Sequence determination of nucleic acids containing 5-methylisocytosine and isoguanine: identification and insight into polymerase replication of the non-natural nucleobases

J. David Ahle, Stephen Barr, A. Michael Chin[1] and Thomas R. Battersby*

Bayer HealthCare LLC, Diagnostics Division, 725 Potter Street, Berkeley, CA 94710, USA and [1]Sequetech Corporation, 935 Sierra Vista Avenue, Mountain View, CA 94043, USA

ABSTRACT

Nucleobase analogs 5-methylisocytosine (^{Me}isoC) and isoguanine (isoG) form a non-natural base pair in duplex nucleic acids with base pairing specificity orthogonal to the natural nucleobase pairs. Sequencing reactions were conducted with oligodeoxyribonucleotides (ODNs) containing d^{Me}isoC and disoG using modified pyrosequencing and dye terminator methods. Modified dye terminator sequencing was generally useful for the sequence identification of ODNs containing the non-natural nucleobases. The two sequencing methods were also used to monitor nucleotide incorporation and subsequent extension by Family A polymerases used in the sequencing methods with a six-nucleobase system that includes d^{Me}isoC and disoG. Nucleic acids containing the six-nucleobase system could be replicated well, but not as well as natural nucleic acids, especially in regions of high d^{Me}isoC–disoG content. Challenges in replication with d^{Me}isoC–disoG are consistent with nucleobase tautomerism in the insertion step and disrupted minor groove nucleobase pair–polymerase contacts in subsequent extension.

INTRODUCTION

Non-natural nucleobase analogs with base pairing specificity orthogonal to the natural base pairs have been designed to expand the sequence and functional diversity of nucleic acids (1–3). One strategy in the design of additional base pairs has been to work within the Watson–Crick pairing rules of size and hydrogen bonding complementarity. In this approach, nucleobase analogs with carbon/nitrogen ring systems isosteric to natural purines or pyrimidines are used to implement hydrogen bonding functionality arrayed in patterns not found in natural DNA (4). The most thoroughly studied of these non-natural pairs is the 5-methylisocytosine–isoguanine (^{Me}isoC–isoG) pair joined by three hydrogen bonds in duplex nucleic acids (Figure 1) (5–7), and capable of acting as a third base pair in PCR amplification (8). The ^{Me}isoC–isoG pair has established technological value in reducing background signal (9) in widely used commercial diagnostic nucleic acid hybridization assays (10,11) approved by the U.S. Food and Drug Administration and other global regulatory authorities. The pair has been used as a component of a real-time quantitative PCR assay (12). Non-natural isoC–isoG or ^{Me}isoC–isoG pairs have also been used as mechanistic probes of the fundamental biological processes of template-directed nucleic acid synthesis (13,14), translation (15), protein-mediated strand exchange of DNA (16) and excision repair (17).
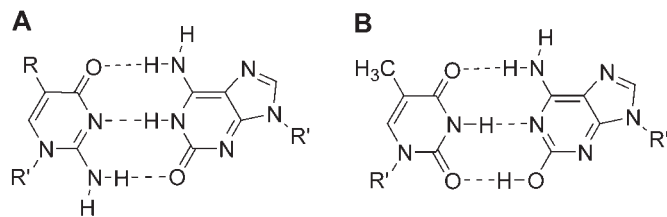
If nucleic acids containing ^{Me}isoC and isoG are to have the utility of natural nucleic acids, the tools and techniques of molecular biology must be available. A powerful tool for characterizing nucleic acids is sequence determination. Non-natural nucleobase positions in nucleic acids have been identified in very limited experiments using various methods, including enzymatic pausing (18), chemical degradation (8,13,14,19) and dye-labeled terminators (20). No reported sequencing method has concurrently identified both nucleobases of a non-natural pair. Here, we describe work to sequence oligodeoxyribonucleotides (ODNs) containing d^{Me}isoC and disoG. We demonstrate that d^{Me}isoC and disoG positions can be unambiguously identified within a single nucleic acid using a dye-labeled terminator method, despite lacking terminators corresponding to the non-natural nucleobases. We have

---

Figure 1. Base pairing with the isoC–isoG non-natural pair. (**A**) IsoC and isoG form a three hydrogen bond pair. (**B**) An isoG tautomer is complementary to T in Watson–Crick pairing. 2′-Deoxy-5-methylisocytidine and 2′-deoxyisoguanosine were used in this work (R = CH₃, R′ = 2′-deoxyribose).

repeatedly used this method to verify synthetic ODN sequences. Another method using pyrosequencing, which detects pyrophosphate generated from the enzymatic addition of a nucleoside triphosphate to a nucleic acid strand, is only partially successful at sequencing ODNs containing d^Me^isoC and disoG. Development of these sequencing systems with non-natural analogs has afforded the additional benefit of probing polymerase molecular recognition. The two sequencing methods were used to monitor nucleotide incorporation and subsequent extension by polymerases with a six-nucleobase system that includes d^Me^isoC and disoG.

MATERIALS AND METHODS

Oligodeoxyribonucleotides

Synthetic ODN sequences containing disoG and d^Me^isoC were synthesized using phosphoramidite chemistry (5,21) and PAGE purified. ODN purity was verified as at least 90%, and nearly always >95%, by capillary electrophoresis analysis (22). The identity of each ODN was confirmed by matrix-assisted laser desorption ionization time-of-flight mass spectrometry and high-performance liquid chromatography (HPLC) analysis of component nucleosides after enzymatic degradation (22).

Pyrosequencing

Pyrosequencing was performed using a PSQ96MA Sequencer (Biotage AB). For optimal results, the concentration of a nucleotide in pyrosequencing reactions should be slightly above the K_m of the enzyme for that particular nucleotide. Lower concentrations cause incomplete incorporation and higher concentrations increase misincorporation. Stock solutions of nucleotides are dispensed stepwise from four reservoirs in a pyrosequencing dispensation cartridge. Concentrations of stock solutions of d^Me^isoCTP and disoGTP required to give nucleotides at the appropriate K_m during extension were roughly determined in separate experiments (data not shown) by dispensing a range (10 μM, 50 μM, 500 μM and 2.5 mM) of nucleotide concentrations in pyrosequencing reactions with templates containing either disoG (for d^Me^isoCTP) or d^Me^isoC (for disoGTP). The lower concentrations were clearly insufficient and generated less pyrophosphate than expected for complete incorporation. Signal height at 500 μM and 2.5 mM was nearly unchanged for both non-natural nucleotides, suggesting pyrosequencing reservoir stock solutions at 500 μM dispense nucleotide near K_m in the reaction solution for the insertion of d^Me^isoCTP opposite

template disoG and disoGTP opposite template d^Me^isoC. The higher 2.5 mM concentration was chosen for these nucleotides in the cartridge reservoirs to guard against incomplete incorporation at the expense of possibly slightly increasing misincorporation of d^Me^isoC and disoG. For comparison, standard cartridge concentrations used in pyrosequencing were measured by A_{260} at 0.3–0.6 mM for dCTP, dGTP and dTTP, and ∼2.5 mM for α-S-dATP.

Most of the pyrophosphate impurities visible in pyrosequencing were removed from d^Me^isoCTP and disoGTP by HPLC purification with a YMC-Pack ODS-AM column (120 Å, 5 μm, 250 × 4.6 mm). Approximately 0.2 μmol of nucleotide (20 μl) was injected on a Series 1100 HPLC (Hewlett Packard) and purified with a binary gradient (solvent A = 0.2 M triethylammonium acetate, pH 6.8; solvent B = 95% solvent A, 5% acetonitrile) at 1.0 ml/min: 4% solvent B hold for 10 min, then increase solvent B to 100% over 25 min. The eluate containing the d^Me^isoCTP or disoGTP was collected by monitoring at 260 nm. Eluate containing d^Me^isoCTP was immediately adjusted to pH 8.3 with triethylamine. The solvent and the volatile buffer were removed under vacuum and the nucleotide was redissolved in water. The purification yielded nucleotides with no detectable impurity peaks upon reinjection and analysis with this HPLC method, and removed most of the pyrophosphate undetectable by UV monitoring, but visible in pyrosequencing.

Fifteen ODN templates were designed to vary the nearest-neighbor positions around d^Me^isoC and disoG positions. Only three of the four natural nucleobases were used in each template to leave an available reservoir for either d^Me^isoCTP or disoGTP. Complementary nucleotides were sequentially dispensed for each template. In instances of incomplete incorporation, sequential dispensations of a single nucleotide were used in a subsequent experiment to examine whether extending the time available for incorporation would increase incorporation. One dispensation of an out-of-sequence non-complementary nucleotide was performed with each ODN as a negative control (Supplementary Figure S3). Pyrosequencing data presented are peak heights of emitted light detected and are the average of two replicates.

Preparation of ODNs for cycle sequencing

Synthetic ODNs were ligated with T4 DNA ligase (Amersham Biosciences) to a DNA fragment, which was assembled from three component ODNs (170 nt total, sequences in Supplementary Material). The ODN to be sequenced (1.73 nmol) was ligated to a 5′-phosphorylated 50mer ODN (1.44 nmol) using a reverse complementary linker ODN (2.02 nmol) that formed a 6 nt duplex with each of the ODNs to be ligated. Simultaneously, the 50mer was ligated to a 5′-phosphorylated 57mer (1.20 nmol) through an analogous linker ODN (1.68 nmol), and the 57mer was in turn ligated to a 5′-phosphorylated 63mer (1.00 nmol) through another linker ODN (1.40 nmol). An annealing step was first performed in 1× 100 μl ligation buffer (50 mM Tris–HCl, pH 7.5, 10 mM MgCl₂, 2 mM spermidine) by incubating the solution at 55°C (2 min) and reducing the temperature (0.67°C/min) to 22°C, then holding (2 min). To the ODN solution was added 10× ligation buffer (10 μl), 100 mM ATP (4 μl), 500 mM DTT (4 μl), water (28 μl), 50% PEG-8000 (48 μl) and T4 DNA ligase (6 μl, 6 U).

The ligation reaction was incubated at 20°C for 14 h. The reaction was quenched with 0.5 M EDTA (6 μl) and the nucleic acid was precipitated by adding pH 4.8 ammonium acetate (137 μl) and ethanol (687 μl) and cooling at −20°C for 1 h. After spinning in a microcentrifuge at 4°C (20 000 r.c.f., 30 min) and rinsing three times with cold 80% ethanol, the pellet was dissolved and the ligation product separated on a 5% polyacrylamide gel. The product band was excised and isolated by electroelution. The ligation product was purified using a NAP-25 column and then ethanol precipitated again.

Cycle sequencing

The template produced by ligation was included in cycle sequencing reactions with d^{Me}isoCTP (21), disoGTP (21), and either BigDye 3.0 (Applied Biosystems) or BigDye 3.1 (Applied Biosystems) kits. Cycle sequencing was performed on a 7700 Sequence Detector (Applied Biosystems) in 9600 emulation mode with 17.6 μl of Ready Reaction Mix using 25 nM template and 160 nM primer in 44 μl reactions for 25 cycles (96°C, 10 s; 60°C, 240 s). The reactions were then purified with DTR spin columns (Edge Biosystems). The sequencing reactions were analyzed on a 310 Genetic Analyzer (Applied Biosystems) using POP-6 polymer gel in a 61 cm × 50 μm uncoated capillary (50°C, 200 V/cm). Concentrations of d^{Me}isoCTP and disoGTP (10–1000 μM) were examined in optimization matrix experiments with the goals of minimal signal attenuation and no mispaired terminator signals opposite d^{Me}isoC or disoG template positions.

RESULTS

Pyrosequencing

In pyrosequencing (23,24), nucleoside triphosphates are singly dispensed into a solution containing primer, template and exo(−) Klenow fragment of DNA polymerase I at 28°C. Incorporation of a complementary nucleotide produces an enzymatically mediated cascade resulting in the generation of visible light. The amount of light generated is proportional to the pyrophosphate produced during incorporation. Excess nucleotide is enzymatically destroyed before subsequent nucleotide dispensations. The ODN templates used here all have natural nucleobases at the first four template positions and these positions always yielded relative signal heights typical of pyrosequencing with natural nucleobases. In evaluating these signals, it is important to note that dispensations of α-S-dATP used in pyrosequencing typically result in peak heights ∼20% higher than the other nucleotides (25). These first four positions form a baseline for comparison of replication performance with non-natural nucleobases. Because of the template design, non-natural positions were challenged with natural nucleotides upon dispensation of the complementary nucleotide opposite the fourth template position of the ODNs containing d^{Me}isoC or disoG positions. Significant incorporation of a natural nucleotide opposite the d^{Me}isoC or disoG at the fifth template position would add to the signal for incorporation opposite the fourth position and give a signal greater than one equivalent.

Pyrosequencing reactions performed quite differently depending on whether disoG or d^{Me}isoC was present in the 9 nt template region o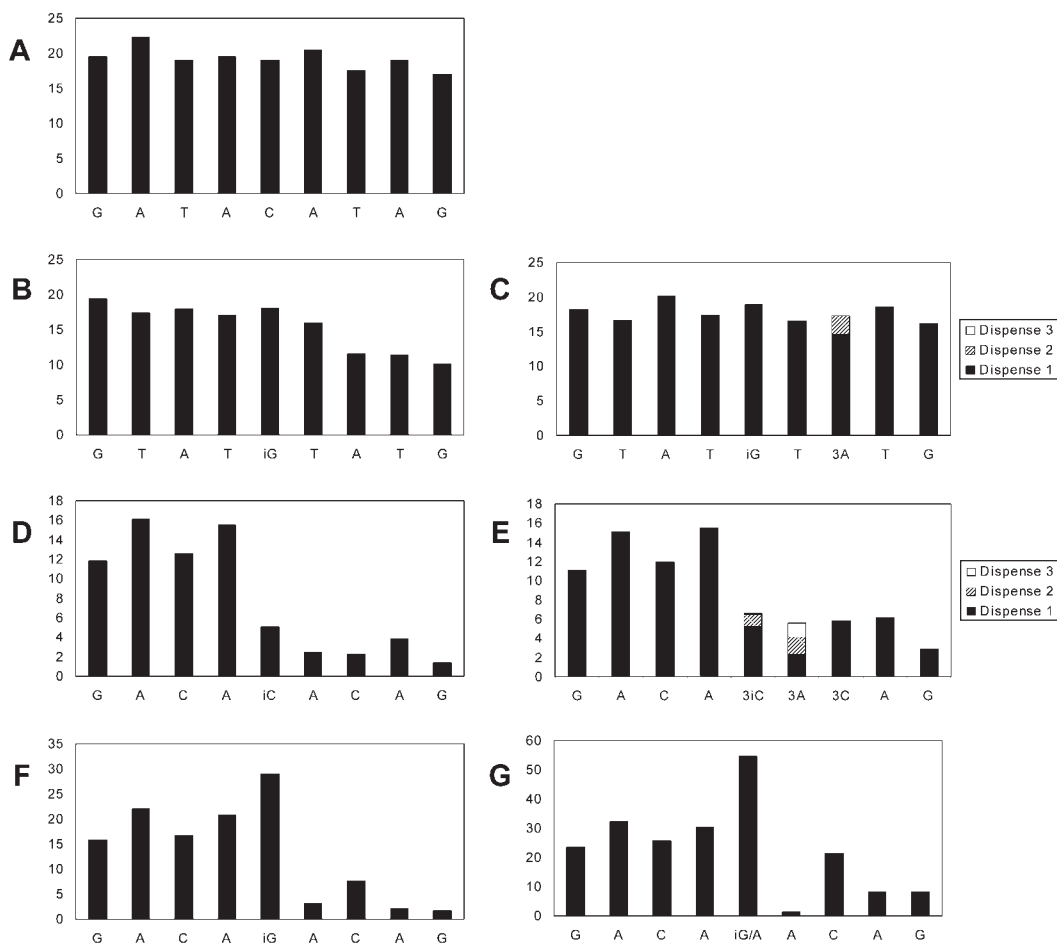f the individual ODNs (Figure 2 and Supplementary Figures S1 and S2). Complementary disoG nucleotide was always readily incorporated opposite d^{Me}isoC in the template, while the natural nucleotides were not significantly incorporated opposite d^{Me}isoC (Figure 2B). However, further extension following the d^{Me}isoC–disoG pair was slowed and was often incomplete after a single dispensation of nucleotide. Sequential dispensations of the same complementary nucleotide allowed more cumulative time for incorporation and usually improved the incomplete incorporation observed with a single dispensation at positions following a disoG–d^{Me}isoC pair (Figure 2C). Misincorporation upon dispensing disoGTP was observed when disoGTP was dispensed at d^{Me}isoC positions followed by dT (Figure 2F); more than one equivalent of pyrophosphate was produced, indicating that disoG nucleotide was incorporated opposite d^{Me}isoC and then further incorporated opposite the following dT. When α-S-dATP was mixed 1:1 with disoGTP and dispensed at a template d^{Me}isoC position followed by dT, two equivalents of pyrophosphate were produced and incorporation opposite the position following dT was improved (Figure 2G), suggesting that disoG and dA were incorporated opposite disoC and dT, respectively. This implies that dA is incorporated more readily opposite template dT positions than disoG.

In contrast, significantly less than one equivalent of pyrophosphate was produced when d^{Me}isoCTP was dispensed at disoG template positions (Figure 2D). The proportion of template disoG paired with d^{Me}isoC was quite variable in different sequence contexts (Supplementary Figure S1). Extending the time available for incorporation through multiple dispensations of d^{Me}isoCTP increased the incorporation of d^{Me}isoC only slightly (Figure 2E). Further extension of the fraction of templates incorporating d^{Me}isoC opposite template disoG was slow and was improved by multiple nucleotide dispensations at positions following d^{Me}isoC–disoG (Figure 2E). Significant misincorporation of any natural nucleotide opposite template disoG was not observed, and d^{Me}isoC nucleotide was not visibly incorporated opposite any of the natural nucleobases.

Natural nucleotides were not significantly incorporated opposite d^{Me}isoC or disoG positions of any template. No misincorporation following the fourth template position is visible in any of the pyrosequencing reactions, which cover all possible natural nucleotide misincorporations opposite non-natural template positions. Interestingly, even dT was not misincorporated opposite template disoG.

Dye terminator sequencing

Sequencing reactions with a thermophilic polymerase and dye terminator chemistry were also conducted with ODNs containing disoG and d^{Me}isoC. Two sequencing kits, BigDye 3.0 and BigDye 3.1 (Applied Biosystems), were used. Amplitaq FS in BigDye 3.0 is a *Taq* polymerase with two point mutations: the F766Y mutation increases the acceptance of dideoxy nucleotides and G46D eliminates the 5′-exonuclease activity (26). The identity of the polymerase from the BigDye 3.1 kit is undisclosed, but the kit almost certainly includes a Family A polymerase with two analogous mutations (26). The polymerases from the two kits had qualitatively similar performance in sequencing reactions with d^{Me}isoC and disoG.
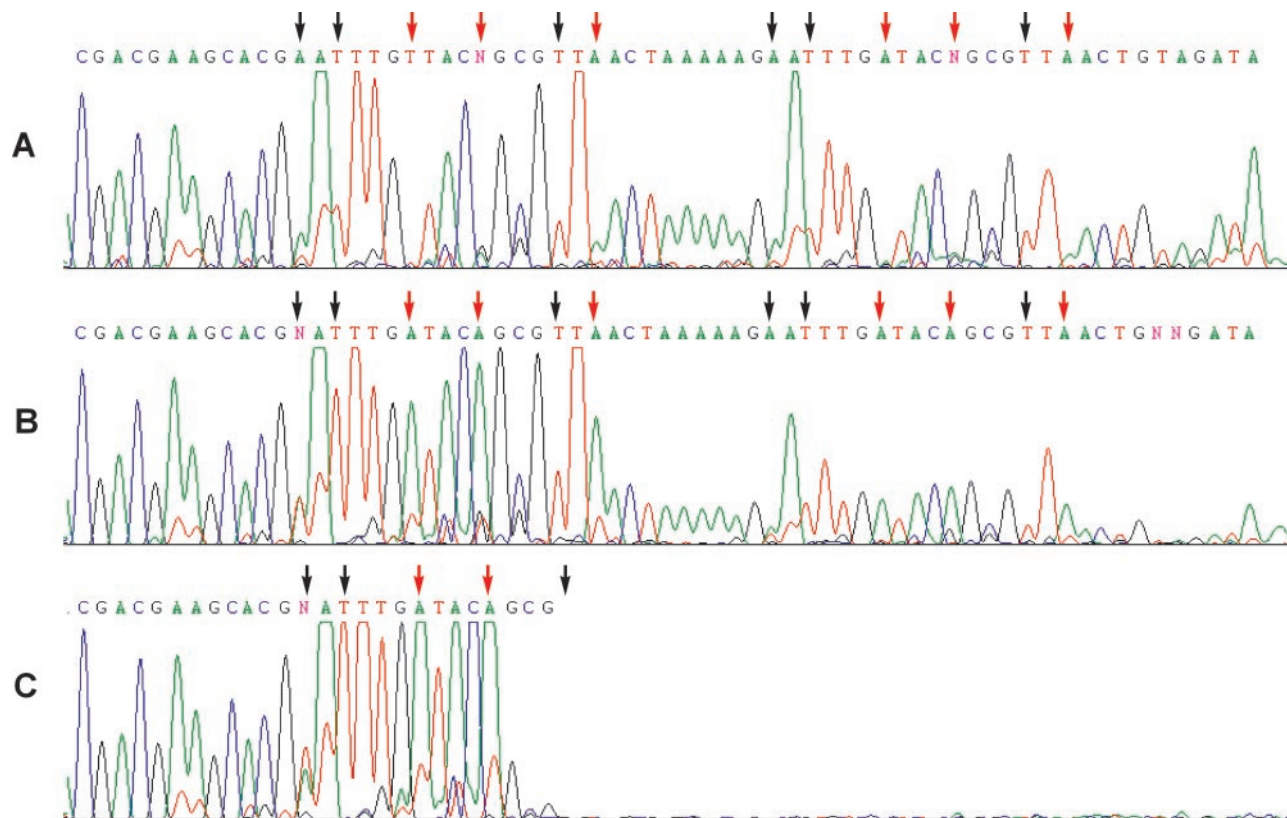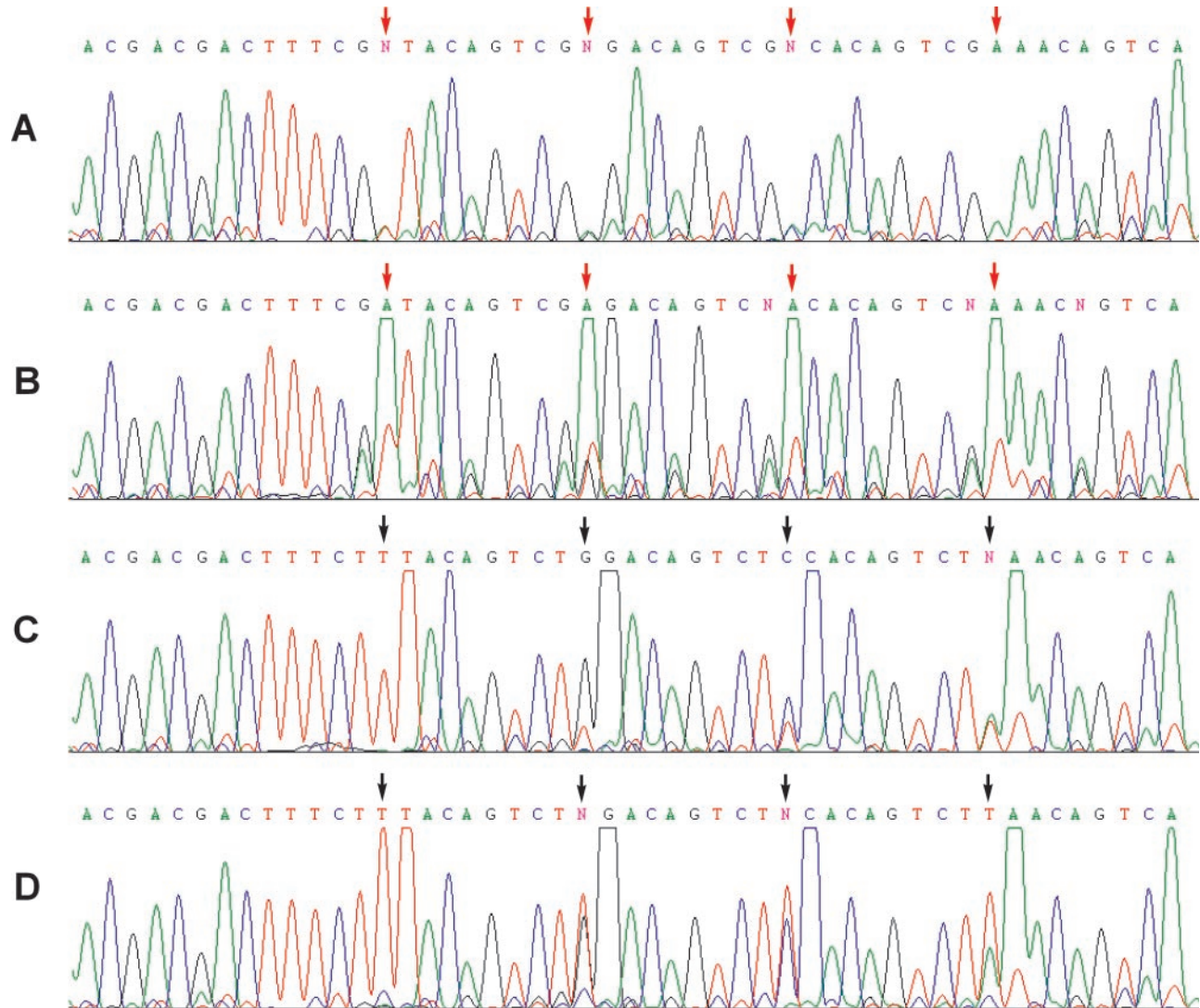
Figure 3. Dye terminator sequencing (BigDye 3.1 kit) of an ODN containing d^{Me}isoC and disoG used to optimize concentrations of d^{Me}isoCTP and disoGTP. The template sequence in the region shown is 3′-GCTGCTTCGTGCiGTiGAACiCATGiCCGCiGAiC TGATTTTTCiGTiGAACiCATGiCCGCiGAiCTGACATCTA-5′. Changing the concentrations of d^{Me}isoCTP and disoGTP caused differences in dye terminator signals at template disoG (black arrow) and d^{Me}isoC (red arrow) positions. With increasing concentrations of d^{Me}isoCTP, signals from ddT were suppressed opposite template disoG positions. With increasing concentrations of disoGTP, signals from ddA were suppressed opposite d^{Me}isoC template positions. These changes are indicative of competition for incorporation at non-natural template positions between a non-natural nucleotide and a specific natural nucleotide. (**A**) Sequencing with 100 μM d^{Me}isoCTP and 400 μM disoGTP largely suppressed terminator incorporation opposite the non-natural positions. Modest signal attenuation caused by unwanted strand termination was apparent. (**B**) As the concentrations of disoGTP and d^{Me}isoCTP were decreased (10 μM d^{Me}isoCTP and 10 μM disoGTP), ddA terminators were incorporated opposite template d^{Me}isoC positions and ddT terminators were incorporated opposite template disoG positions. Signal attenuation caused by unwanted strand termination was increased. (**C**) In the absence of disoGTP and d^{Me}isoCTP, more ddA was incorporated opposite template disoC and more ddT was incorporated opposite template d^{Me}isoG. Forced misincorporation at all non-natural template positions led to premature termination of the sequencing reaction, unlike in (A) and (B) where it was possible to incorporate some non-natural nucleotide. Similar reactions were performed using the BigDye 3.0 kit (Supplementary Figures S4 and S5).

signal attenuation was always observed upon encountering the multiple non-natural nucleobase positions. In contrast, replication of the template was completely terminated in the absence of disoGTP and d^{Me}isoCTP (Figure 3C). These changes in incorporation and extension with varying disoGTP and d^{Me}isoCTP concentrations were primarily a result of the change in nucleotide concentration, and not a generally inhibitory effect, such as an effective reduction in the free Mg^{2+} concentration. If an increase in disoGTP or d^{Me}isoCTP caused a general inhibition, then the dye signals for the >170 natural nucleobases preceding the non-natural template positions would also be attenuated. General attenuation was observed only at very high concentrations of disoGTP and d^{Me}isoCTP (data not shown).

A series of sequencing reactions was conducted to examine the influence of sequence context (Figure 4 and Supplementary Figures S6 and S7). These experiments were performed with 42mer template ODNs containing all 16 natural nucleotide nearest-neighbor contexts possible for d^{Me}isoC and analogous template ODNs for disoG. The ODNs were used as templates in

sequencing reactions in the presence or absence of disoGTP and d^{Me}isoCTP. In the absence of disoGTP and d^{Me}isoCTP, extension required misincorporation of natural nucleobases opposite the d^{Me}isoC and disoG template positions in order to proceed; ddA was always incorporated opposite template d^{Me}isoC positions (Figure 4B) and ddT was always incorporated opposite template disoG positions (Figure 4D). In the presence of disoGTP and d^{Me}isoCTP, the complementary non-natural nucleotide was paired opposite d^{Me}isoC and disoG in all sequence contexts, verified by diminished terminator signals opposite the non-natural template positions (Figure 4A and C). Additionally, a noticeable signal that may correspond to the polymerase skipping over a fraction of template disoG positions was often visible opposite disoG (Figure 4C).

Notable features of the terminator signal intensities were evident in the sequencing reactions. Interestingly, no apparent signal attenuation was visible with these ODN templates containing only isolated non-natural nucleobases, either in the presence or absence of disoGTP and d^{Me}isoCTP. Extremely

Figure 4. Dye terminator sequencing of two ODNs used to examine d^{Me}isoC and disoG in all possible natural nearest-neighbor contexts (Supplementary Figures S6 and S7 contain additional sequences). An ODN containing four d^{Me}isoC positions (red arrows) was sequenced in a BigDye 3.1 reaction in the presence of (**A**) 300 μM disoGTP and 0 μM d^{Me}isoCTP or (**B**) 0 μM disoGTP and 0 μM d^{Me}isoCTP. The template sequence in the region shown is 3′-TGCTGCTGAAAGCiCATGT-CAGCiCCTGTCAGCiCGTGTCAGCiCTTTGTCAG-5′. Very large ddA terminator signals were always observed opposite d^{Me}isoC in the absence of disoGTP, perhaps indicating a transition state base pairing geometry different from complementary pairings. An ODN containing four disoG positions (black arrows) was sequenced in a BigDye 3.1 sequencing reaction in the presence of (**C**) 100 μM d^{Me}isoCTP and 0 μM disoGTP or (**D**) 0 μM d^{Me}isoCTP and 0 μM disoGTP. The template sequence in the region shown is 3′-TGCTGCTGAAAGAiG ATGTCAGAiGCTGTCAGAiGGTGTCAGAiGTTGTCAG-5′. Terminator ddT was always incorporated opposite disoG in the absence of d^{Me}isoCTP. Extremely large terminator signals were observed at the position following disoG for all conditions. Additionally, a noticeable signal that may correspond to the polymerase skipping over a fraction of template disoG positions was visible opposite disoG. Little, if any, signal attenuation caused by unwanted strand termination was visible upon encountering isolated d^{Me}isoC or disoG positions, either in the presence or absence of complements disoGTP and d^{Me}isoCTP. This contrasts with Figure 3C, in which several proximate non-natural template positions caused premature termination of sequencing reactions lacking disoGTP and d^{Me}isoCTP.

large terminator signals were always observed at the position following incorporation of either d^{Me}isoC or dT nucleotides opposite template disoG positions, indicating the partitioning of dideoxy and deoxy nucleotides opposite these positions was outside the usual range for natural nucleobase templates. In the absence of disoGTP, extremely large ddA terminator signals were observed opposite d^{Me}isoC positions. The perturbed ratio of dideoxy terminator to deoxynucleotide for incorporation of A nucleotides opposite d^{Me}isoC suggests that transition state base pairing geometry of this pair may be different from complementary pairs.

Sequencing reactions were also conducted to verify the specificity of incorporation of d^{Me}isoC and disoG nucleotides (Supplementary Figure S8). Addition of d^{Me}isoCTP and disoGTP to sequencing reactions with templates lacking d^{Me}isoC and disoG caused no discernable differences in dye terminator patterns from standard sequencing reactions. Addition of d^{Me}isoCTP had no effect on sequencing reactions of templates containing d^{Me}isoC, even in the absence of diso-GTP. Similarly, addition of disoGTP had no effect on sequencing reactions of templates containing disoG in the absence of d^{Me}isoCTP. These reactions demonstrate that d^{Me}isoC and

disoG were not significantly incorporated opposite natural nucleobases and were not self-paired.

DISCUSSION

We have demonstrated the first generally useful method to determine sequences of nucleic acids containing both constituents of a non-natural nucleobase pair. Our dye terminator method has been routinely used in a single reaction with d[Me]isoCTP and disoGTP to verify the known sequences of diverse synthetic ODNs containing d[Me]isoC and disoG. Additionally, the method may have future application in determining unknown sequences of nucleic acids, such as ODNs generated from *in vitro* selection (27) experiments using a six-nucleobase lexicon with d[Me]isoC and disoG. More than one reaction is necessary to unambiguously identify d[Me]isoC and disoG positions in nucleic acids of unknown sequence containing both nucleobases. In a first reaction, d[Me]isoCTP and isoGTP are present at concentrations sufficient to suppress misincorporation at d[Me]isoC and disoG positions. In subsequent sequencing reactions, the concentration of d[Me]isoCTP or disoGTP is reduced (both nucleotide concentrations may also be reduced simultaneously in a single reaction), permitting ddA and ddT nucleotides to be incorporated opposite some of the d[Me]isoC and disoG positions, respectively. Suppression of specific terminator signals at increased d[Me]isoCTP or disoGTP concentrations, in addition to the signature large terminator signal following disoG template positions, should allow the identification of the non-natural positions.

The sequencing experiments also demonstrate how replication with the six-nucleobase lexicon falls short of the performance of the natural nucleobases, providing an opportunity to probe features of nucleobases important for polymerase recognition. The pyrosequencing method with d[Me]isoC and disoG suffered from three defects. First, extension in the positions following d[Me]isoC–disoG pairs was significantly slowed. Second, d[Me]isoC nucleotide was not readily incorporated opposite disoG template nucleobases. Third, disoG nucleotide was incorporated opposite template dT positions more readily than natural nucleotides are misincorporated opposite natural nucleobases. The more successful dye terminator sequencing method also had some difficulty with the non-natural nucleobases. Extension at several positions following a d[Me]isoC–disoG pair was clearly inhibited, leading to modest signal attenuation upon encountering additional proximate disoG and d[Me]isoC positions.

Tautomerism

Some of the peculiarities in the replication of the d[Me]isoC–disoG pair may be a consequence of tautomerism. A 2O–H tautomer of isoG (Figure 1B), complementary to T, has long been suspected of confounding replication of isoC–isoG (13,14,28–30). One problem that may result from isoG tautomerism is the difficulty of incorporating d[Me]isoC nucleotide opposite template disoG positions in the pyrosequencing reactions. Two observations suggest that the deficient incorporation of d[Me]isoC is indeed the result of interaction between paired nucleobases and not a protein–nucleobase interaction at insertion. First, crystal structures of complexes of Family A

polymerases, DNA duplex and dNTP lack direct contacts with nucleobases at the insertion site (31–33). Second, 3-deazaadenine (34) and nonpolar nucleobase analogs, unable to form minor groove hydrogen bonding contacts (35,36), are nonetheless efficiently incorporated opposite template dT by diverse polymerases. Tautomerism of the template isoG is implicated because replication should only be affected by tautomerism at the template nucleobase; an unsuitable tautomer as triphosphate would simply be selectively excluded. It is possible that isoG may not readily interconvert between tautomeric forms at the polymerase active site in the lower temperature pyrosequencing method, leading to problematic d[Me]isoC incorporation when isoG is locked in an alternate tautomeric form. The evident incorporation of d[Me]isoC opposite template disoG positions by the thermophilic polymerases may be the result of relatively more rapid interconversion of tautomers in the polymerase active site or a shifted tautomeric equilibrium [although the tautomeric equilibrium of disoG has been reported as unperturbed by variations in this temperature range (30)]. Curiously, if the 2O–H tautomer of isoG was present in the templates, it did not lead to significant dT incorporation opposite disoG in pyrosequencing (Supplementary Figure S1C–F). Pyrosequencing and dye terminator sequencing suggest that the incorporation of dT nucleotide opposite template disoG by Family A polymerases, while apparently facile for a misincorporation event (14,30,37,38), is probably much slower than the incorporation of d[Me]isoC nucleotide opposite disoG.

In our experiments, the misincorporation of disoG nucleotide opposite template dT occurred much more readily than the misincorporation of dT nucleotide opposite template disoG. The 2O–H tautomer of isoG has also been invoked to explain previously observed incorporation of disoG nucleotide opposite dT template positions (13,14,39,40). Misincorporation of disoG nucleotide opposite dT positions adjacent to template d[Me]isoC was also apparent in our pyrosequencing reactions. However, this misincorporation was evidently suppressed in the presence of competing dA nucleotide in the pyrosequencing and dye terminator reactions, suggesting a preference for incorporation of disoG over dA opposite template d[Me]isoC positions (39). The comparative ease of this misincorporation, however, may still lead to relatively high mutation rates in the six-nucleobase system.

Extension following d[Me]isoC–disoG positions

Another irregularity in d[Me]isoC–disoG replication is the relatively poor extension following d[Me]isoC–disoG pairs, reminiscent of slow extension following natural nucleobase mismatches (41). This is the first report of hindered extension following correctly matched d[Me]isoC–disoG pairs. Hindered extension in dye terminator sequencing of templates with proximate non-natural nucleobases and in pyrosequencing suggests that d[Me]isoC–disoG pairs, despite adopting a Watson–Crick pairing conformation in duplex nucleic acids (7), do not provide specific contacts necessary for efficient incorporation at subsequent positions. The lone pairs of electrons at N3 on purines and O2 on pyrimidines are symmetrically positioned about a pseudo 2-fold axis of natural base pairs (42) and can act as hydrogen bond acceptors to confirm correct nucleobase pairing. Crystal structures of complexes

of polymerase, duplex nucleic acid and dNTP have revealed minor groove hydrogen bonding interactions between the protein and hydrogen bond acceptors on post-insertion nucleobase pairs (31–33,43,44). Mismatched pairs, with associated conformational changes in the base pairing, cannot form these interactions and therefore disrupt the polymerase active site (45). The d[Me]isoC–disoG pair also cannot satisfy all polymerase minor groove hydrogen bonding sites because d[Me]isoC lacks the O_2 acceptor found in the natural pyrimidines. Slow extension beyond d[Me]isoC–disoG is likely a result of disruption of the polymerase active site by failure of the non-natural pair to form these contacts. The termination of extension observed at lower non-natural nucleotide concentrations (Figure 3) suggests that the disruption of these contacts hinders polymerase function more severely as the number of mismatched positions near the insertion site increases. Comparison of extension beyond mismatched positions in Figure 3 with Figure 4, in which non-natural nucleobases isolated in templates display no visible signal attenuation in the absence of complementary non-natural nucleotide, indicates that the thermophilic polymerases scanned ≤8 bp of the duplex preceding an insertion site. Our results are consistent with several studies that have found these interactions important in post-insertion extension (34,35,46).

However, despite deficient minor groove protein contacts, the steric equivalence of the d[Me]isoC–disoG pair to natural nucleobase pairs may provide an advantage in avoiding steric clashes with the protein. Extension by the thermophilic polymerases following complementary d[Me]isoC–disoG pairs proceeded more successfully than extension following mismatches involving the non-natural bases. Mispairing with d[Me]isoC or disoG at several proximate template positions caused complete termination of subsequent extension. In contrast, incorporation leading to d[Me]isoC–disoG pairs at these positions always allowed extension, although accompanied by signal attenuation. Hence, the ability of the thermophilic polymerases to extend several nucleotides following a non-natural pair was not as good as following natural complementary pairs, but better than following the mismatched pairs generated in our experiments.

Nucleobase structure and replication with the d[Me]isoC–disoG pair

Sequencing with d[Me]isoC and disoG analogs has highlighted the biological relevance of structural features of nucleobases in polymerase-mediated replication. The analogs reinforce the importance of interactions between polymerase and duplex near the site of insertion during extension. Minor groove interactions observed in natural duplexes with Family A polymerases are unable to form between the protein and d[Me]isoC–disoG pairs in a duplex and this is a likely cause of the slow extension following d[Me]isoC–disoG pairs. There may also be a steric screening of base pairs in this region of the duplex, because duplex d[Me]isoC–disoG pairs unable to form usual minor groove interactions nevertheless allow extension to proceed more smoothly than duplexes containing mismatches of the natural nucleobases with d[Me]isoC or disoG. Successful utilization of the disoG nucleobase demonstrates that mispairing resulting from nucleobase tautomerism can be minimized to yield workable replication, at least in applications, such as

sequencing, which do not demand high fidelity. In addition to helping understand polymerase–nucleic acid interaction, these observations should prove useful in the effective design of nucleobase analogs intended for use in polymerase-mediated replication.

Our sequencing experiments illustrate the utility of replication with a six-nucleobase system that includes d[Me]isoC and disoG. However, the experiments also reveal potential limitations of the pair. The d[Me]isoC–disoG pair, while isosteric to natural nucleobase pairs, does not have minor groove hydrogen bond acceptors, and subsequent extension following d[Me]isoC–disoG pairs, with either d[Me]isoC or disoG in the template strand, does not proceed as readily as with natural nucleobases. This will likely have undesirable consequences not seen in the sequencing experiments. Polymerases with significant 3′-exonuclease activity may display severe pausing at d[Me]isoC or disoG template positions. Nucleic acids containing d[Me]isoC and disoG should suffer relatively high mutation rates compared with high fidelity natural replication systems. *In vitro* selection experiments may be biased toward yielding nucleic acids with lower d[Me]isoC–disoG content simply because templates with high d[Me]isoC–disoG content are less readily replicated. Furthermore, although the d[Me]isoC–disoG pair demonstrates that it is possible to minimize potential mispairing problems stemming from tautomerism to yield a workable six-nucleobase sequencing system, nucleobases with tautomeric ambiguity may still be problematic as components of systems requiring higher fidelity. Interconversion between tautomers at the polymerase insertion site appears slow, and tautomeric ambiguity will probably slow replication. Appropriate engineering of the environment of the polymerase active site (47) or the carbon–nitrogen heterocycles of disoG (48) may be effective for higher fidelity replication of the d[Me]isoC–disoG pair.

REFERENCES

1. Benner,S.A., Alleman,R.K., Ellington,A.D., Ge,L., Glasfeld,A., Leanz,G.F., Krauch,T., MacPherson,L.J., Moroney,S.E., Piccirilli,A.J. and Weinhold,E. (1987) Natural selection, protein engineering, and the last rioborganism: rational model building in biochemistry. *Cold Spring Harb. Symp. Quant. Biol.*, **52**, 53–63.
2. Ogawa,A.K., Wu,Y., McMinn,D.L., Liu,J., Schultz,P.G. and Romesberg,F.E. (2000) Efforts toward the expansion of the genetic alphabet: information storage and replication with unnatural hydrophobic base pairs. *J. Am. Chem. Soc.*, **122**, 3274–3287.
3. Ishikawa,M., Hirao,I. and Yokoyama,S. (2000) Synthesis of 3-(2-deoxy-β-D-ribofuranosyl)pyridin-2-one and 2-amino-6-(*N,N*-dimethylamino)-9-(2-deoxy-β-D-ribofuranosyl)purine derivatives for an unnatural base pair. *Tetrahedron Lett.*, **41**, 3931–3934.

4. Piccirilli,J.A., Krauch,T., Moroney,S.E. and Benner,S.A. (1990) Enzymatic incorporation of a new base pair into DNA and RNA extends the genetic alphabet. *Nature*, **343**, 33–37.

5. Horn,T., Chang,C.-A. and Collins,M.L. (1995) Hybridization properties of the 5-methyl-isocytidine/isoguanosine base pair in synthetic oligonucleotides. *Tetrahedron Lett.*, **36**, 2033–2036.

6. Seela,F. and Wei,C. (1999) The base-pairing properties of 7-deaza-2′-deoxyisoguanosine and 2′-deoxyisoguanosine in oligonucleotide duplexes with parallel and antiparallel chain orientation. *Helv. Chim. Acta*, **82**, 726–745.

7. Chen,X., Kierzek,R. and Turner,D.H. (2001) Stability and structure of RNA duplexes containing isoguanosine and isocytidine. *J. Am. Chem. Soc.*, **123**, 1267–1274.

8. Johnson,S.C., Sherrill,C.B., Marshall,D.M., Moser,M.J. and Prudent,J.R. (2004) A third base pair for the polymerase chain reaction: inserting isoC and isoG. *Nucleic Acids Res.*, **32**, 1937–1941.

9. Collins,M.L., Irvine,B., Tyner,D., Fine,E., Zayati,C., Chang,C.-A., Horn,T., Ahle,D., Detmer,J., Shen,L.-P. *et al.* (1997) A branched DNA signal amplification assay for quantification of nucleic acid targets below 100 molecules/ml. *Nucleic Acids Res.*, **25**, 2979–2984.

10. Gleaves,C.A., Welle,J., Campbell,M., Elbeik,T., Ng,V., Taylor,P.E., Kuramoto,K., Aceituno,S., Lewalski,E., Joppa,B. *et al.* (2002) Multicenter evaluation of the Bayer VERSANT HIV-1 RNA 3.0 assay: analytical and clinical performance. *J. Clin. Virol.*, **25**, 205–216.

11. Elbeik,T., Surtihadi,J., Destree,M., Gorlin,J., Holodniy,M., Jortani,S.A., Kuramoto,K., Ng,V., Valdes,R.,Jr, Valsamakis,A. and Terrault,N.A. (2004) Multicenter evaluation of the performance characteristics of the Bayer VERSANT HCV RNA 3.0 assay (bDNA). *J. Clin. Microbiol.*, **42**, 563–569.

12. Sherrill,C.B., Marshall,D.J., Moser,M.J., Larsen,C.A., Daudé-Snow,L. and Prudent,J.R. (2003) Nucleic acid analysis using an expanded genetic alphabet to quench fluorescence. *J. Am. Chem. Soc.*, **126**, 4550–4556.

13. Switzer,C., Moroney,S.E. and Benner,S.A. (1989) Enzymatic incorporation of a new base pair into DNA and RNA. *J. Am. Chem. Soc.*, **111**, 8322–8323.

14. Switzer,C.Y., Moroney,S.E. and Benner,S.A. (1993) Enzymatic recognition of the base pair between isocytidine and isoguanosine. *Biochemistry*, **32**, 10489–10496.

15. Bain,J.D., Switzer,C., Chamberlin,A.R. and Benner,S.A. (1992) Ribosome-mediated incorporation of a non-standard amino acid into a peptide through expansion of the genetic code. *Nature*, **356**, 537–539.

16. Rice,K.P., Chaput,J.C., Cox,M.M. and Switzer,C. (2000) RecA protein promotes strand exchange with substrates containing isoguanine and 5-methyl isocytosine. *Biochemistry*, **39**, 10177–10188.

17. Moser,M.J. and Prudent,J.R. (2003) Enzymatic repair of an expanded genetic information system. *Nucleic Acids Res.*, **31**, 5048–5053.

18. Sismour,A.M., Lutz,S., Park,J.-H., Lutz,M.J., Boyer,P.L., Hughes,S.H. and Benner,S.A. (2004) PCR amplification of DNA containing non-standard base pairs by variants of reverse transcriptase from Human Immunodeficiency Virus-1. *Nucleic Acids Res.*, **32**, 728–735.

19. Liu,D., Moran,S. and Kool,E.T. (1997) Bi-stranded, multisite replication of a base pair between difluorotoluene and adenine: confirmation by 'inverse' sequencing. *Chem. Biol.*, **4**, 919–926.

20. Ohtsuki,T., Kimoto,M., Ishikawa,M., Mitsui,T., Hirao,I. and Yokoyama,S. (2001) Unnatural base pairs for specific transcription. *Proc. Natl Acad. Sci. USA*, **98**, 4922–4925.

21. Jurczyk,S.C., Kodra,J.T., Rozzell,J.D., Benner,S.A. and Battersby,T.R. (1998) Synthesis of oligonucleotides containing 2′-deoxyisoguanosine and 2′-deoxy-5-methylisocytidine using phosphoramidite chemistry. *Helv. Chim. Acta*, **81**, 793–811.

22. Wang,J., Jiang,J. and Battersby,T.R. (2002) Chemical stability of 2′-deoxy-5-methylisocytidine during oligodeoxynucleotide synthesis and deprotection. *Nucleosides Nucleotides Nucleic Acids*, **21**, 417–426.

23. Ronaghi,M., Karamohamed,S., Pettersson,B., Uhlén,M. and Nyrén,P. (1996) Real-time DNA sequencing using detection of pyrophosphate release. *Anal. Biochem.*, **242**, 84–89.

24. Ronaghi,M. (2001) Pyrosequencing sheds light on DNA sequencing. *Genome Res.*, **11**, 3–11.

25. Pyrosequencing Technical Note 103 (2000) Estimation of SNP allele frequencies.

26. Spurgeon,S.L. and Brandis,J.W. (2004) New DNA sequencing enzymes. In Kieleczawa,J. (ed.) *DNA Sequencing*. Jones and Bartlett, Sudbury, MA, pp. 35–54.

27. Tuerk,C. and Gold,L. (1990) Systematic evolution of ligands by exponential enrichment: RNA ligands to bacteriophage T4 DNA polymerase. *Science*, **249**, 505–510.

28. Roberts,C., Bandaru,R. and Switzer,C. (1997) Theoretical and experimental study of isoguanine and isocytosine: base pairing in an expanded genetic system. *J. Am. Chem. Soc.*, **119**, 4640–4649.

29. Robinson,H., Gao,Y.-G., Bauer,C., Roberts,C., Switzer,C. and Wang,A.H.-J. (1998) 2′-Deoxyisoguanosine adopts more than one tautomer to form base pairs with thymidine observed by high-resolution crystal structure analysis. *Biochemistry*, **37**, 10897–10905.

30. Maciejewska,A.M., Lichota,K.D. and Kuśmierek,J.T. (2003) Neighbouring bases in template influence base-pairing of isoguanine. *Biochem. J.*, **369**, 611–618.

31. Doublié,S., Tabor,S., Long,A.M., Richardson,C.C. and Ellenberger,T. (1998) Crystal structure of a bacteriophage T7 DNA replication complex at 2.2 Å resolution. *Nature*, **391**, 251–258.

32. Li,Y., Korolev,S. and Waksman,G. (1998) Crystal structures of open and closed forms of binary and ternary complexes of the large fragment of *Thermus aquaticus* DNA polymerase I: structural basis for nucleotide incorporation. *EMBO J.*, **17**, 7514–7525.

33. Li,Y. and Waksman,G. (2001) Crystal structures of ddATP-, ddTTP-, ddCTP-, and ddGTP-trapped ternary complex of Klentaq1: insights into nucleotide incorporation and selectivity. *Protein Sci.*, **10**, 1225–1233.

34. Hendrickson,C.L., Devine,K.G. and Benner,S.A. (2004) Probing minor groove recognition contacts by DNA polymerases and reverse transcriptases using 3-deazaz-2′-deoxyadenosine. *Nucleic Acids Res.*, **32**, 2241–2250.

35. Morales,J.C. and Kool,E.T. (1999) Minor groove interactions between polymerase and DNA: more essential to replication than hydrogen bonding? *J. Am. Chem. Soc.*, **121**, 2323–2324.

36. Morales,J.C. and Kool,E.T. (2000) Varied molecular interactions at the active sites of several DNA polymerases: nonpolar nucleoside isosteres as probes. *J. Am. Chem. Soc.*, **122**, 1001–1007.

37. Kamiya,H., Ueda,T., Ohgi,T., Matsukage,A. and Kasai,H. (1995) Misincorporation of dAMP opposite 2-hydroxyadenine, an oxidative form of adenine. *Nucleic Acids Res.*, **23**, 761–766.

38. Bukowska,A.M. and Kuśmierek,J.T. (1996) Miscoding properties of isoguanine (2-oxoadenine) studied in an AMV reverse transcriptase *in vitro* system. *Acta Biochim. Pol.*, **43**, 247–254.

39. Tor,Y. and Dervan,P.B. (1993) Site-specific enzymatic incorporation of an unnatural base, *N*6-(6-aminohexyl)isoguanosine, into RNA. *J. Am. Chem. Soc.*, **115**, 4461–4467.

40. Kamiya,H. and Kasai,H. (2000) Two DNA polymerases of *Escherichia coli* display distinct misinsertion specificities for 2-hydroxy-dATP during DNA synthesis. *Biochemistry*, **39**, 9508–9513.

41. Huang,M.-H., Arnheim,N. and Goodman,M.F. (1992) Extension of base mispairs by Taq DNA polymerase: implications for single nucleotide discrimination in PCR. *Nucleic Acids Res.*, **20**, 4567–4573.

42. Seeman,N.C., Rosenberg,J.M. and Rich,A. (1976) Sequence-specific recognition of double helical nucleic acids by proteins. *Proc. Natl Acad. Sci. USA*, **73**, 804–808.

43. Kiefer,J.R., Mao,C., Braman,J.C. and Beese,L.S. (1998) Visualizing DNA replication in a catalytically active *Bacillus* DNA polymerase crystal. *Nature*, **391**, 304–307.

44. Hsu,G.W., Ober,M., Carell,T. and Beese,L.S. (2004) Error-prone replication of oxidatively damaged DNA by a high-fidelity DNA polymerase. *Nature*, **431**, 217–221.

45. Johnson,S.J. and Beese,L.S. (2004) Structures of mismatch replication errors observed in a DNA polymerase. *Cell*, **116**, 803–816.

46. Morales,J.C. and Kool,E.T. (2000) Functional hydrogen-bonding map of the minor groove binding tracks of six DNA polymerases. *Biochemistry*, **39**, 12979–12988.

47. Chaput,J.C. and Switzer,C. (2000) Non-enzymatic transcription of an isoG·isoC base pair. *J. Am. Chem. Soc.*, **122**, 12866–12867.

48. Martinot,T.A. and Benner,S.A. (2004) Artificial genetic systems: exploiting the 'aromaticity' formalism to improve the tautomeric ratio for isoguanosine derivatives. *J. Org. Chem.*, **69**, 3972–3975.