li Y, Mitaxov V, Waksman G. *Proc Natl Acad Sci USA* 1999;96:9491–9496. [PubMed: 10449720]
27. Suzuki M, Baskin D, Hood L, Loeb LA. *Proc Natl Acad Sci USA* 1996;93:9670–9675. [PubMed: 8790389]
28. Summerer D, Rudinger NZ, Detmer I, Marx A. *Angew Chem Int Ed Engl* 2005;44:4712–4715. [PubMed: 15995989]
29. Jestin JL, Kristensen P, Winter G. *Angew Chem Int Ed* 1999;38:1124–1127.
30. Xia G, Chen L, Sera T, Fa M, Schultz PG, Romesberg FE. *Proc Natl Acad Sci U S A* 2002;99:6597–6602. [PubMed: 12011423]
31. Ghadessy FJ, Ong JL, Holliger P. *Proc Natl Acad Sci USA* 2001;98:4552–4557. [PubMed: 11274352]
32. Ghadessy FJ, Ramsay N, Boudsocq F, Loakes D, Brown A, Iwai S, Vaisman A, Woodgate R, Holliger P. *Nat Biotechnol* 2004;22:755–759. [PubMed: 15156154]
33. Fa M, Radeghieri A, Henry AA, Romesberg FE. *J Am Chem Soc* 2004;126:1748–1754. [PubMed: 14871106]
34. Leconte AM, Chen L, Romesberg FE. *J Am Chem Soc* 2005;127:12470–12471. [PubMed: 16144377]
35. Tawfik DS, Griffiths AD. *Nature Biotechnol* 1998;16:652–656. [PubMed: 9661199]
36. Loakes D, Brown DM. *Nucleic Acids Res* 1994;22:4039–4043. [PubMed: 7937128]
37. Loakes D. *Nucleic Acids Res* 2001;29:2437–2447. [PubMed: 11410649]
38. Smith CL, Simmonds AC, Felix IR, Hamilton AL, Kumar S, Nampalli S, Loakes D, Hill F, Brown DM. *Nucleosides Nucleotides* 1998;17:541–554.
39. Too K, Brown DM, Holliger P, Loakes D. *Coll Czech Chem Commun* 2006;71:899–911.
40. d’Abbadie M, Hofreiter M, Vaisman A, Loakes D, Gasparutto D, Cadet J, Woodgate R, Paabo S, Holliger P. *Nat Biotechnol* 2007;25:939–943. [PubMed: 17632524]
41. Eom SH, Wang J, Steitz TA. *Nature* 1996;382:278–281. [PubMed: 8717047]
42. Creighton S, Bloom LB, Goodman MF. *Methods Enzymol* 1995;262:232–256. [PubMed: 8594351]
43. Huang MM, Arnheim N, Goodman MF. *Nucleic Acids Res* 1992;20:4567–4573. [PubMed: 1408758]

44. Sagher D, Strauss B. *Biochemistry* 1983;22:4518–4526. [PubMed: 6354260]
45. Randall SK, Eritja R, Kaplan BE, Petruska J, Goodman MF. *J Biol Chem* 1987;262:6864–6870. [PubMed: 3571289]
46. Takeshita M, Chang CN, Johnson F, Will S, Grollman AP. *J Biol Chem* 1987;262:10171–10179. [PubMed: 2440861]
47. Zahn KE, Belrhali H, Wallace SS, Doubie S. *Biochemistry* 2007;46:10551–10561. [PubMed: 17718515]
48. Tae EL, Wu Y, Xia G, Schultz PG, Romesberg FE. *J Am Chem Soc* 2001;123:7439–7440. [PubMed: 11472182]
49. Yu C, Henry AA, Romesberg FE, Schultz PG. *Angew Chem Int Ed* 2002;41:3841–3844.
50. Gallego J, Loakes D. *Nucleic Acids Res* 2007;35:2904–2912. [PubMed: 17438041]
51. Klewer DA, Zhang P, Bergstrom DE, Davisson VJ, LiWang AC. *Biochemistry* 2001;40:1518–1527. [PubMed: 11327810]
52. Smirnov S, Matray TJ, Kool ET, de los Santos C. *Nucleic Acids Res* 2002;30:5561–5569. [PubMed: 12490724]
53. Reineks EZ, Berdis AJ. *Biochemistry* 2004;43:393–404. [PubMed: 14717593]
54. Goodey NM, Benkovic SJ. *Nat Chem Biol* 2008;4:474–482. [PubMed: 18641628]
55. Lockless SW, Ranganathan R. *Science* 1999;286:295–299. [PubMed: 10514373]
56. Suel GM, Lockless SW, Wall MA, Ranganathan R. *Nat Struct Biol* 2003;10:59–69. [PubMed: 12483203]
57. Lockless SW, Muir TW. *Proc Natl Acad Sci U S A* 2009;106:10999–11004. [PubMed: 19541616]
58. Astatke M, Grindley ND, Joyce CM. *Journal Of Molecular Biology* 1998;278:147–165. [PubMed: 9571040]
59. Suzuki M, Yoshida S, Adman ET, Blank A, Loeb LA. *J Biol Chem* 2000;275:32728–32735. [PubMed: 10906120]
60. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. *Proc Natl Acad Sci U S A* 2006;103:5869–5874. [PubMed: 16581913]
61. Guerois R, Nielsen JE, Serrano L. *J Mol Biol* 2002;320:369–387. [PubMed: 12079393]
62. Patel PH, Loeb LA. *J Biol Chem* 2000;275:40266–40272. [PubMed: 11005812]
63. Ong JL, Loakes D, Jaroslawski S, Too K, Holliger P. *J Mol Biol* 2006;361:537–550. [PubMed: 16859707]
64. Li Y, Korolev S, Waksman G. *EMBO J* 1998;17:7514–7525. [PubMed: 9857206]
65. Lu H, He K, Kool ET. *Angew Chem Int Ed* 2004;43:5834–5836.
66. Bebenek K, Joyce CM, Fitzgerald MP, Kunkel TA. *Journal Of Biological Chemistry* 1990;265:13878–13887. [PubMed: 2199444]
67. Loakes D, Guo MJ, Brown DM, Salisbury SA, Smith CL, Felix IR, Kumar S, Nampalli S. *Nucleosides, Nucleotides & Nucl Acids* 2000;19:1599–1614.
68. Loakes D, Hill F, Brown DM, Ball S, Reeve MA, Robinson PS. *Nucleosides, Nucleotides & Nucl Acids* 1999;18:2685–2695.
69. Pon, RT. *Protocols for oligonucleotides and analogs*. Vol. 20. Humana Press; 1993. p. 469
70. Ghadessy FJ, Holliger P. *Methods Mol Biol* 2007;352:237–248. [PubMed: 17041269]
71. Zhao H, Giver L, Shao Z, Affholter JA, Arnold FH. *Nature Biotechnol* 1998;16:258–261. [PubMed: 9528005]
72. Smith JM, Thomas DJ. *Comput Appl Biosci* 1990;6:93–99. [PubMed: 2361188]
73. Smith J, Singh M. *BioTechniques* 1996;20:1082–1087. [PubMed: 8780878]
74. Delaglio F, Grzesiek S, Vuister GW, Zhu G, Pfeifer J, Bax A. *J Biomol NMR* 1995;6:277–293. [PubMed: 8520220]
75. Goddard, TD.; Kneller, DG. *University of California*. San Francisco, CA, USA: 2001.
76. Halgren TA. *J Comp Chem* 1996;17:490–519.
77. Marchler-Bauer A, Anderson JB, Chitsaz F, Derbyshire MK, DeWeese-Scott C, Fong JH, Geer LY, Geer RC, Gonzales NR, Gwadz M, He S, Hurwitz DI, Jackson JD, Ke Z, Lanczycki CJ, Liebert CA,

- Liu C, Lu F, Lu S, Marchler GH, Mullokandov M, Song JS, Tasneem A, Thanki N, Yamashita RA, Zhang D, Zhang N, Bryant SH. *Nucleic Acids Res* 2009;37:D205–210. [PubMed: 18984618]
78. Thompson JD, Higgins DG, Gibson TJ. *Nucleic Acids Res* 1994;22:4673–4680. [PubMed: 7984417]
79. Edgar RC. *BMC Bioinformatics* 2004;5:113. [PubMed: 15318951]
80. Li Y, Korolev S, Waksman G. *EMBO J* 1998;17:7514–7525. [PubMed: 9857206]
81. Urs UK, Murali R, Krishna Murthy HM. *Acta Crystallogr D Biol Crystallogr* 1999;55:1971–1977. [PubMed: 10666572]
82. Korolev S, Nayal M, Barnes WM, Di Cera E, Waksman G. *Proc Natl Acad Sci U S A* 1995;92:9264–9268. [PubMed: 7568114]

Figure 1.

(a) Chemical structures of the base moieties of dA and hydrophobic base analogue 5-nitroindole (d5NI) and 5-nitroindole-3-carboxamide derivative (d5NIC) (b) Bypass of dA, d5NI and d5NIC by wild-type *Taq* polymerase. Polymerase activity was assayed for its extension ability on (from left to right) an unmodified DNA template (x=dA), the same template with d5NI respectively d5NIC inserted at the +1 position (x= d5NI; x=d5NIC). While d5NI stalls *Taq* almost completely (with < 10% full length extension product⁴¹), there is up to 50% bypass of d5NIC

Figure 2.

Principle of the CSR selection strategy with d5NI/d5NIC modified primers. CSR is based on a simple feedback loop, in which a polymerase replicates only its own encoding gene. Compartmentalization in the aqueous compartments of a water-in-oil emulsion³⁵ serves to isolate individual self-replication reactions from each other. Two independent emulsion compartments are shown. Polymerases (such as Pol1 (left compartment)) that are capable of utilizing flanking primers modified by d5NIC (x=d5NIC) are able to replicate their one encoding gene (*pol1*) and produce “offspring” i.e. increase their copy number in the post-selection population, while polymerases like Pol2 (right compartment) that are unable to utilize such primers disappear from the gene pool.

Figure 3.

(a) Activity of selected polymerase 5D4 compared to wtTaq in bypass of template d5NI and d5NIC. While Taq largely stalls, 5D4 readily bypasses template both template d5NIC and d5NI. (c) PCR amplification (20x (94°C 30sec, 50°C 30sec, 72°C 30 sec)) using (from left to right) unmodified standard primers (1,4), or primers in which either only the 3' terminal base is substituted by d5NI or d5NIC (2,5) or primers with both 3'-d5NIC as well as internal d5NIC substitutions (3,6,7). While both polymerases show comparable activity with unmodified primers (1, 4), only 5D4 yields amplification products with d5NI(C)-modified primers (2, 5–7) including (3, 7) requiring the extension and bypass no less than of four d5NIC substitutions every round of replication. PCR activity with d5NI-modified primers although detectable is much poorer than with d5NIC modified primers. m, ϕ X174 HaeIII digest marker.

Figure 4. 5D4 polymerase structure

Mutations from the *Taq* consensus in the selected polymerase 5D4 are mapped on a ribbon and surface representation of *Taq* DNA polymerase (1TAU.pdb⁴¹). 5D4 mutations that derive from *Th* gene segments are coloured light blue. Point mutations unique to 5D4 (or shared with less than two other selected polymerases) are coloured orange. 5D4 point mutations that are shared by a majority of selected polymerases (supporting information Table 1) are coloured red.

Figure 6.

(a) Formation and extension of a d5NIC-d5NIC self-pair is shown for both Taq and 5D4. While d5NIC-TP is incorporated efficiently opposite template d5NIC by both polymerases, further extension of the self-pair by Taq is stalled, while 5D4 efficiently extends the self-pair. (b) Extension of 3' d5NI(C) (% fully extended primer, 5 min) by 5D4 is plotted versus template bases N, N+1.

Figure 7.

Bypass of hydrophobic and universal base analogues. Primer extension (% of fully extended versus input primer) on DNA templates comprising hydrophobic and universal base analogues X= DPC, NP, DFT, IN, BI at the +1 position (and the +1 and +2 position for X=d5NI-d5NI, d5NIC-d5NIC) is plotted against incubation time for activity normalized polymerase concentrations of wtTaq (open diamonds) and 5D4 (black circles). Chemical structures for DPC, NP, DFT, IN, BI are shown. For the small HBAs, DPC and NP, *Taq* and 5D4 show similar activity, while 5D4 greatly out-performs *Taq* on larger HBAs. In particular 5D4 is able to bypass analogues such as IN, BI, [d5NI(C)]₂ as well as an abasic site that stall *Taq* under the experimental conditions.

Figure 8. Formation and extension of base-pairs between large HBAs and an abasic site

Primer extension on DNA templates comprising an abasic site (φ) at the +1 position by *Taq* and 5D4 is compared for incorporation of d5NITP (left) and dPyTP (right), followed by dATP chase. Chemical structures for the pyrene base (Py) and the tetrahydrofuran abasic site (φ) are shown. Incorporation of d5NITP and dPyTP is efficient for both *Taq* and 5D4. However, *Taq* is unable to extend either the d5NI-φ nor the dPy-φ pair, while 5D4 is able to extend both. In the case of dPy, extension with dATP leads to the formation of a dA-φ pair on the unoccupied φ sites (lower band) by both *Taq* and 5D4. This dA-φ pair is also not extended by *Taq*, while it is extended by 5D4, although, remarkably, 5D4 preferentially extends the dPy-φ pair (upper band). P is the unextended primer.

5' - TAATACGACTCACTATAGGGAGA (X)
 3' - ATTATGCTGAGTGATATCCCTCT (Y) GTCA

Figure 9. Extension of HBA base-pairs

Primer extension (% of fully extended product vs. input primer) on DNA templates comprising hydrophobic base-pairs comprising isocarbostryl (ICS) and 7-azaindole (7AI) self- (ICS:ICS; 7AI:7AI) and hetero-pairs (ICS:7AI, 7AI:ICS) is plotted against incubation time for activity normalized polymerase concentrations of wtTaq (open diamonds) and 5D4 (closed circles) in the presence and absence of 1mM Mn²⁺ (Ta_qMn, (grey diamonds); 5D₄Mn, (grey circles)). While Taq only poorly extends base-pairs comprising 3'-ICS (ICS:ICS, ICS:7AI), 5D4 is able to extend these (especially the ICS:7AI heteropair). Extension unnatural base-pairs comprising 3' 7AI (7AI:ICS, 7AI:7AI) by 5D4 is less efficient but superior to Taq.

Figure 10.

NMR spectroscopy analysis of primer-template duplexes containing d5NI and d5NIC base analogues. (a) sequence of the symmetric primer-template duplexes studied by NMR spectroscopy. (b) NMR models of primer-template duplexes containing d5NI (*tni*) and d5NIC (*mic*) at the 3' end of the primer strand. For clarity only one half of the symmetric duplex structures are shown. The models are oriented to illustrate the stacking of d5NI (orange) and d5NIC (green) on the -1 base-pair (G5:C14) and the intercalation of the nitroindole heterocycles between A4 and G5 of the template strand. The carboxamide group of d5NIC is oriented towards the C14-d5NIC15 backbone and establishes a hydrogen-bonding interaction with the d5NIC15 phosphate (dotted red line).

Figure 11.

(a) Aromatic-H1' region of the D₂O NOESY spectra (250 ms mixing time/25°C) of *tni*. Intra-residue pyrimidine H6-H1' and purine/d5NI H8-H1' cross-peaks are labelled with residue name and number, intraresidue pyrimidine H5-H6 crosspeaks are labelled with residue number, and sequential NOE connectivities are indicated with arrows, (only T1 to C6 and G13 to d5NI15 assignments shown). Cross-peaks (a) to (k) are assigned as follows: (a) d5NI15 H2-A4 H1'; (b) C14 H6-d5NI15 H7; (c) A4 H2-d5NI15 H1' and A4 H2-d5NI15 H7 (overlapped); (d) d5NI15 H8-C14 H5; (e) d5NI15 H6-C14 H5; (f) d5NI15 intraresidue H2-H1' & H2-H7 (overlapped); (g) d5NI15 intraresidue H3-H1' & H3-H7 (overlapped); (h) d5NI15 intraresidue H7-H8 (overlapped with H8-H1'); (i) d5NI15 intraresidue H7-H6; (j) G5 H8-C6 H5; (k) G13 H8-C14 H5. The absence of the sequential A4 H2-G5 H1' NOE is marked with an "X". (b) Aromatic region of the H₂O NOESY spectra (120 ms mixing time/9°C) of *tnic*, showing the interactions between the d5NIC H2, H3, H6, H8 and carboxamide (e and hb) protons.

Figure 12. Structure of the polymerase active site

Structure of the active site of the closed ternary complex of Taq polymerase (3KTQ⁶⁴) with the mutated residue I614 shown in red. Conserved residues (> 97% identity) are shown as the green surface together with the SCA network in shown as the orange surface comprising F667. The template strand is shown in teal and nascent primer strand in purple. Both I614 and F667 make close contacts with the incoming dNTP, shown in elemental colours.