SI Appendix

Efficient and Sequence-Independent Replication of DNA Containing a Third Base Pair Establishes a Functional Six-letter Genetic Alphabet

Denis A. Malyshev,[1] Kirandeep Dhami,[1] Henry T. Quach,[1] Thomas Lavergne,[1] Phillip Ordoukhanian,[2] Ali Torkamani,[3] and Floyd E. Romesberg[1]*

[1]Department of Chemistry, [2]Center for Protein and Nucleic Acids Research, [3]The Scripps Translational Science Institute, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, CA, 92037

E-mail: floyd@scripps.edu.

TABLE OF CONTENTS

Materials. Phosphoramidites and nucleosides of d**NaM** and d**5SICS** were purchased from Berry & Associates, Inc. (Dexter, MI). Phosphoramidites were used directly for DNA synthesis (see below), while nucleosides were first phosphorylated under Ludwig conditions (see Refs. 1 and 2). Biotinylated d**MMO2** (Fig. S5) was prepared as described previously (3). *Taq*, Deep Vent(exo–), Vent(exo–), Deep Vent, Vent, Phusion High Fidelity, and One*Taq* DNA polymerases were purchased from New England Biolabs (Ipswich, MA). KOD polymerase was purchased from EMD Chemicals (Billerica, MA). A mixture of dNTPs was purchased from Fermentas (Glen Burnie, MD). SYBR Green I Nucleic Acid Gel Stain (10,000×) was purchased from Invitrogen (Carslbad, CA). dsDNA was purified by DNA Clean & Concentrator-5 (Zymo Research Corp., Irvine, CA) without agarose gel purification unless stated otherwise and quantified by the Quant-iT dsDNA HS Assay (Invitrogen). Streptavidin Sepharose High Performance resin (*i.e.* "streptavidin beads") was purchased from GE Healthcare (Piscataway, NJ).

DNA synthesis. See Table S1 for complete oligonucleotide sequences. Fully natural primers were purchased from Integrated DNA Technologies (Coralville, Iowa). Reagents and solvents for synthesis of unnatural oligonucleotides were obtained from Glen Research (Sterling, VA) or Applied Biosystems (Foster City, CA). Templates `ACT`**`Y`**`GTG`, `GTC`**`Y`**`GGT`, `AGC`**`Y`**`CGT`, `CCG`**`Y`**`GAA`, and `NNN`**`Y`**`NNN` (**Y** = **5SICS,** N **=** randomized nucleotide) were prepared as dsDNA as previously reported (4). All other templates were prepared as ssDNA oligonucleotides (d**NaM** strand) using standard automated DNA synthesis with ultra-mild phosphoramidites (Glen Research) on controlled pore glass supports (0.20 μmol, 1000 Å, Glen Research) and an ABI Expedite 8905 synthesizer. After automated synthesis, the oligonucleotides were cleaved from the support, deprotected by incubation in conc. aqueous ammonia overnight at room temperature, and purified by 8 M denaturing 8% polyacrylamide gel electrophoresis followed by "crush and soak" overnight extraction with buffer (200 mM NaCl, 1 mM EDTA, 10 mM Tris pH 8), and finally desalted over Sephadex G-25 (NAP-10 Columns, GE Healthcare). The concentration of single stranded oligonucleotides was determined by UV absorption at 260 nm.

PCR conditions with different polymerases (Table S2). PCR was carried out in 1× ThermoPol buffer, except in the case of One*Taq* and *Phusion* for which One*Taq* standard buffer was used, and in the case of KOD polymerase for which KOD Buffer #1 (pH 8.0) was used, in a total volume of 50 μL with the following cycling conditions: initial denaturation 96 °C, 1 min; 96 °C, 10 s; 60 °C, 15 s; 68 °C, 1 or 4 min; 1 μM of each primer; 0.5× SYBR Green in an MyiQ Thermal Cycler (Bio-Rad). Other conditions are given in Table S2. Upon completion, a 5 μL aliquot was analyzed on a 2% agarose gel. The remaining solution was purified, quantified, and Sanger sequenced. Raw sequencing traces are shown in Figs. S1 and S2.

Sanger Sequencing. The cycle sequencing reactions (20 μL) were performed on a 9800 Fast Thermal Cycler (Applied Biosystems) with the Cycle Sequencing Mix (0.5 μl) of the BigDye Terminator v3.1 Cycle Sequencing Kit (Applied Biosystems) containing approximately 10 fmol (1 ng) template and 6 pmol primer. After 25 cycles of PCR (Initial denaturation 98 °C, 1 min; 98 °C, 10 s; 60 °C, 15 s; 68 °C, 1.5 min), the residual dye terminators were removed from the reaction with Agencourt CleanSEQ (Beckman-Coulter, Danvers, MA). The products were eluted off the beads with deionized water and sequenced directly on a 3730 DNA Analyzer (Applied Biosystems). Sequencing traces were collected using Applied Biosystems Data Collection software v3.0 and analyzed with the Applied Biosystems Sequencing Analysis v5.2 software. To avoid using specialized polymers and dye terminator reaction kits designed for short PCR products, we utilized a 5´-poly-dT-tailed sequencing primer strategy to improve the quality of the sequencing reads for the 130-150 nucleotide DNA templates (5), sequences of poly-dT primers are given in Table S1.

Analysis of Sanger Sequencing Traces. Sanger Sequencing traces was analyzed as described (4) to determine the retention of the unnatural base pair. Briefly, the presence of an unnatural nucleotide leads to a sharp termination of the sequencing profile, while mutation to a natural nucleotide results in "read-through" (for example, compare sequencing traces for KOD and Deep Vent polymerase in Fig. S1). The extent of this read-through after normalization is inversely correlated with the retention of the unnatural base pair. Raw sequencing traces were analyzed by first adjusting the start and stop points for the Sequencing Analysis software (Applied Biosystems) and then determining the average signal intensity individually for each channel (A, C, G and T) for peaks within the defined points (35-45 nucleotides in length). This was done separately for the parts of the sequencing trace before (section *L*) and after the unnatural nucleotide (section *R*). Calibration experiments were carried out with control mixtures containing defined percentages of different templates with and without the unnatural base pair (the mixtures contained 50-100% of the template with the unnatural base pair; see the Supporting Information of Ref. 4 for a complete description). The R/L ratio over the percentage of the natural template was plotted and was fit by linear regression. Equations for normalization of the *R*/*L* ratio that are given in Table S5 were used to account for read-through in the unamplified control templates as well as for the amplitude decay, which would otherwise result in the overestimation of fidelity. Finally, the retention of the unnatural base pair (*F*) was calculated as $1 - (R/L)_{norm}$ and the retention of the unnatural base pair per doubling (fidelity, *f*) was calculated as $F^{\frac{1}{\log_2 A}}$, where *A* is an amplification and $\log_2 A$ is the number of doublings. Each sample before and PCR amplification was sequenced in triplicate in each direction to minimize sequencing error (Fig. S3). The weighted mean retention in both directions was calculated (Table 1).

To obtain unamplified control sequencing data for the d**5SICS** strands of templates that were synthesized as d**NaM** stands, ssDNA (**dNaM**) was converted into dsDNA (d**5SICS**-d**NaM**) using the following protocol: ssDNA (0.75μM) was annealed with primer2 (8 μM) by heating to 95 °C, followed by cooling to room temperature over 2 hours in NEBuffer 2 buffer. 5'-overhangs were filled-in to form blunt ends in the presence of dNTPs (300 μM), d**5SICS**TP (100 μM) and Klenow (exo–) (0.05 units/μl) at 37 °C for 1 h. After quenching with EDTA to final concentration of 10 mM, the dsDNA was purified, quantified, and used as a control for sequencing with primer1 to determine the "read-through" that resulted during the sequencing reaction (and thus not during the PCR amplification).

Amplification of templates with two unnatural base pairs. Amplification efficiencies for the templates ACT**YY**GTG, ACT**Y**A**Y**GTG and ACT**Y**GTGACT**Y**GTG (**Y** = **5SICS**) were 96%, 95% and 93% with 1 min extension time and 87%, 82% and 93% with 4 min extension time, respectively. Efficiency (*E*) was calculated from $A = (1+E)^n$ (6), where *A* is the amplification level and *n* is the total number of PCR cycles. To estimate the fidelity of the amplification, we first quantified the sequencing peak height of the natural bases before and after the unnatural nucleotide in template as described above. The average retention in both directions for template ACT**Y**GTGACT**Y**GTG was 96.24% ± 0.09 and 97.32% ± 0.05 for 1 and 4 min, respectively. Assuming equivalent retention for each unnatural base pair, the retention per base pair is $1-\sqrt{1-0.9624} = 0.806$ (80.6%) and $1-\sqrt{1-0.9732} = 0.836$ (83.6%), respectively. This corresponds to 99.47% and 99.59% per doubling, respectively. The fidelities observed with each unnatural nucleotide of the ACT**YY**GTG and ACT**Y**A**Y**GTG templates are difficult to accurately quantify because of sequencing artifacts that cause a consistently detected ringing baseline signal (Fig. S3). Nonetheless, the most significant decrease in signal amplitude detected when sequencing in either direction was at the position

of the first unnatural nucleotide, demonstrating that both unnatural base pairs are retained at a level similar to that observed with the other templates.

PCR selection. For a description of the library construction, see the Materials and Methods section of the main text. The sublibraries (combined in a 1:1:1 ratio) were subjected to One*Taq* PCR amplification, as described above. A second set of reactions was also performed under the same conditions, but with a 4 min extension time. For each extension time, three samples were run: **Sample A** corresponds to the DNA library (0.025 ng; ~4.5×10^8 molecules) amplified without SYBR Green to avoid any possible inhibition and/or mutagenesis due to its intercalation with DNA; **Sample B** was identical to **Sample A,** but amplified with 0.5× SYBR Green to monitor the reaction progress and prevent overamplification and complications associated with correct duplex reannealing (if a library of dsDNA is denatured, reannealing will not be sequence-specific); and **Sample C** corresponds to a negative control lacking template but otherwise identical to **Sample B**. After 13 cycles, each PCR sample was diluted by a factor of 10^3, transferred to PCR tubes with fresh reagents and then subjected to 10 additional cycles (*i.e*. 10 cycles after 10^3 dilution), followed by 2× 20 cycles after 10^6 dilution, and finally 21 additional cycles after 10^6 dilution (84 PCR cycles in total, see Fig. S4*A* for qPCR data). **Samples A** from different levels of amplification were purified, quantified, and analyzed on 10% non-denaturing PAGE, and compared with **Samples B** to confirm similar amplification, while comparison with similarly purified and analyzed **Sample C** was used to confirm that no impurities were present (Fig. S4*B*).

To analyze the PCR reactions, each **Sample A** was subjected to an additional six cycles of PCR during which d**NaM**TP was replaced with biotinylated d**MMO2**TP ((see the Materials and Methods section of the main text, Fig. S5), followed by streptavidin gel-shift analysis (Fig. S7 and Ref. 3). Amplification level and the net retention of the unnatural base pair are given in Table S4. After the 10^{24}-fold amplification, the net retention of the unnatural base pair was 85% and 72% for the population amplified with 4 min and 1 min extension times, respectively. This includes the strands that retained the unnatural base pair at its originally incorporated position, as well as those that lost the originally incorporated unnatural base pair, but gained it via misinsertion of an unnatural triphosphate at another position (during either the amplification or biotinylation step). To determine the retention that resulted from misinsertion, we performed an identical PCR amplification with analogous, fully natural templates. After the full 10^{24}-fold amplification with both 1 and 4 min extension times, thirteen percent of the natural library showed a streptavidin band-shift, indicating (after calibration, see **Calibration of gel shift mobility assay** below and Fig. S8) that 30% of the strands acquired the unnatural base pair (Fig. S7*B*). This corresponds to an error-rate of 5×10^{-5} misinsertions of an unnatural triphosphate for each natural base pair synthesized. The percentage of the population that retained the originally incorporated unnatural base pair was 79% and 60%, for the 4 min and 1 min extension times, respectively. Correspondingly, the fraction of DNA strands that lost the unnatural base pair at its original position but gained it at another, even after the full 10^{24}-fold amplification, is 0.06 ((1-0.79) × 0.30) and 0.12 ((1-0.60) × 0.30) for the 4 min and 1 min extension times, respectively. This small contribution to the total population will not significantly affect the sequence analysis of the amplified libraries.

Calibration of gel shift mobility assay. To quantify the net retention of the unnatural base pair at different levels of amplification, DNA mixtures with a known ratio of unnatural and natural templates were subjected to biotinylation by PCR (see Material and Methods in the main text) and analyzed by mobility-shift assay on 6 % non-denaturing PAGE. Each experiment was run in triplicate (a representative gel assay is shown on Fig. S8*B*), and the shift in mobility (*SM*) was plotted as function of the fraction of

the DNA with an unnatural base pair (Fig. S8*A*). Data points were fit with a quadratic equation, $SM = 0.492U^2 + 0.219U + 0.026$, with R^2 of 0.997. *U* was then calculated as

$$U = \frac{-0.219 + \sqrt{0.219^2 - 4 \times 0.492 \times (0.026 - SM)}}{2 \times 0.492}.$$

Determination of error rate via gel-shift assay. The DNA populations at the different amplification levels were purified, quantified, analyzed by streptavidin mobility gel shift, and normalized using calibration data (see **Calibration of gel shift mobility assay**) to determine unnatural base pair net retention. The net retention (1-*L*) was plotted as a function of $\log_2$(amplification) and the unnatural to natural nucleotide mutation rate was determined via linear regression, from which the net loss per doubling (*l*) was calculated (Fig. S7*A*). To determine the natural to unnatural nucleotide mutation rate, the unnatural base pair in the library used for the PCR selection (1:1:1 mixture of three sublibraries) was converted to a natural base pair via PCR with only natural dNTPs (see **Deep Sequencing** for conditions). The resulting natural library was subjected to the same 84 cycles of PCR selection (see Material and Methods in the main text for conditions) and the DNA populations at different levels of amplification were analyzed by streptavidin mobility gel-shift assay and normalized using the calibration equation described above to obtain the net gain, *G*, of the unnatural base pair by the natural template. Thus, (1–*G*)/*n* corresponds to the retention of given natural nucleotide, where *n* is the length of the portion of the DNA replicated. ln(1–*G*) was plotted as a function of $\log_2$(amplification) to obtain ln(1–*g*) via linear regression, where *g* is the net gain of the unnatural base pair per doubling (Fig. S7*B*). Finally, the error rate (natural to unnatural) was obtained by dividing 1–*g* by *n*.

Libraries separation on streptavidin beads. Streptavidin beads were centrifuged to remove storage solution and then washed with 100 μL of 1× Streptavidin binding buffer (SA-BB, 1× = 50 mM Na HEPES, pH 7.5, 150 mM NaCl, 1 mM EDTA), followed by 100 μL of 1×SA-BB with 10 μM of the Primer1 to block non-specific DNA binding and, finally, with 100 μL of 1×SA-BB. Biotinylated dsDNA library (30 ng) was combined with Streptavidin binding buffer (1× SA-BB, 50 mM Na HEPES, pH 7.5, 150 mM NaCl, 1 mM EDTA), applied to the prewashed beads, and incubated at 25 °C for 1 h with occasional hand-mixing. The bead slurry was separated from the supernatant (containing unbound DNA that lost unnatural base pair) using a centrifuge tube filter with a 0.22 μM nylon membrane (Corning Costar Spin-X, Lowell, MA) by centrifugation at 10,000×*g* for 5 min. Beads with bound biotinylated DNA (that retained the unnatural base pair) were washed first with 1×SA-BB (100 μL), then with mildly denaturing buffer (2×100 μL of 1 M guanidinium chloride, 50 mM Na HEPES, pH 7.5, 150 mM NaCl, 1 mM EDTA, 0.05% [v/v] Triton X-100) and 2×100 μL of 10 mM Tris HCl (pH 7.5), and finally the bound DNA was released by treatment with DTT (50 mM) at 37 °C for 1 h.

Deep Sequencing. To make it possible to use standard sequencing methodologies, DNA products that retained the unnatural base pair after amplification were converted to all natural sequences. To do this and to label each of the populations with a population-specific barcode so that all of the populations could be sequenced together, all of the populations (23 in total, Table S3) were subjected to PCR amplification with natural dNTPs and with a population-specific sequencing primer Primer1-DS-NNNNNN containing a six-nucleotide barcode (see Tables S1 and S3 for primer sequences) under the following conditions: 50 μL of 1× One*Taq* Standard reaction buffer with Mg^{2+} concentration adjusted to 5 mM, 500 μM of each natural dNTP, 1 μM of Primer1-DS-NNNNNN and Primer2-DS, 0.5× SYBR Green and 0.02 unit/μL of One*Taq* in an MyiQ Thermal Cycler (Bio-Rad) under following thermal cycling conditions: 96 °C, 30 s;

60 °C, 30 s; 68 °C, 8 min; 10 cycles. Following PCR, 1 μL of ExoSAP-IT (Affymetrix) was added and the mixture was incubated for 30 min at 37 °C to degrade any unused long primers. Lastly, the products were purified on 4% agarose gel. The product fragment (210 bp) was excised, purified, and quantified. All DNA populations labeled with a population-specific barcode were pooled in an equimolar ratio, and their purity was confirmed using a 2100 Bioanalyzer (Agilent Technologies). The equimolar DNA pool was sequenced using standard Illumina protocols, in which the DNA was denatured and diluted to 6 pM, loaded onto a version 2 Illumina flowcell, and sequenced in a single read for 100 bases on a HiSeq2000 (Illumina). The population-specific barcode sequences were determined in an independent 7 base read.

References

1. Ludwig J., Eckstein F. (1989) Rapid and efficient synthesis of nucleoside 5'-0-(1-thiotriphosphates), 5'-triphosphates and 2',3'-cyclophosphorothioates using 2-chloro-4H-1,3,2-benzodioxaphosphorin-4-one. *J Org Chem* 54:631.
2. Lavergne T., Malyshev D. A., Romesberg F. E. (2012) Major groove substituents and polymerase recognition of a class of predominantly hydrophobic unnatural base pairs. *Chemistry* 18:1231-1239.
3. Seo Y. J., Malyshev D. A., Lavergne T., Ordoukhanian P., Romesberg F. E. (2011) Site-specific labeling of DNA and RNA using an efficiently replicated and transcribed class of unnatural base pairs. *J Am Chem Soc* 133:19878-19888.
4. Malyshev D. A., Seo Y. J., Ordoukhanian P., Romesberg F. E. (2009) PCR with an Expanded Genetic Alphabet. *J Am Chem Soc* 131:14620-14621.
5. Binladen J., Gilbert M. T., Campos P. F., Willerslev E. (2007) 5'-tailed sequencing primers improve sequencing quality of PCR products. *Biotechniques* 42:174-176.
6. Cha R. S., Thilly W. G. (1993) Specificity, efficiency, and fidelity of PCR. *PCR Methods Appl* 3:S18-29.

Table S1. Sequences of DNA used in this study. Primer binding regions are underlined. **Y = 5SICS**.

Name	Sequence (5' to 3')	Remarks
Primer1	CACACAGGAAACAGCTATGAC	Primer for PCR
Primer2	GAAATTAATACGACTCACTATAGG	Primer for PCR
Primer1-poly-dT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT TTCACACAGGAAACAGCTATGAC	Primer for Sanger sequencing†
Primer2-poly-dT	TTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTTT TTTTGAAATTAATACGACTCACTATAGG	Primer for Sanger sequencing†
Primer1-DS-NNNNNN	CAAGCAGAAGACGGCATACGAGAT**NNNNNN**GTGACTGGAGTTCAGACGTGTGCTCTT CCGATCTGGAAACAGCTATGACGAATTC	Primers for deep sequencing with a population-specific barcode shown in bold (23 primers in total, see Table S3)
Primer2-DS	AATGATACGGCGACCACCGAGATCTACACTCTTTCCCTACACGACGCTCTTCCGATC TACGACTCACTATAGGAAGC	Universal primer for deep sequencing
ACT**Y**GTG	TTTCACACAGGAAACAGCTATGACCCGGGTTATTACATGCGCTGGCACTTGCCCGTA CGGCGGTTGCACT**Y**GTGATGGGGACCGGCTTCTTGGAGCCCATGGTATATCTCCTTC TTAAAGTTAACCCTATAGTGAGTCGTATTAATTTC	
GTC**Y**GGT	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGGTCCCGTGAAGCTCGTC**Y**GGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	
AGC**Y**CGT	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGGTCCCGTGAAGC**Y**CGTCTGGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	
CCG**Y**GAA	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGGTCCCG**Y**GAAGCTCGTCTGGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	
NNN**Y**NNN	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGGTCCCGNNN**Y**NNNCGTCTGGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	N = randomized natural nucleotide
GTA**Y**TGT	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGTTTCCTTAAAGATAGTA**Y**TGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	
AGA**Y**AGT	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGTTTCCTTAAAGA**Y**AGTATTGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	
CCT**Y**AAA	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGTTTCCT**Y**AAAGATAGTATTGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	
GGT**Y**TCC	GAAATTAATACGACTCACTATAGGGTTAAGCTTAACTTTAAGAAGGAGATTTACTAT GGT**Y**TCCTTAAAGATAGTATTGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGG TCATAGCTGTTTCCTGTGTG	
ACT**YY**GTG	AGAAATTAATACGACTCACTATAGGGTTAAGCTTAACAAGGAGATTTACTATGGATT ACT**YY**GTGATAGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGGTCATAGCTGT TTCCTGTGTG	
ACT**Y**A**Y**GTG	AGAAATTAATACGACTCACTATAGGGTTAAGCTTAACAAGGAGATTTACTATGGATT ACT**Y**A**Y**GTGATAGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGGTCATAGCTG TTTCCTGTGTG	
ACT**Y**GTGACT**Y**G TG	AGAAATTAATACGACTCACTATAGGGTTAAGCTTAACAAGGAGATTTACTATGGATT ACT**Y**GTGACT**Y**GTGATAGTGAATTCCAAGTGCTAGCGCATGTAATAACCCGGGTCAT AGCTGTTTCCTGTGTG	
Sublibrary1	CACACAGGAAACAGCTATGACGAATTC***GA*** N_{30} **Y** N_{10} ***GA***AAGCTTCCTATAGTGAGTCGTATTAATTTC	N = randomized natural nucleotide. Sublibrary-specific barcode sequences are bold and italicized
Sublibrary2	CACACAGGAAACAGCTATGACGAATTC***CT*** N_{20} **Y** N_{20} ***CT***AAGCTTCCTATAGTGAGTCGTATTAATTTC	
Sublibrary3	CACACAGGAAACAGCTATGACGAATTC***AG*** N_{10} **Y** N_{30} ***AG***AAGCTTCCTATAGTGAGTCGTATTAATTTC	

†Binladen J., Gilbert M. T., Campos P. F., Willerslev E. (2007) 5'-tailed sequencing primers improve sequencing quality of PCR products. *Biotechniques* 42:174-176.

Table S2. Polymerase screen.

Polymerase	dNTP (μM)	Amplification	Retention (%)[a]	Fidelity (%)[b]
Taq[c]	700	1.4×10^2	45	89.4
Deep Vent (exo–)[c]	700	1.2×10^2	83	97.4
Deep Vent[c]	700	0.74×10^2	88	97.9
Vent (exo–)[c]	700	2.3×10^2	48	91.0
Vent[c]	700	1.0×10^2	85	97.6
KOD[c]	700	0.22×10^2	17	67.0
Phusion[c]	700	1.5×10^2	93	99.0
Deep Vent[d]	200	na[e]	-	-
Taq[d]	200	3.2×10^3	48	93.9
Taq + Deep Vent[d]	200	5.0×10^3	99	99.9
Taq + Vent[d]	200	4.3×10^3	99	99.9
Taq + Phusion[d]	200	2.6×10^3	95	99.6

[a]Percentage of amplified product that retained the unnatural base pair, calculated as the average retention determined by sequencing in both directions.

[b]Fidelity (*f*) was determined by sequencing (see Materials and Methods) and is defined as the retention of the unnatural base pair per doubling, calculated as $R = f^n$, where R is the retention of the unnatural base pair and n is the number of doublings ($n = \log_2(A)$, and A is the amplification.

[c]Conditions: 0.2 ng of NNN**Y**NNN template (**Y** = **5SICS**, see Table S1 for full sequence), d**5SICS**TP and d**NaM**TP = 100 μM, 6 mM $MgSO_4$, 0.02 U/μL polymerase, cycling conditions: 96 °C, 10 s; 60 °C, 15 s; 68 °C, 4 min, 14 cycles.

[d]Conditions: 0.02 ng of the NNN**Y**NNN template (**Y** = **5SICS**), d**5SICS**TP and d**NaM**TP = 100 μM, 3 mM $MgSO_4$, 0.01 U/μL exo(+) polymerase and/or 0.02 U/μL *Taq*, cycling conditions: 96 °C, 10 s; 60 °C, 15 s; 68 °C, 1 min, 20 cycles.

[e]no amplification product detected.

Table S3. Population-specific six-nucleotide barcodes within Primer1-DS-NNNNNN (see Table S1 for full sequence). As described in the text, each population isolated during the PCR selection was labeled with a unique barcode (see **Deep Sequencing** in the SI Materials and Methods) for deep sequencing and subsequent binning by population. ubp = unnatural base pair.

Population-specific barcode	DNA sample		
	Amplification	**Extension time**	**Population**
CGTGAT	none[a]	-	ubp retained
ACATCG	none[a]	-	ubp lost
GCCTAA	10^3	1 min	ubp retained
TGGTCA	10^3	1 min	ubp lost
CACTGT	10^3	4 min	ubp retained
ATTGGC	10^3	4 min	ubp lost
GATCTG	10^6	1 min	ubp retained
TCAAGT	10^6	1 min	ubp lost
CTGATC	10^6	4 min	ubp retained
AAGCTA	10^6	4 min	ubp lost
GTAGCC	10^{12}	1 min	ubp retained
TACAAG	10^{12}	1 min	ubp lost
TTGACT	10^{12}	4 min	ubp retained
GGAACT	10^{12}	4 min	ubp lost
TGACAT	10^{18}	1 min	ubp retained
GGACGG	10^{18}	1 min	ubp lost
CTCTAC	10^{18}	4 min	ubp retained
GCGGAC	10^{18}	4 min	ubp lost
TTTCAC	10^{24}	1 min	ubp retained
GGCCAC	10^{24}	1 min	ubp lost
CGAAAC	10^{24}	4 min	ubp retained
CGTACG	10^{24}	4 min	ubp lost
CCACTC	none[b]	-	-

[a] Naïve library was biotinylated, separated on the beads and sequenced after the unnatural base pair was converted to a natural one.

[b] Naïve library was sequenced directly (the unnatural base pair was converted to a natural).

Table S4. Amplification and net retention of the d**5SICS**-d**NaM** base pair during PCR selection.

# of PCR cycles	**Extension time (min)**	**Amplification**	**Net retention of unnatural base pair (%)**
13	1	1.4×10^3	94
	4	2.0×10^3	98
23	1	1.7×10^6	92
	4	1.8×10^6	96
43	1	1.3×10^{12}	81
	4	2.1×10^{12}	89
63	1	1.3×10^{18}	79
	4	2.0×10^{18}	84
84	1	1.9×10^{24}	72
	4	3.6×10^{24}	85

Table S5. Equations used for normalization of *R*/*L* values for different templates. **Y** = **5SICS**

Template	Direction	Read-through[a]	Equation used for data analysis[b]
ACT**Y**GTG	Primer1	1.8	$R/L = 0.55(R/L)_{norm} + 1.8$
	Primer2	3.7	$R/L = 0.55(R/L)_{norm} + 3.7$
GTC**Y**GGT	Primer1	1.8	$R/L = 0.55(R/L)_{norm} + 1.8$
	Primer2	1.6	$R/L = 0.55(R/L)_{norm} + 1.6$
AGC**Y**CGT	Primer1	1.0	$R/L = 0.55(R/L)_{norm} + 1.0$
	Primer2	1.4	$R/L = 0.55(R/L)_{norm} + 1.4$
CCG**Y**GAA	Primer1	1.5	$R/L = 0.55(R/L)_{norm} + 1.5$
	Primer2	1.0	$R/L = 0.55(R/L)_{norm} + 1.0$
NNN**Y**NNN	Primer1	2.0	$R/L = 0.55(R/L)_{norm} + 2.0$
	Primer2	3.3	$R/L = 0.55(R/L)_{norm} + 3.3$
GTA**Y**TGT	Primer1	3.2	$R/L = 0.55(R/L)_{norm} + 3.2$
	Primer2	2.1	$R/L = 0.55(R/L)_{norm} + 2.1$
AGA**Y**AGT	Primer1	5.9	$R/L = 0.55(R/L)_{norm} + 5.9$
	Primer2	10.6	$R/L = 0.55(R/L)_{norm} + 10.6$
CCT**Y**AAA	Primer1	1.9	$R/L = 0.55(R/L)_{norm} + 1.9$
	Primer2	3.7	$R/L = 0.55(R/L)_{norm} + 3.7$
GGT**Y**TCC	Primer1	13.0[c]	-[c]
	Primer2	1.0	$(R/L) = 0.55(R/L)_{norm} + 1.0$

[a] Read-through with the control unamplified template.
[b] Determined from linear fit of calibration data with *y*-intercept corresponding to read-through with unamplified control template.
[c] Only one direction (Primer2) was used to gauge PCR fidelity for the GGT**Y**TCC template due to high read-though in the direction of Primer1 that was also observed with the unamplified control template and thus an artifact of the Sanger sequencing.

Fig. S1. Raw sequencing traces for PCR with a single DNA polymerase. Conditions: 0.2 ng of the NNN**Y**NNN template (**Y** = **5SICS)**; dNTPs = 700 µM, d**NaM**TP/d**5SICS**TP = 100/100 µM, 6 mM $MgSO_4$, 0.02 U/µL of the polymerase, 4 min extension time, 14 cycles. Other conditions are given is Table S1. The position of the unnatural nucleotide is indicated with red arrow.

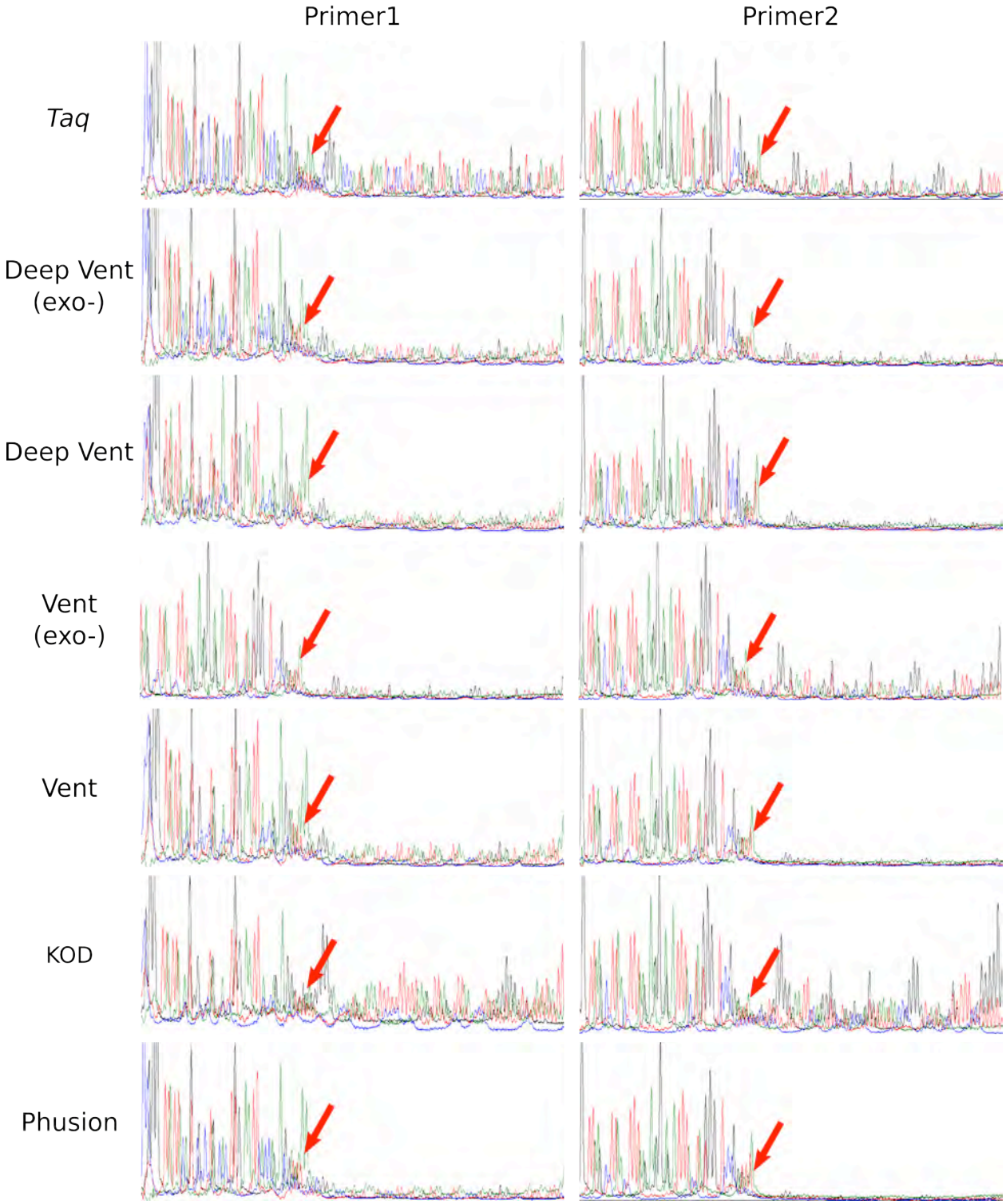


Fig. S2. Raw sequencing traces for PCR with *Taq* and an exo(+) DNA polymerase. Conditions: 0.02 ng of the NNN**Y**NNN template (**Y** = **5SICS)**; dNTPs = 200 µM, d**NaM**TP/d**5SICS**TP = 100/100 µM, 3 mM $MgSO_4$, 0.02 U/µL of *Taq* and/or 0.01 U/µL of exo(+) polymerase, 1 min extension time, 20 cycles. Other conditions are given is Table S1. The position of the unnatural nucleotide is indicated with red arrow.

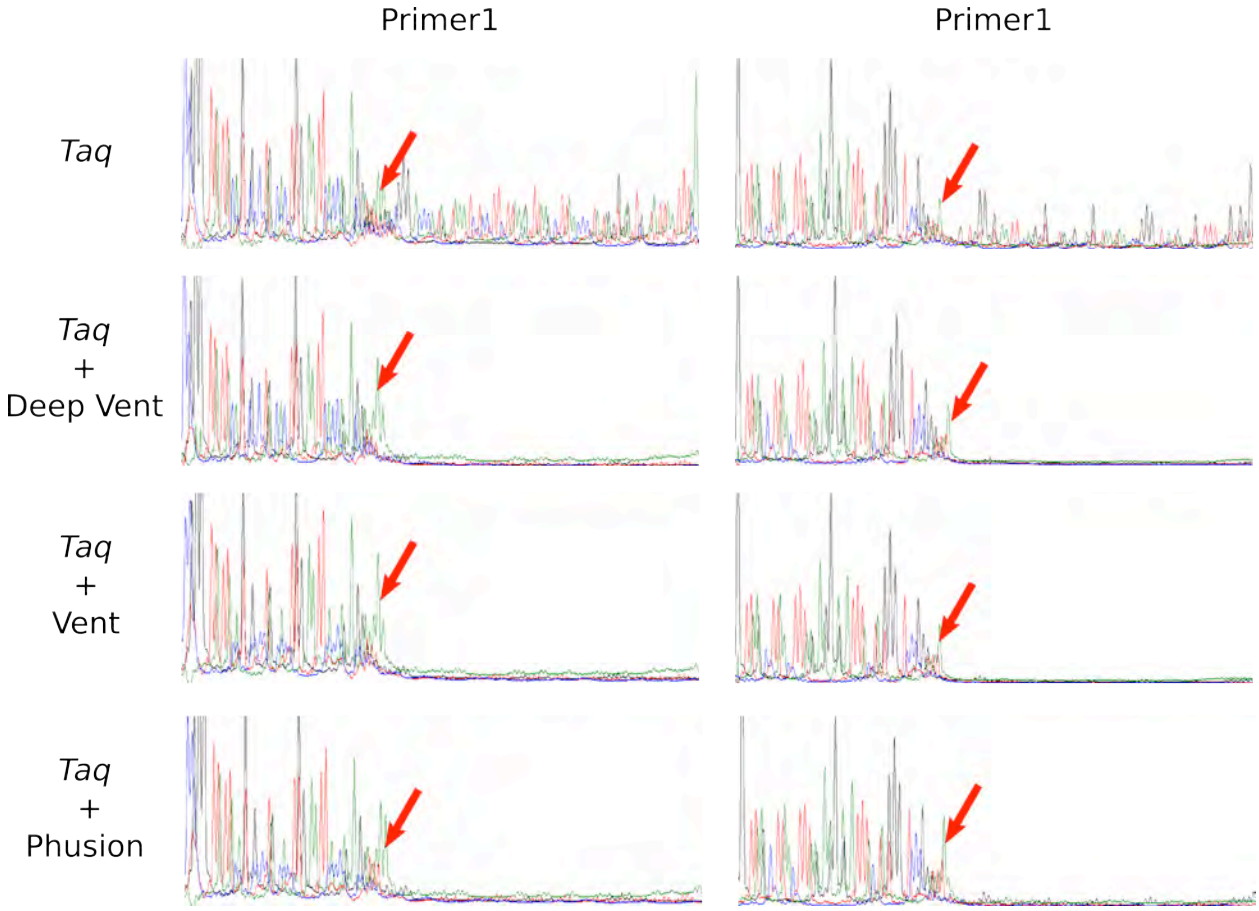


Fig. S3. Raw sequencing traces for PCR with One*Taq* DNA polymerase and different templates. See **One*Taq* PCR** in the Materials and Methods in the main text for conditions. Triplicate data shown. The position(s) of the unnatural nucleotide is(are) indicated with red arrow(s). **Y** = **5SICS**

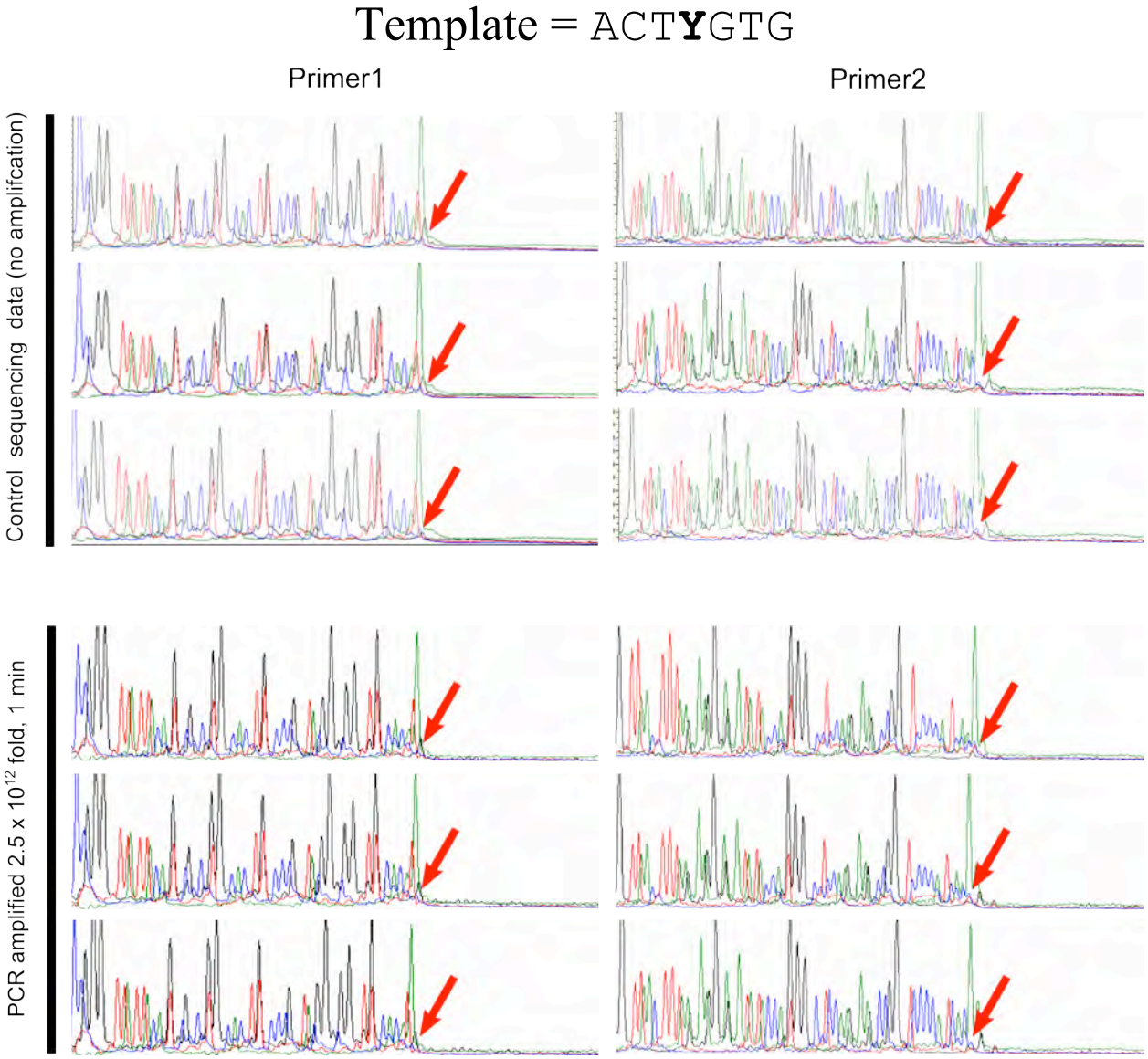


Fig. S3 cont.

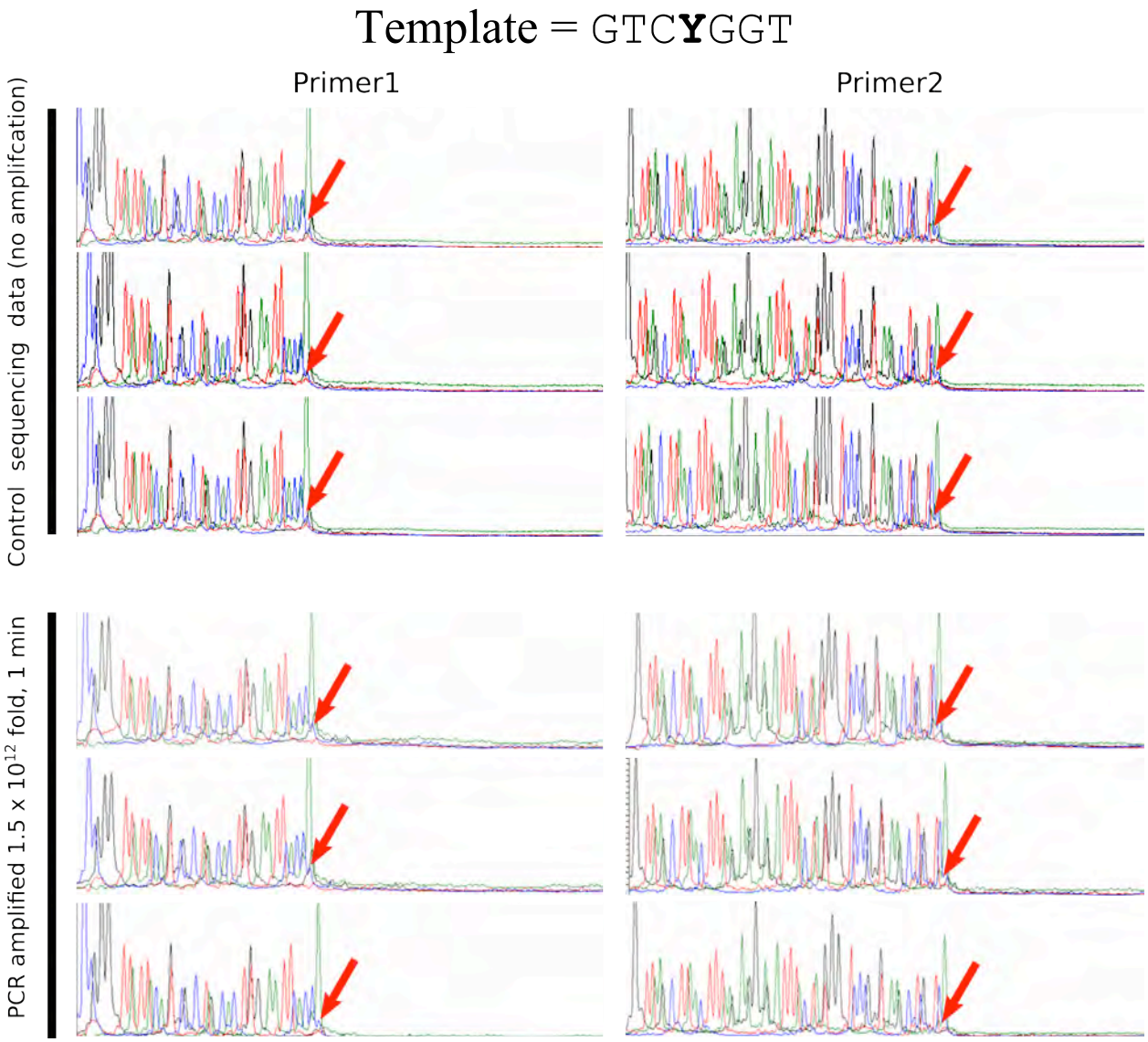

Template = GTCYGGT
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 1.5 x 10^12 fold, 1 min

Fig. S3 cont.

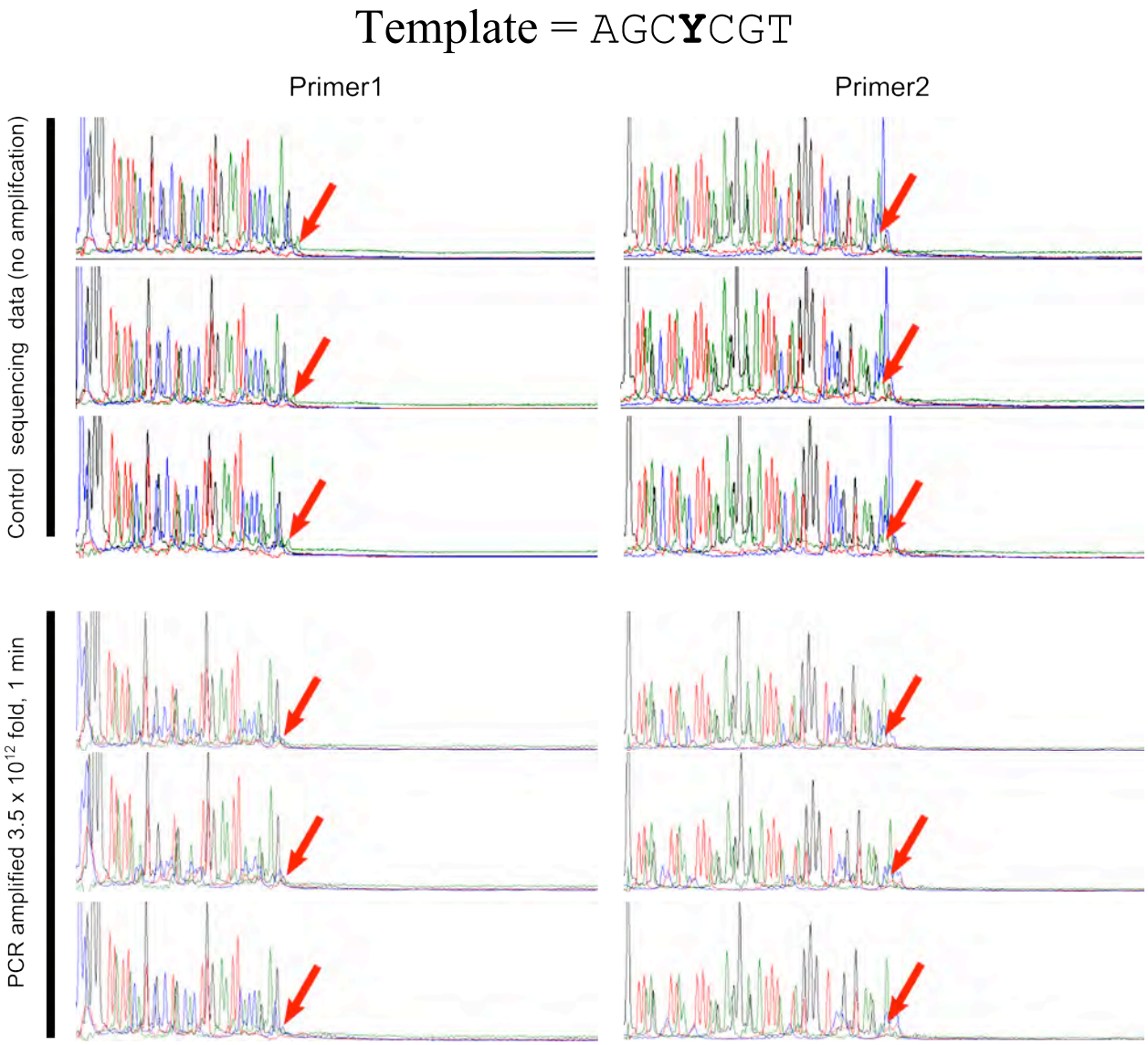

Template = AGCYCGT
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 3.5 x 10^12 fold, 1 min

Fig. S3 cont.

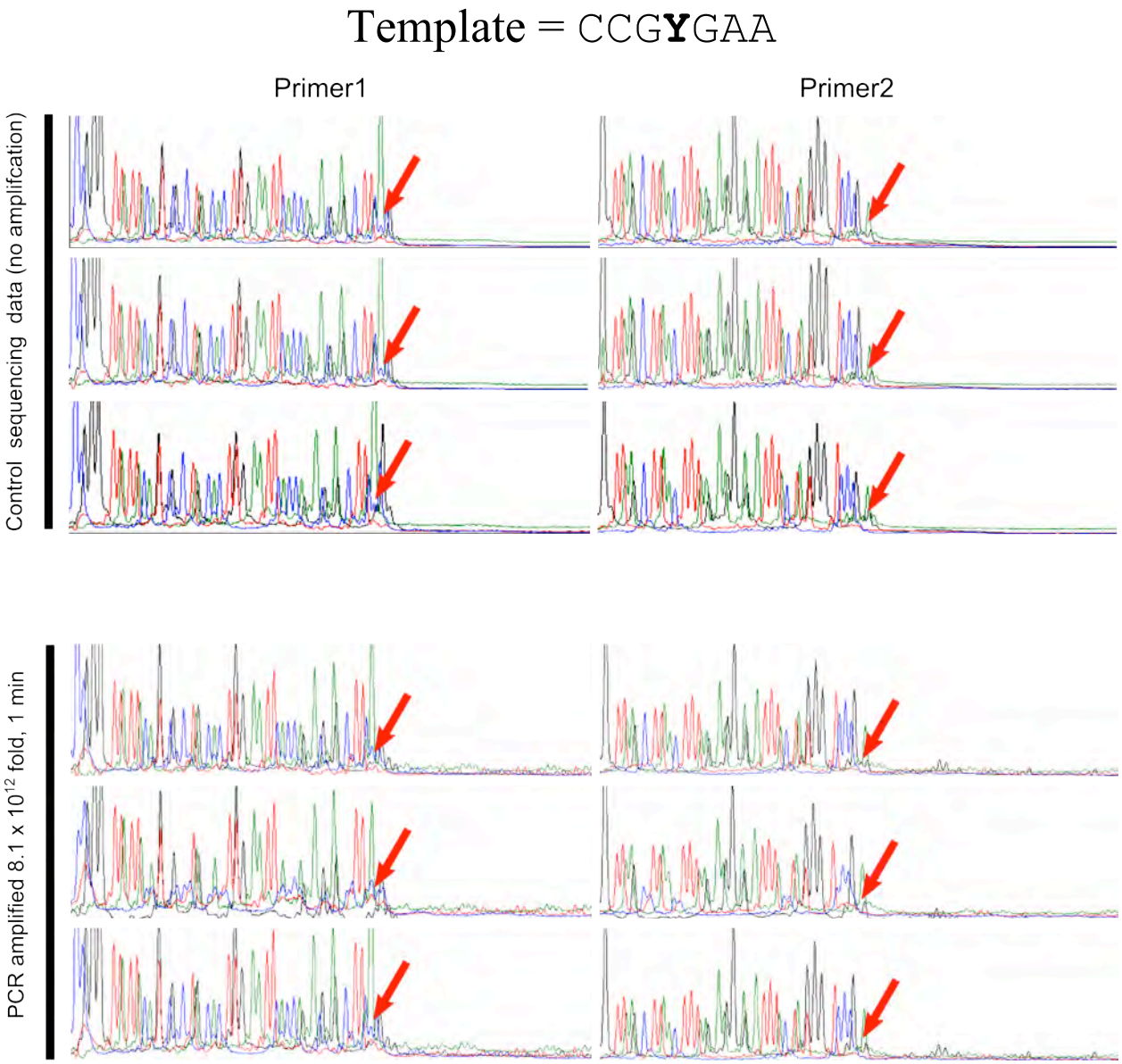

Template = CCGYGAA
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 8.1 x 10^12 fold, 1 min

Fig. S3 cont.


Template = NNNYNNN
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 4.8 x 10^12 fold, 1 min
PCR amplified 1.6 x 10^13 fold, 4 min

Fig. S3 cont.

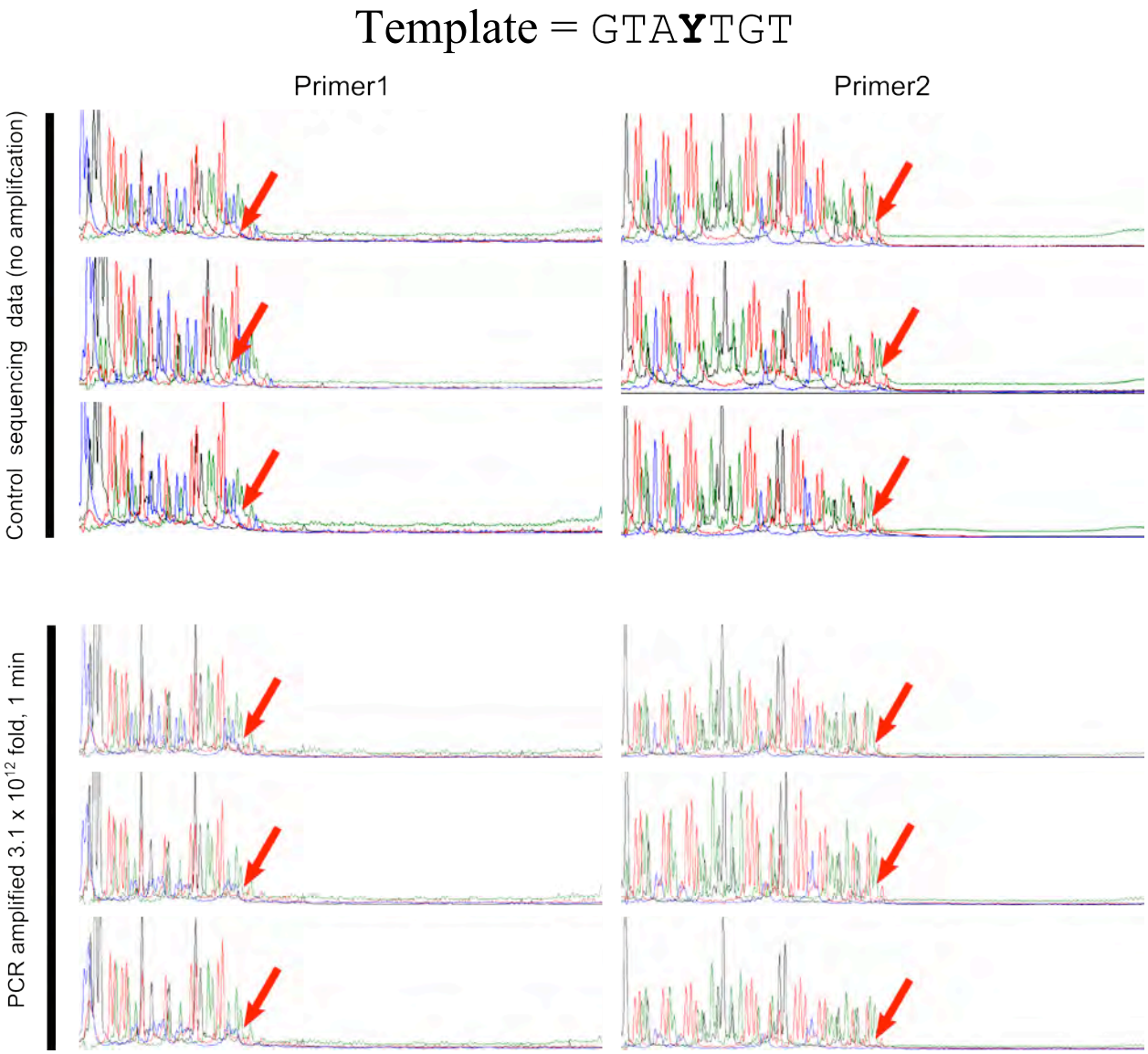

Template = GTAYTGT
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 3.1 x 10^12 fold, 1 min

Fig. S3 cont.

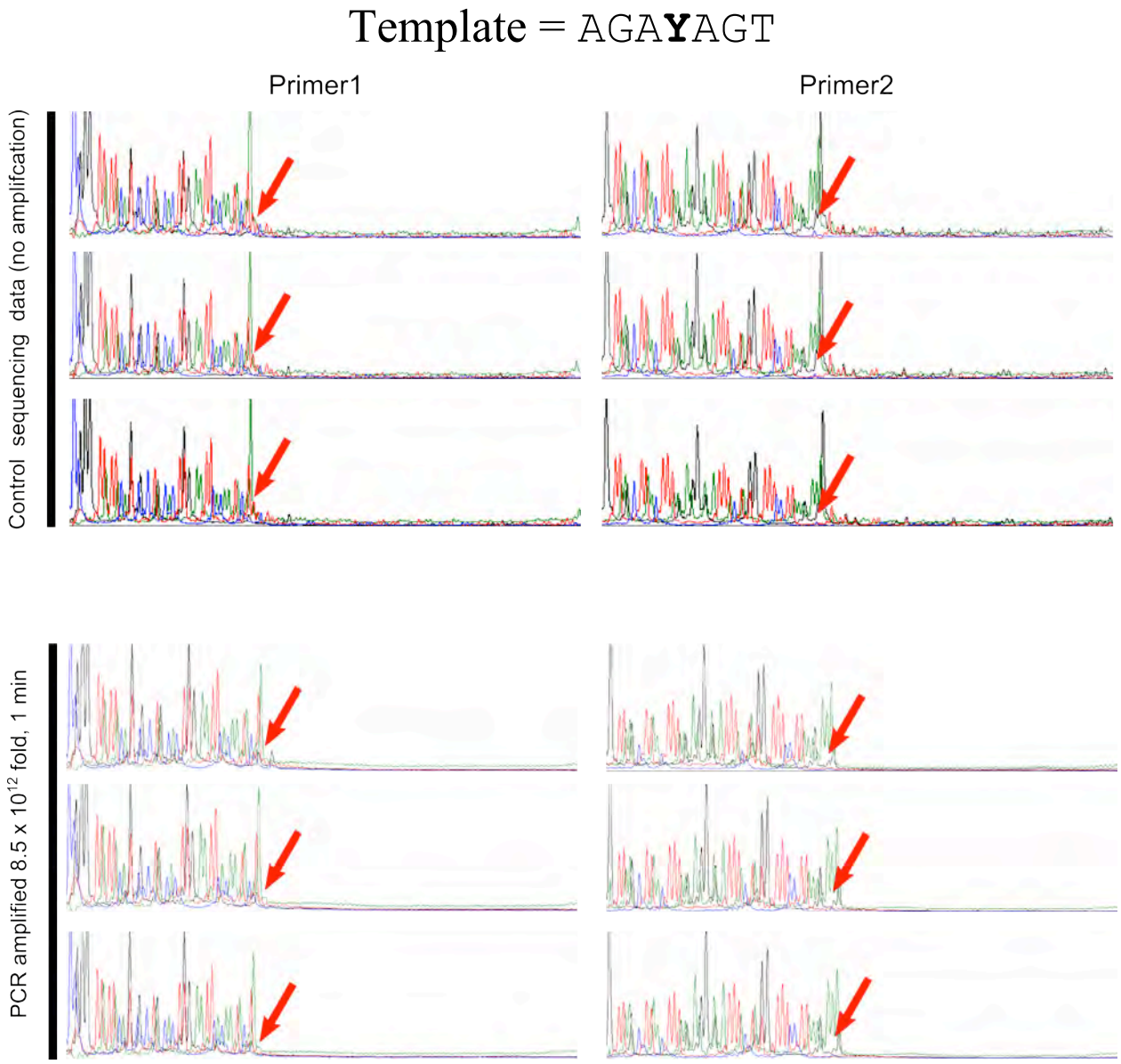

Template = AGAYAGT
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 8.5 x 10^12 fold, 1 min

Fig. S3 cont.

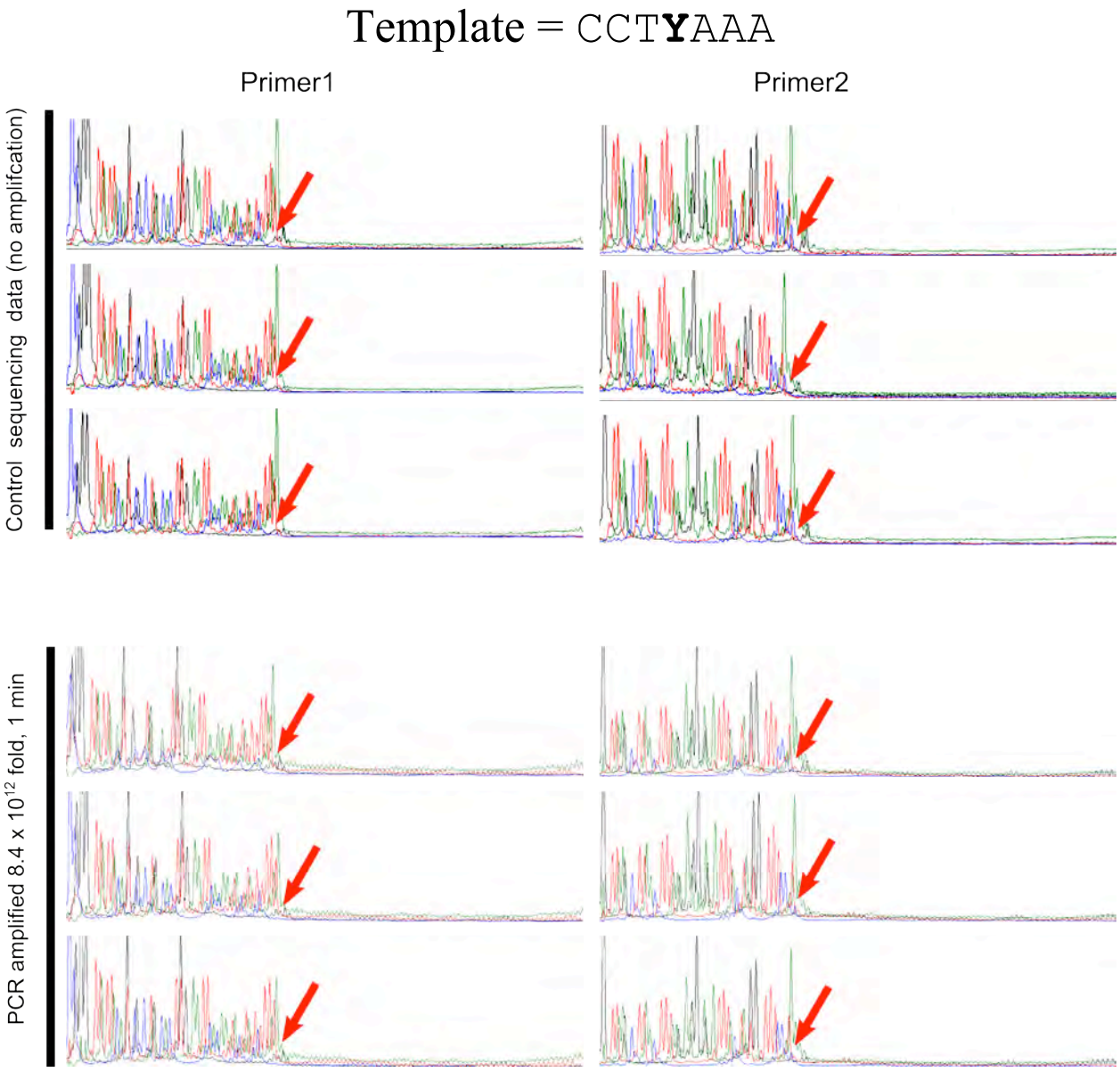

Template = CCTYAAA
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 8.4 x 10^12 fold, 1 min

Fig. S3 cont.

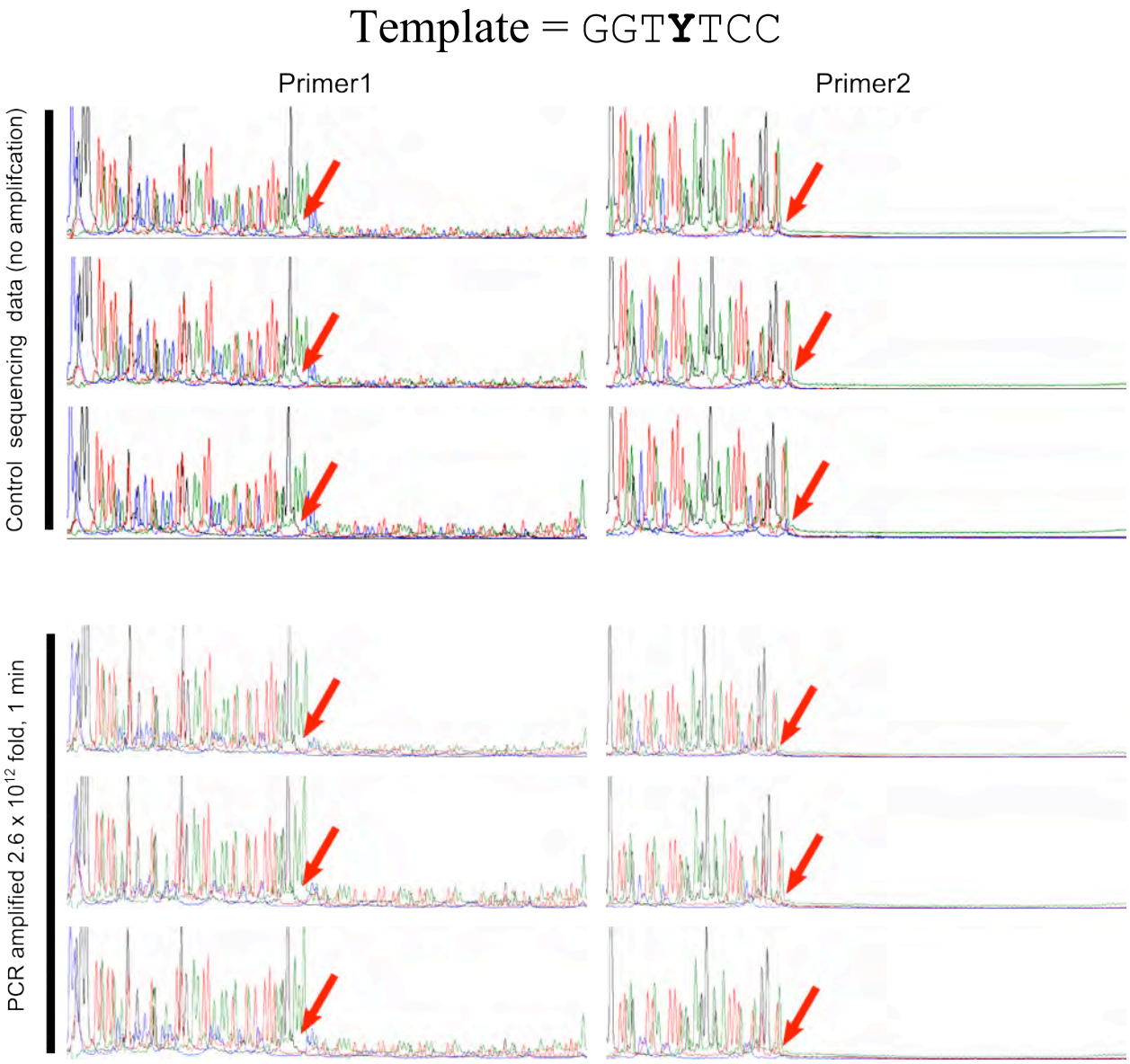

Template = GGTYTCC
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 2.6 x 10^12 fold, 1 min

Fig. S3 cont.

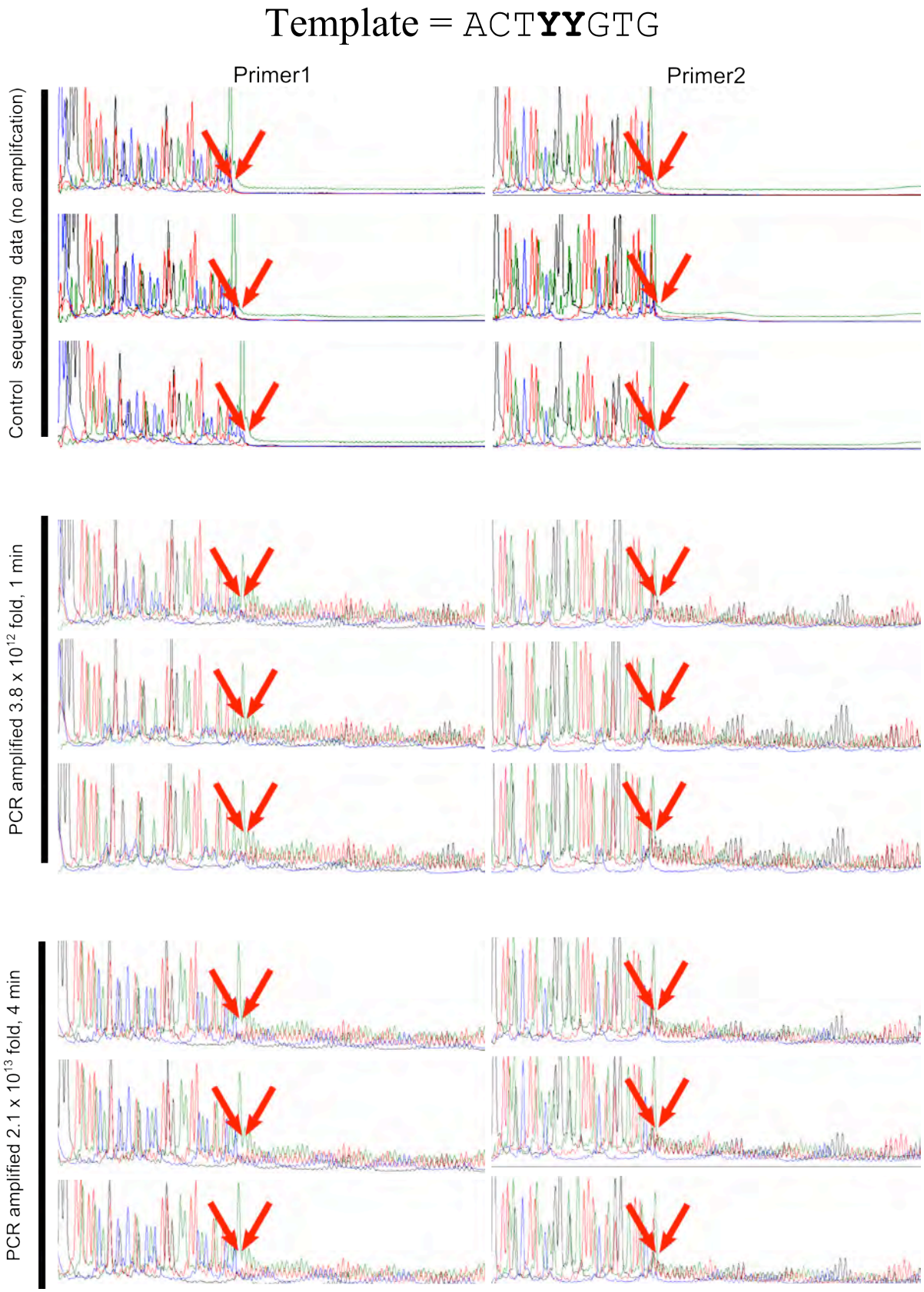

Template = ACTYYGTG
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 3.8 x 10^12 fold, 1 min
PCR amplified 2.1 x 10^13 fold, 4 min

Fig. S3 cont.

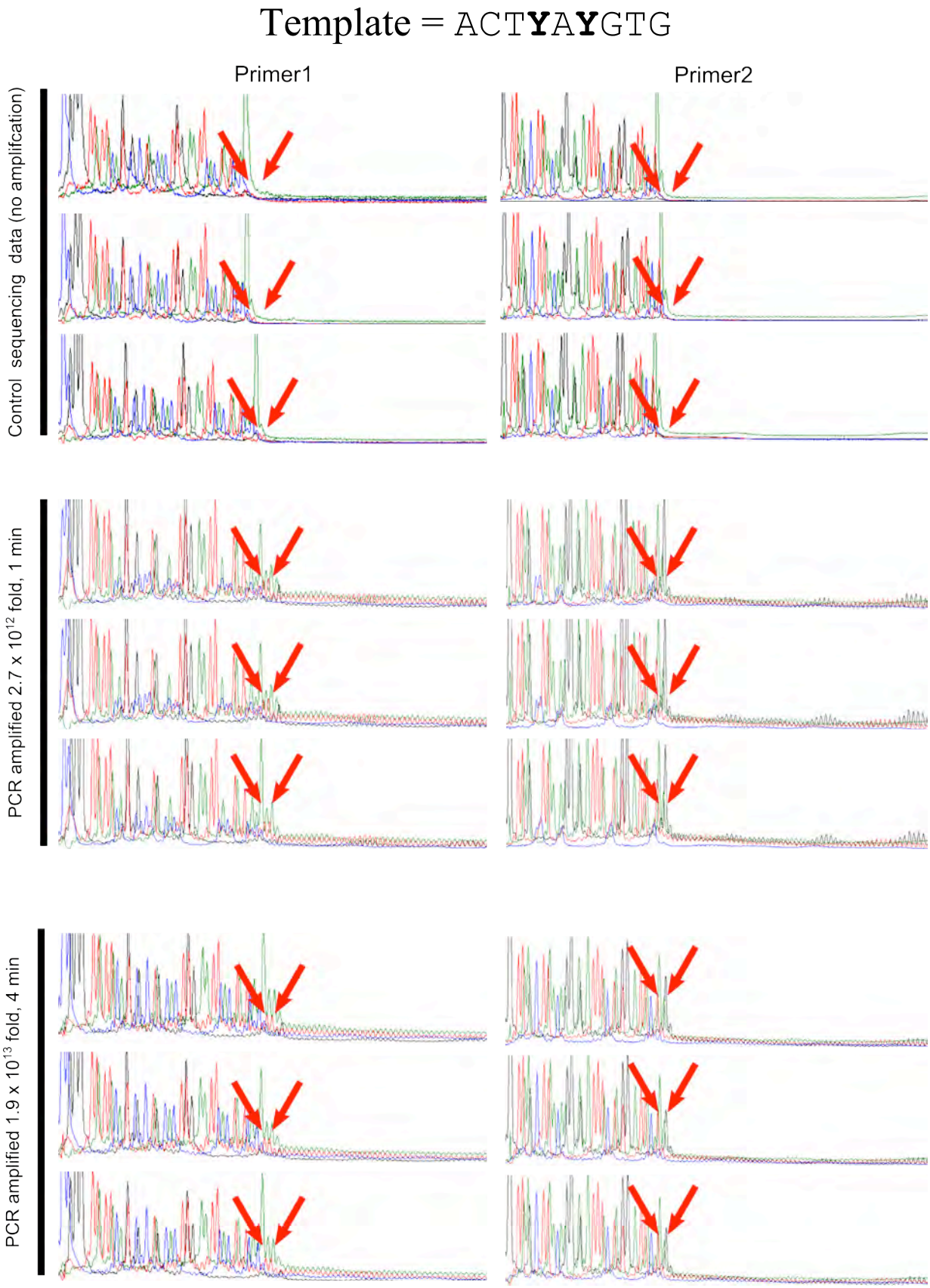

Template = ACTYAYGTG
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 2.7 x 10^12 fold, 1 min
PCR amplified 1.9 x 10^13 fold, 4 min

Fig. S3 cont.

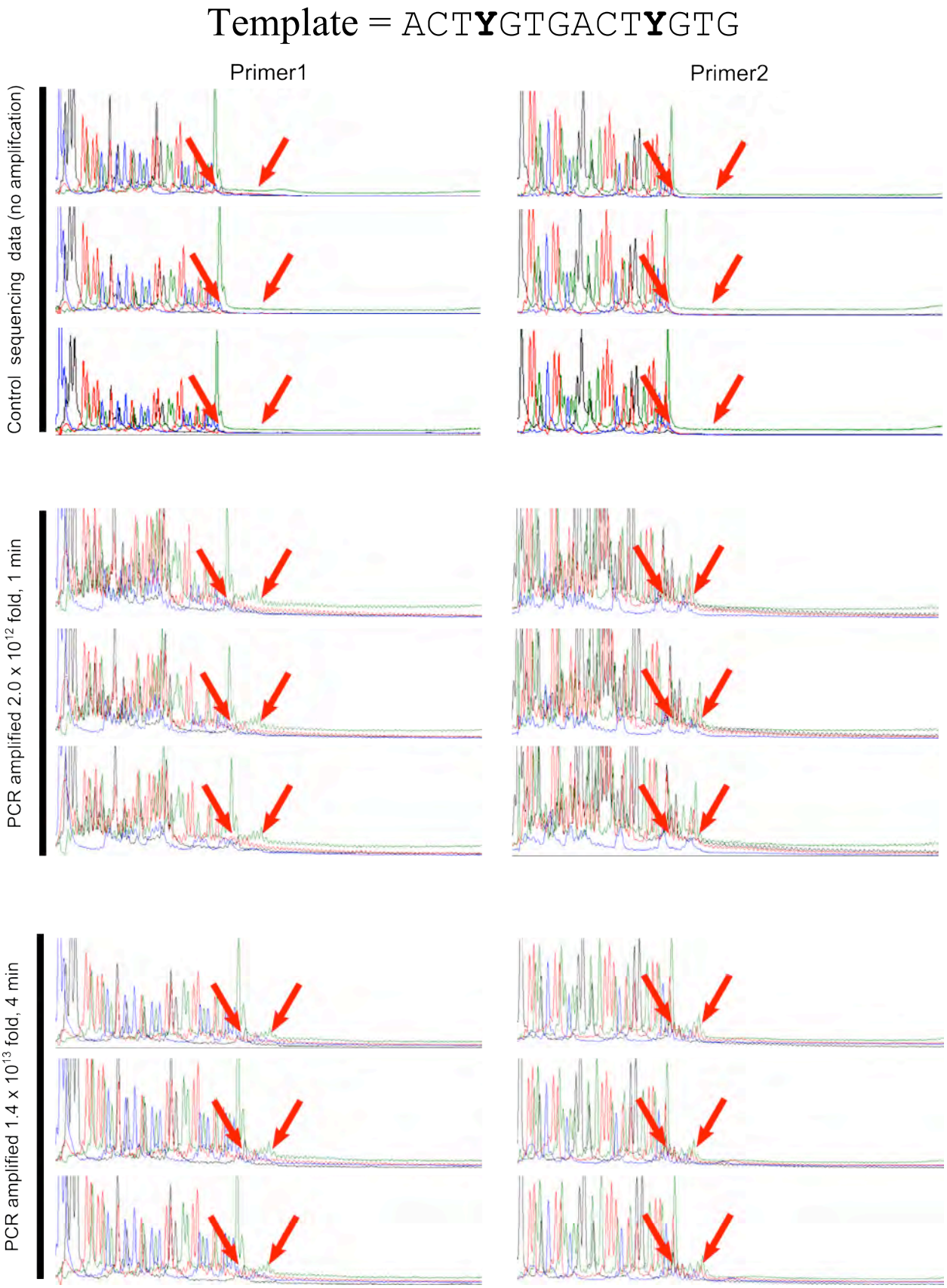

Template = ACTYGTGACTYGTG
Primer1
Primer2
Control sequencing data (no amplifcation)
PCR amplified 2.0 x 10^12 fold, 1 min
PCR amplified 1.4 x 10^13 fold, 4 min

Fig. S4. (*A*) qPCR data for PCR selection. Fold dilutions are indicated. (*B*) Gel analysis of all samples. M is a 50 bp ladder.

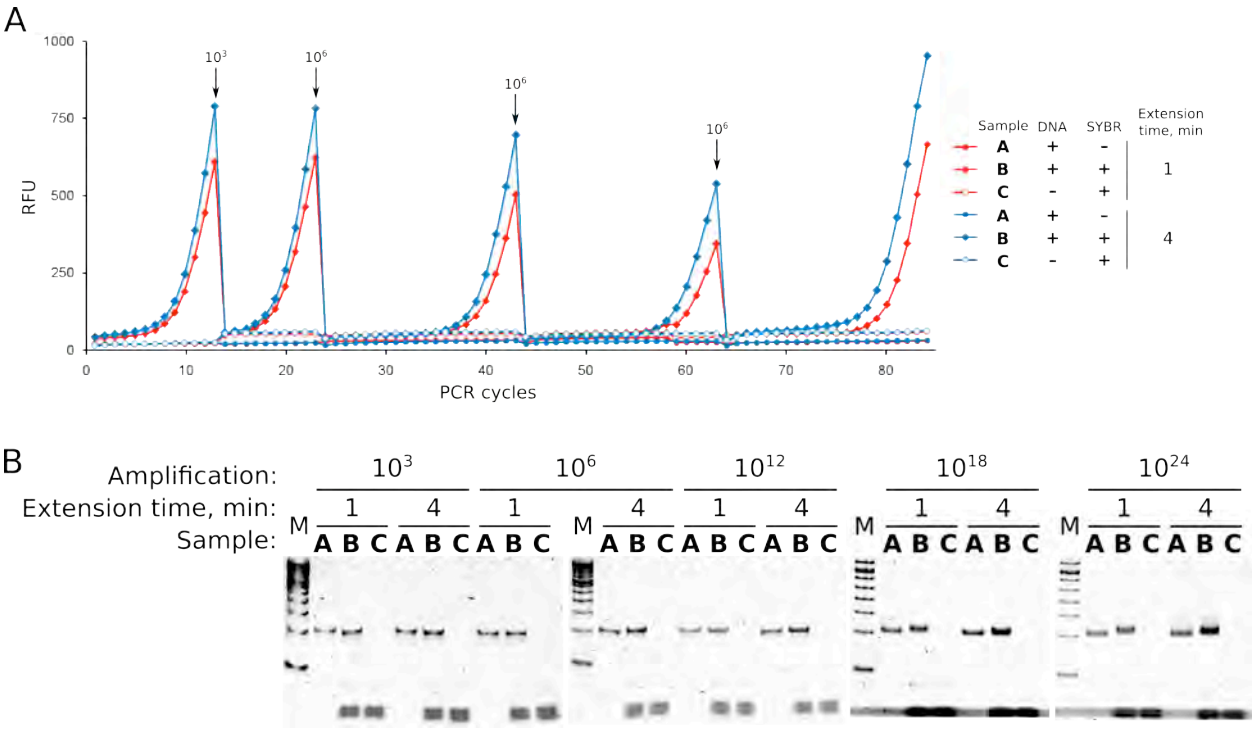
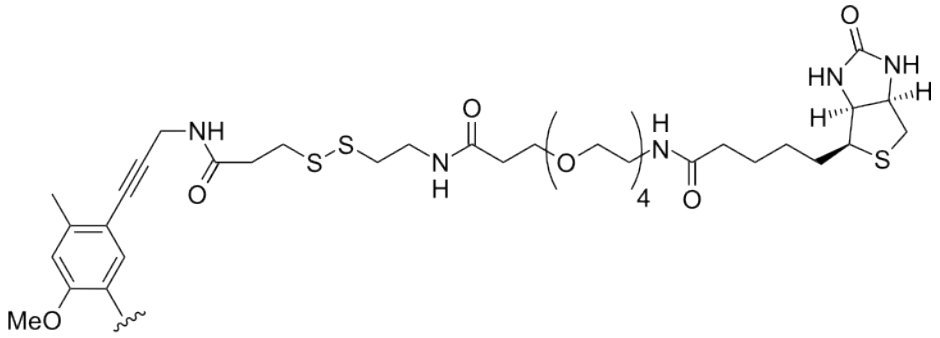


Fig. S5. Biotinylated d**MMO2** utilized to biotinylate DNA duplex at the position of the unnatural base pair (see Materials and Methods in the main text). Only nucleobase is shown, sugar and phosphate are omitted for clarity.



biotinylated d**MMO2**

Fig. S6. Analysis of amplification bias with 4 min extension time. In all cases, "retained" and "lost," refer to the populations that retained the unnatural base pair during amplification and those that lost it, respectively. (*A*) Single nucleotide bias. $f_{rel}^{N}(n)$ – 1 values are shown for each natural nucleotide (indicated along top) as a function of position relative to d**NaM** in the amplified library. Amplification level is shown along the far right edge. (*B*) Normalized pairwise correlations, C(*n*,*n'*). Only positive values of C(*n*,*n'*), which indicate amplification-dependent biases, are shown. For visualization, the discrete data are represented with continuous functions (surfaces). (*C*) 5'- and 3'-dinucleotide bias. $f_{rel}^{NN'}(n,n')$ – 1 are shown on the left and right, respectively, and are represented in a circular format, where the sequences read from the middle outward, with **X** representing d**NaM**. For example, for each dinucleotide distribution, the upper right quadrant corresponds to either 5'-NA**X** or **X**AN-3', where N is (clockwise) A, C, G, or T. Correspondingly, the bottom right quadrant corresponds to either 5'-NC**X** or **X**CN-3', the bottom left quadrant corresponds to either 5'-NG**X** or **X**GN-3', and the top left quadrant corresponds to either 5'-NT**X** or **X**TN-3'. Amplification level is indicted by shade of grey, as shown at the bottom.

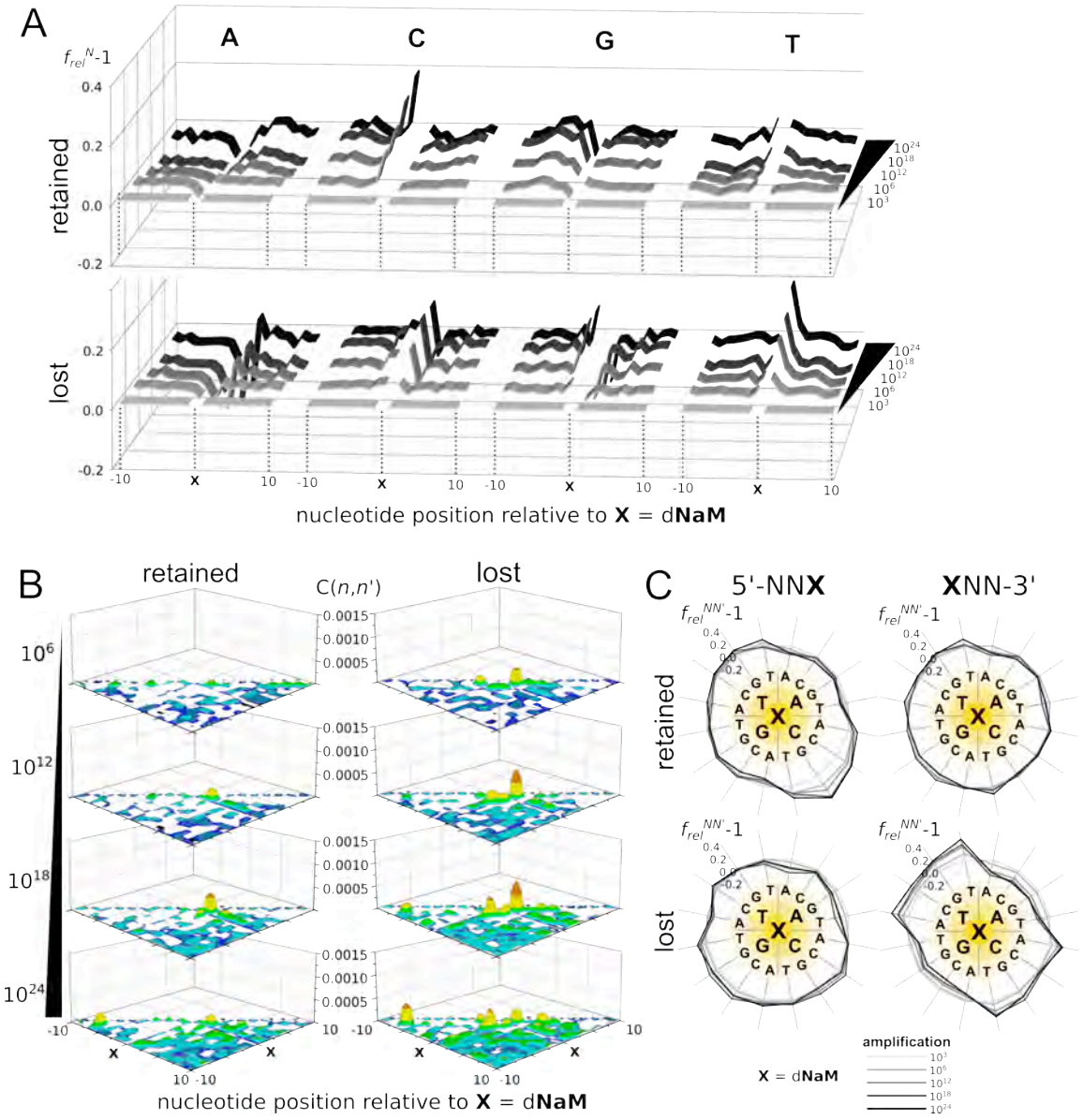


Fig. S7. (*A*) Net retention of the unnatural base pair (1–*L*) as a function of amplification during PCR selection, where *L* is the net loss of the unnatural base pair for the entire amplification. Populations with 1 min and 4 min extension time are shown in blue and red, respectively. The slope of the curve after linear regression gives ln(1–*l*), where *l* is a net loss of the unnatural base pair per doubling, and $1-L=(1-l)^{\log_2(A)}$. (*B*) ln(1–*G*) as a function of PCR amplification starting with fully natural template. *G* is the net gain of the unnatural base pair during the entire amplification. The slope of the curve after linear regression gives ln(1–*g*), where *g* is the net gain of the unnatural base pair per doubling, and $1-G=(1-g)^{\log_2(A)}$. Division of *g* by *n*, the length of the replicated DNA, provides the error rate per natural nucleotide. Gel data for panel (*A*) are shown in panel (*C*). M is a 50 bp ladder, SA is a streptavidin.

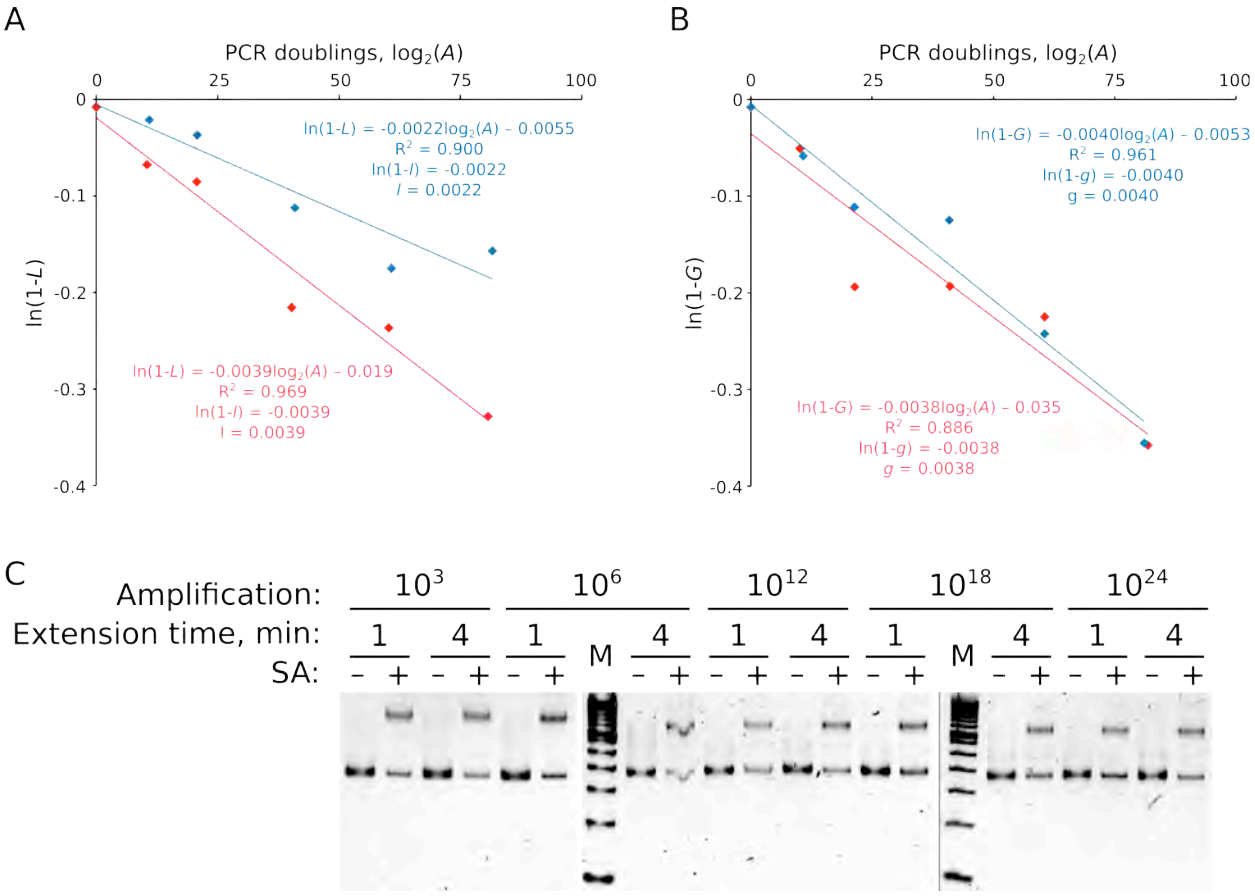


Fig. S8. (*A*) Streptavidin shift as a function of templates with the unnatural base pair (*U*). Each sample was run in triplicate and representative data are shown in panel (*B*).