

Published in final edited form as:

Angew Chem Int Ed Engl. 2010 August 9; 49(34): 5921–5924. doi:10.1002/anie.201001607.

Directed Evolution of DNA Polymerases for Next Generation Sequencing

Aaron M. Leconte,

Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037 (USA)

Maha P. Patel,

Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037 (USA)

Lauryn E. Sass,

Helicos BioSciences Corporation, One Kendall Square, Building 700, Cambridge, MA 02139 (USA)

Peter McInerney,

Helicos BioSciences Corporation, One Kendall Square, Building 700, Cambridge, MA 02139 (USA)

Mirna Jarosz,

Helicos BioSciences Corporation, One Kendall Square, Building 700, Cambridge, MA 02139 (USA)

Li Kung,

Helicos BioSciences Corporation, One Kendall Square, Building 700, Cambridge, MA 02139 (USA)

Jayson L. Bowers,

Helicos BioSciences Corporation, One Kendall Square, Building 700, Cambridge, MA 02139 (USA)

Philip R. Buzby,

Helicos BioSciences Corporation, One Kendall Square, Building 700, Cambridge, MA 02139 (USA)

J. William Efcavitch, and

Helicos BioSciences Corporation, One Kendall Square, Building 700, Cambridge, MA 02139 (USA)

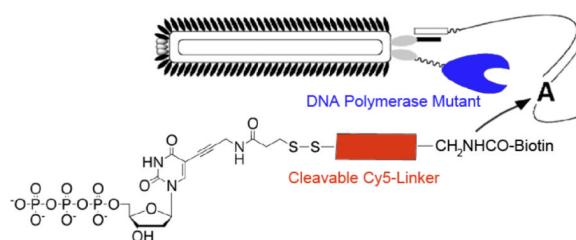
Floyd E. Romesberg*

Department of Chemistry, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA 92037 (USA)

Abstract

*Fax: (+1) 858-784-7472, floyd@scripps.edu, Homepage: <http://www.scripps.edu/chem/romesberg>.

Supporting information for this article is available on the WWW under <http://www.angewandte.org> or from the author.



We present the application of an activity-based phage display method to identify DNA polymerases tailored for next generation sequencing applications. Using this approach, we identify a mutant of Taq DNA polymerase that incorporates the fluorophore-labeled dA, dT, dC, and dG substrates ~50 to 400-fold more efficiently into scarred primers in solution and that also demonstrates significantly improved performance under actual sequencing conditions.

Keywords

DNA Sequencing; Taq DNA Polymerase; Directed Evolution; Nucleotides

Low cost whole genome sequencing technologies promise to revolutionize biomedical research and usher in an era of personalized medicine. The reductions in cost and time that are needed to make routine genome sequencing practical will require new methodologies, the most developed of which are “sequencing-by-synthesis” (SBS) methods that involve direct detection of polymerase-mediated synthesis.[1] Helicos BioSciences (Cambridge, MA) has commercialized a single-molecule SBS platform in which spatially separated oligonucleotides are sequenced in parallel through the use of total internal reflection illumination microscopy to detect the addition of deoxynucleotide triphosphates bearing fluorophores attached via cleavable linkers (dNTP-FI, Figures 1a and S1).[2] After incorporation and detection of the modified dNTP, the disulfide bond of the linker is cleaved to release the fluorophore, and after capping with iodoacetamide, a spectroscopically dark primer is regenerated which is then ready for an additional round of dNTP-FI incorporation. Removal of the fluorophore leaves behind a portion of the linker, referred to as a ‘scar’ (Figure 1b), and sequential rounds of sequencing eventually results in a run of modified nucleotides at the primer terminus. Ultimately, polymerase recognition of modified nucleotides, both during incorporation of the dNTP-FI, and after incorporation as part of scarred primer terminus, limits this and other SBS methods.

A variety of approaches have been pursued for developing polymerases with novel activities, including screening variants produced by rational design[3] or random mutagenesis,[4] and selections based on *in vitro* compartmentalization[5] or phage display. [6] In previous work,[6] we demonstrated that polymerase mutants with specific, non-natural catalytic properties can be isolated from large libraries of mutants through the use of an activity-based selection system that relies on the co-display of DNA mutants and their oligonucleotide substrate on M13 phage (Figure 1c). Phage production is optimized such that each phage particle displays one or zero polymerase mutants via fusion to a phagemid-encoded pIII, and four to five “acidic peptides” via fusions to the phage genome-encoded pIII. The displayed acidic peptides are used to attach oligonucleotide primers to the surface of the phage particle via a covalently linked “basic peptide.” Because all of the pIII proteins are localized to one end of the phage particle, a displayed polymerase preferentially extends the primers that are covalently attached to the same phage particle. Biotinylation of the dNTP substrate, natural or modified, enables selective recovery of the active polymerases and their respective genes using streptavidin beads. Thus, libraries can be enriched in mutants possessing a desired activity, such as the recognition of modified substrates. Using

this directed evolution approach we previously evolved polymerases that can synthesize RNA,[6a] DNA containing C2'-O-methyl modified nucleotides,[6b] or DNA containing unnatural hydrophobic nucleobase analogs.[6c] While the selection system is compatible with either Klenow fragment (Kf, the N-terminal truncation of *E. coli* polymerase I) or Stoffel fragment (Sf, the N-terminal truncation of Taq) DNA polymerases, all previous successes have been with Sf, which may be related to a link between thermostability and evolvability.[7] Also, initial efforts to evolve Kf mutants tailored for the Helicos nucleotides failed, thus our attention turned to the evolution of Sf.

A library of Sf mutants was generated by synthetic shuffling,[8] which involved assembly PCR with degenerate oligonucleotides encoding residues found in six homologous polymerases: *Thermus aquaticus*; *Thermus thermophilus* (91% amino acid identity to *T. aquaticus*); *Thermus caldophilus* (86%); *Thermus filiformis* (81%); *Spirochaeta thermophila* (54%); and *Thermomicrobium roseum* (54%). Mutations were restricted to twenty-one residues within 14 Å of the incoming dNTP (based on a ternary structure of Taq (PDB ID 1qsy[9])). This approach allowed many mutations to be introduced, and since every mutation is found in nature, simultaneously minimized the chances that any mutation would compromise the structure or basic activity of the enzyme. The resulting 14Å library contains mutations at 21 unique sites within the fingers and palm region, which are prominently involved in dNTP binding and incorporation,[10] resulting in a final library of 10⁸ chimeric Sf variants (for details, see Supporting Information).

To optimize selection pressure, we measured the steady-state rates at which Sf extends a natural primer by incorporation of each dNTP-FI against its cognate base in the template (see Table S1 in Supporting Information), and we found that the incorporation of dUTP-FI opposite dA is the least efficient. Thus, to apply a selection pressure for this step, we synthesized biotinylated dUTP-FI (Figure 1a). 10¹¹ phage bearing both a polymerase mutant and a primer-template duplex containing dA at the first templating position were prepared as described previously.[6] Four rounds of selection were performed where phage immobilization required the more efficient extension of the primer with the biotinylated dUTP-FI.

From a preliminary screen of 300 members of the enriched library, mutants were selected based on their ability to recognize dUTP-FI. Six mutants were further characterized based on their ability to recognize each different modified dNTP under both steady state (Table 1) and sequencing-like conditions using a scarred primer (Table S2). Six mutants were selected for further characterization under both steady state (Table 1) and sequencing-like conditions using a scarred primer (Table S2 in Supporting Information). The three most active polymerase mutants, Sf168, Sf197, and Sf267, showed an approximately 10- to 50-fold increased efficiency for dUTP-FI incorporation and a 7- to 80-fold increased efficiency for incorporation of the other three modified dNTPs.

Sf has relatively low affinity for DNA,[11] so its practical use for sequencing is limited by the need for prohibitively high concentrations of enzyme. Thus, the mutations found in Sf168, Sf197, and Sf267 (see Table S3), were cloned into full length Taq, which has a higher affinity for DNA.[11] As expected, the concentration of each Taq mutant needed to saturate the primer-template decreased more than 25-fold, making it likely that the selected mutants will be suitable for practical applications. A preliminary survey of dNTP-FI incorporation kinetics indicated that Taq197 (corresponding to the mutations from Sf197) was better optimized for the modified substrates than was Taq168 or Taq267 (corresponding to Sf168 and Sf267, respectively), thus our focus turned to the further characterization of Taq197. Using pre-steady state kinetics, we found that Taq197 incorporates each dNTP-FI 48- to 377-fold more efficiently into scarred primer termini than wild type Taq, with no

apparent bias toward the identity of the dNTP-FI or the sequence of the primer (Table 2). At least for dUTP-FI incorporation, the data also suggest that GC-rich sequences, which are commonly difficult to sequence, are not problematic.

To examine the fidelity of Taq197, we characterized the misincorporation of the three incorrect dNTP-FIs opposite dA in the template under pre-steady state conditions. Importantly, as with wild type Taq, Taq197 does not measurably synthesize any mispairs, even after 90 minutes. Thus, based on the detection limit of the assay (see Supporting Information), we set an upper limit of $5.6 \times 10^{-4} \text{ min}^{-1}$ for the rate of mispair formation, making correct incorporation of modified nucleotides more than 5,000-fold more efficient than mispair formation. These data suggest the fidelity of Taq197 has not been significantly compromised and that it should be sufficient for sequencing applications.

Finally, we evaluated the performance of Taq and Taq197 in single molecule sequencing reactions using a set of 30 oligonucleotides derived from the sequence of the M13 phage genome (Figure 2). After 32 base addition cycles (a total of 8 additions of each of the four dNTP-FIs), Taq produced a median strand length of two nucleotides. In contrast, Taq197 produced a median strand length of six nucleotides, with significantly longer lengths also observed. The conditions employed in these reactions were optimized for mesophilic polymerases, thus optimization for Taq197 will likely significantly increase the absolute read length. Regardless, the data reveal that the improved ability of Taq197 to accept the modified nucleotides translates into significantly improved performance in single molecule sequencing.

Taq197 has 14 mutations relative to its wild type progenitor (Figure 3 and Supporting Information). Based on the crystal structure of the ternary complex of the wild type enzyme and natural substrates,[9] it seems likely that at least one of the mutations, T664A, alters direct interactions with the modified substrates. Thr664 is located in the developing major groove of the DNA, 6.1 Å away from the site where the linker is attached to the incoming dNTP (Figure 3a); mutation to the smaller Ala residue may enable the polymerase to better tolerate the bulky linker. In addition to these direct interactions, there are a number of more subtle changes; five amino acids in the O-helix and three in the N-helix, which packs on the O-helix, are mutated to other hydrophobic residues (Figure 3b). These mutations appear to participate in three clusters of packing interactions that likely contribute to improved positioning of the O-helix, which has been shown to close over, and make specific contacts with, the incoming dNTP during DNA synthesis.[9–10,12] Thus, while additional experiments are required to fully deconvolute the specific contributions of individual mutations to the improved activities of Taq197, the polymerase appears to have acquired an expanded substrate repertoire by optimizing both direct and indirect contacts with the modified substrates.

Taq197 is the first example of a DNA polymerase optimized by directed evolution for next generation sequencing. Considering that many of the most promising next generation sequencing methods rely on DNA polymerase recognition of modified substrates, and that they would be aided by polymerase optimization, the methods detailed here should be broadly applicable to other next generation sequencing platforms. In addition to its benefit to sequencing methods, many other emerging technologies, including DNA labeling and SELEX, would be greatly facilitated by an increased ability to replicate DNA containing similarly modified nucleotides, and Taq197, or a further optimized progeny, is likely to facilitate these technologies, as well.[13]

Supplementary Material

Refer to Web version on PubMed Central for supplementary material.

Acknowledgments

This work was supported by National Institutes of Health award GM060005 (F.E.R.) and Helicos BioSciences Corporation.

References

1. a) Reviewed in Shendure J, Ji H. *Nat. Biotech.* 2008; 26:1135–1145. b) Bentley DR, et al. *Nature.* 2008; 456:53–59. [PubMed: 18987734]
2. a) Braslavsky I, Herbert B, Kartalov E, Quake SR. *Proc. Natl. Acad. Sci. USA.* 2003; 100:3960–3964. [PubMed: 12651960] b) Harris TD, et al. *Science.* 2008; 320:106–109. [PubMed: 18388294] c) Bowers J, et al. *Nat. Methods.* 2009; 6:593–595. [PubMed: 19620973]
3. a) Kranaster R, Marx A. *Nucleic Acids Symp. Ser. (Oxf.).* 2008;477–478. b) Kranaster R, Marx A. *Angew. Chem. Int. Ed. Engl.* 2009; 48:4625–4628. [PubMed: 19444846] c) Chen F, Gaucher EA, Leal NA, Hutter D, Havemann SA, Govindarajan S, Ortlund EA, Benner SA. *Proc. Natl. Acad. Sci. U.S.A.* 2010; 107:1948–1953. [PubMed: 20080675]
4. a) Camps M, Loeb LA. *Methods Mol. Biol.* 2003; 230:11–18. [PubMed: 12824565] b) Loh E, Choe J, Loeb LA. *J. Biol. Chem.* 2007; 282:12201–12209. [PubMed: 17301051]
5. a) Ghadessy FJ, Ong JL, Holliger P. *Proc. Natl. Acad. Sci. U.S.A.* 2001; 98:4552–4557. [PubMed: 11274352] b) Loakes D, Gallego J, Pinheiro VB, Kool ET, Holliger P. *J. Am. Chem. Soc.* 2009; 131:14827–14837. [PubMed: 19778048]
6. a) Fa M, Radeghieri A, Henry AA, Romesberg FE. *J. Am. Chem. Soc.* 2004; 126:1748–1754. [PubMed: 14871106] b) Leconte AM, Chen L, Romesberg FE. *J. Am. Chem. Soc.* 2005; 127:12470–12471. [PubMed: 16144377] c) Xia G, Chen L, Sera T, Fa M, Schultz PG, Romesberg FE. *Proc. Natl. Acad. Sci. USA.* 2002; 99:6597–6602. [PubMed: 12011423]
7. Bloom JD, Labthavikul ST, Otey CR, Arnold FH. *Proc. Natl. Acad. Sci. U.S.A.* 2006; 103:5869–5874. [PubMed: 16581913]
8. a) Minshull J, Govindarajan S, Cox T, Ness JE, Gustafsson C. *Methods.* 2004; 32:416–427. [PubMed: 15003604] b) Ness JE, Kim S, Gottman A, Pak R, Krebber A, Borchert TV, Govindarajan S, Mundorff EC, Minshull J. *Nat. Biotechnol.* 2002; 20:1251–1255. [PubMed: 12426575] c) Castle LA, et al. *Science.* 2004; 304:1151–1154. [PubMed: 15155947]
9. Li Y, Mitaxov V, Waksman G. *Proc. Natl. Acad. Sci. U.S.A.* 1999; 96:9491–9496. [PubMed: 10449720]
10. Loh E, Loeb LA. *DNA Repair (Amst.).* 2005; 4:1390–1398. [PubMed: 16230053]
11. Datta K, LiCata VJ. *Nucleic Acids Res.* 2003; 31:5590–5597. [PubMed: 14500822]
12. Li Y, Korolev S, Waksman G. *EMBO J.* 1998; 17:7514–7525. [PubMed: 9857206]
13. a) Turcatti G, Romieu A, Fedurco M, Tairi AP. *Nucleic Acids Res.* 2008; 36:e25. [PubMed: 18263613] b) Mitra RD, Shendure J, Olejnik J, Olejnik EK, Church GM. *Anal. Biochem.* 2003; 320:55–65. [PubMed: 12895469] c) Weisbrod SH, Marx A. *Chem. Commun. (Camb.).* 2008:5675–5685. [PubMed: 19009049]

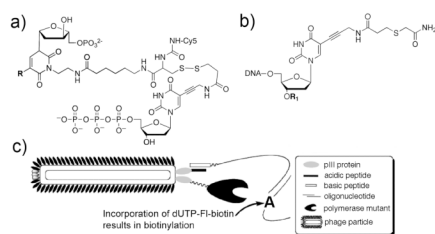


Figure 1.

SBS substrates and selection system. a) Modified dUTP used in sequencing (R=H) or selection experiments (R=CH₂NHCO-biotin). b) Representative scarred nucleotide. R₁ = H at primer terminus and DNA when in remainder of primer. c) Activity based phage display selection system. Active polymerase mutants incorporate a biotinylated nucleotide substrate into their intramolecularly attached oligonucleotide, enabling recovery using streptavidin beads.

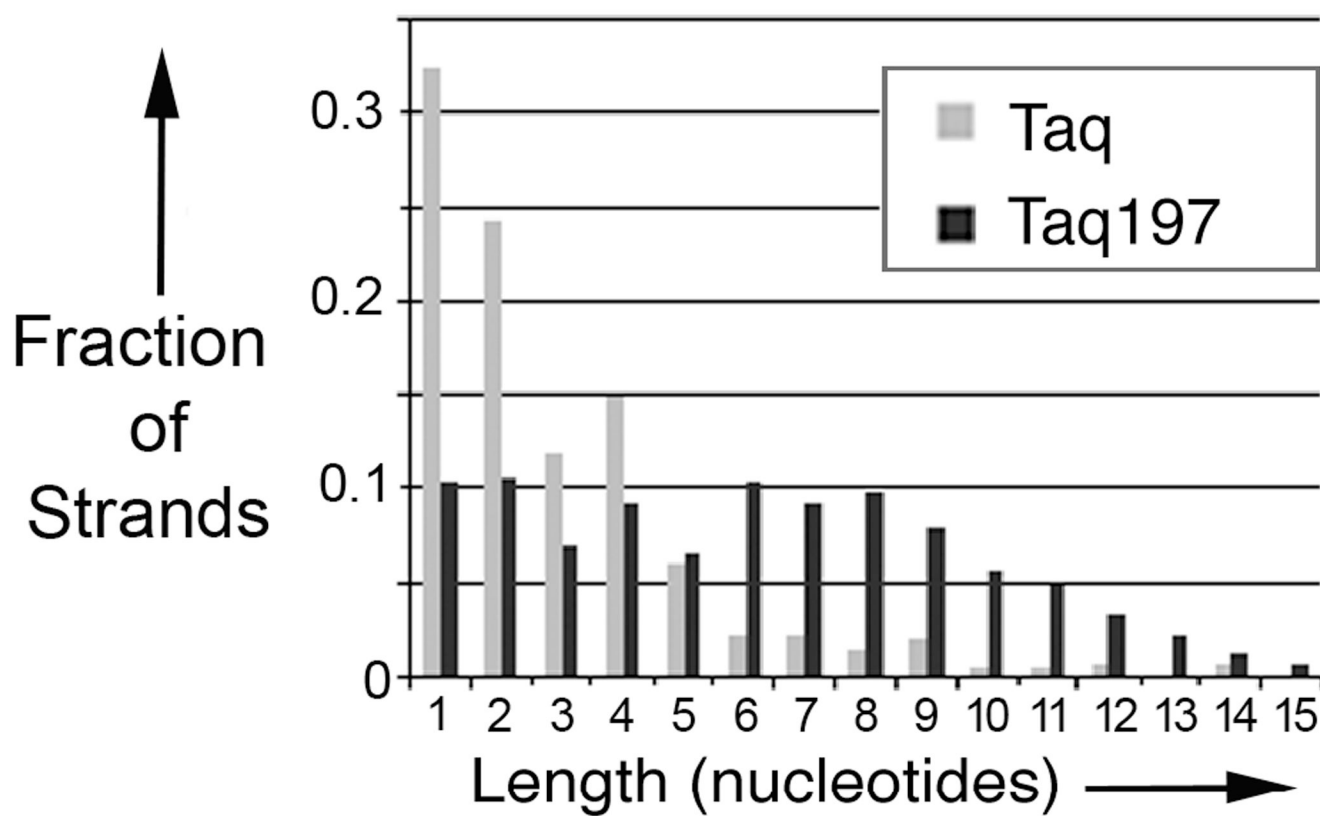


Figure 2. Performance of Taq197 (black) compared to wild type Taq (gray) in single molecule DNA sequencing after 32 base addition cycles (for details, see Supporting Information).

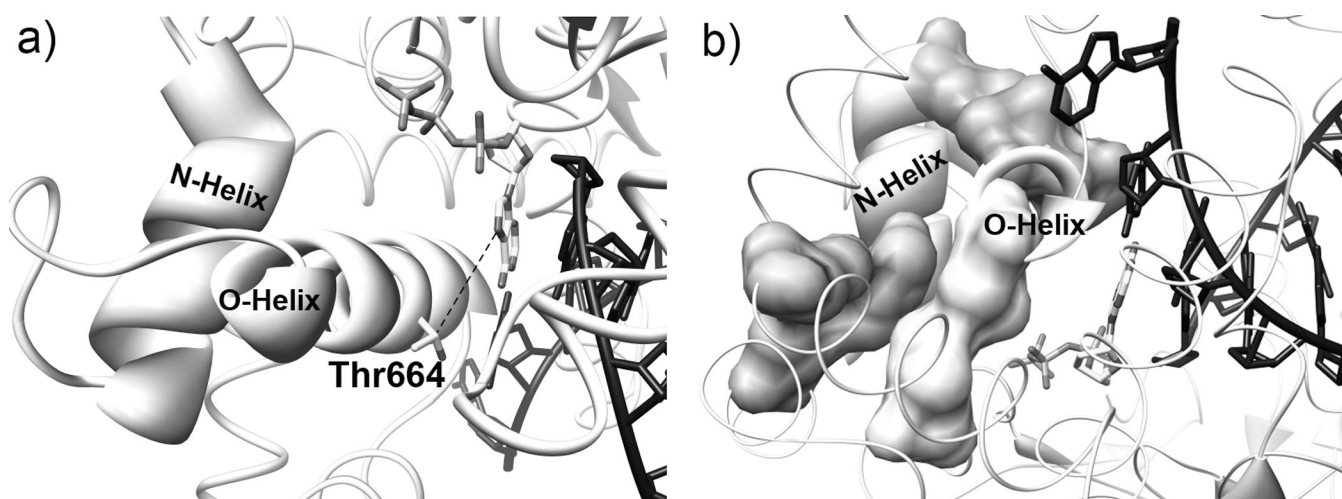


Figure 3. Taq polymerase (PDB ID 1qsy) showing residues mutated in Taq197. The N- and O-helices are labeled, the DNA strands are shown in black, and the incoming dNTP is rendered as stick. a) Thr664 is located 6.1 Å (dashed line) from the major groove of the incoming dNTP. b) Three distinct networks of mutated residues (surfaces highlighted) alter the packing between the N-helix and the O-helix, which packs on the incoming dNTP.

Table 1Steady-state rate of dUTP-FI incorporation by Sf wt and Sf mutants^[a]

5'-dTAATACGACTCACTATAGGGAGA
 3'-dATTATGCTGAGTGATATCCCTCTAGCTAGGTTACGGCAGGATCGC

Polymerase	k_{cat} (min ⁻¹)	K_{M} (μM ⁻¹)	$k_{\text{cat}}/K_{\text{M}}$ (min ⁻¹ M ⁻¹)	relative $k_{\text{cat}}/K_{\text{M}}^a$
Sf wt	1.7 ± 0.4	5.2 ± 1.0	3.3 × 10 ⁵	1.0
Sf281	6 ± 1.7	15.5 ± 5.2	3.9 × 10 ⁵	1.2
Sf292	12 ± 1	11.4 ± 1.8	1.1 × 10 ⁶	3.2
Sf247	12 ± 2	10.8 ± 0.9	1.1 × 10 ⁶	3.4
Sf197	15 ± 1	5.6 ± 2.3	2.7 × 10 ⁶	8.2
Sf267	15 ± 2	3.4 ± 1.3	4.4 × 10 ⁶	13.5
Sf168	21 ± 4	1.3 ± 0.4	1.6 × 10 ⁷	49.4

^[a] Error estimates (standard deviation) were determined from three independent measurements; see Supporting Information for reaction conditions.

Table 2

Rate of incorporation (k_{pol}) of dNTP-FI against cognate base (N_3) by Taq and Taq197 onto scarred primer termini^[a]

5'-TCTGCCCTGN₁N₁N₁

3'-AGACGGGACN₂N₂N₃TACTATCATTGTACTATCATTGTACTATCA^[b]

dNTP-FI	(N ₁ N ₁ N ₁ / N ₂ N ₂ N ₂) ^[c]	Taq (min ⁻¹)	Taq197 (min ⁻¹)	Fold Increase
N=A	UUU/ AAA	0.011 ± 0.002	1.4 ± 0.2	133
C	AAA/ UUU	0.002 ± 0.0002	0.66 ± 0.12	377
G	UUU/ AAA	0.14 ± 0.05	6.6 ± 1.8	48
U	CCC/ GGG	0.05 ± 0.02	2.9 ± 1.2	58

^[a]Error estimates (standard deviation) were determined from three independent measurements; conditions were as follows: 20 mM Tris pH 8.8, 10 mM NaCl, 10 mM KCl, 10 mM (NH₄)₂SO₄, 0.1% Triton X-100, 10 mM MgCl₂, 37 °C, 40 nM enzyme, 10 nM DNA, 100 nM dNTP-FI.

^[b]Primer and template have been truncated; complete sequences are shown in Table S2 (Supporting Information).

^[c]N₁ nucleotides are scarred (see Figure 1b).