AmpliTaq® DNA Polymerase, FS Dye-Terminator Sequencing: Analysis of Peak Height Patterns

BioTechniques 21:694-699 (October 1996)

L.T. Parker, H. Zakeri, Q. Deng[1], S. Spurgeon[2], P.-Y. Kwok and D.A. Nickerson[1]
Washington University School of Medicine, St. Louis, MO; [1]University of Washington School of Medicine, Seattle, WA; [2]Perkin-Elmer/Applied Biosystems Division, Foster City, CA, USA

ABSTRACT

Taq *DNA polymerases in which the phenylalanine is substituted by a tyrosine at position 667 (Taq F667Y) are members of a new class of DNA polymerases that incorporate chain-terminating dideoxyribonucleoside triphosphates (ddNTPs) much more efficiently than the wild-type* Taq *DNA polymerase. Improved incorporation of ddNTPs into DNA during cycle sequencing using AmpliTaqÒ DNA polymerase, FS (Taq-FS, a member of the Taq F667Y family), and dye-labeled primers results in nearly uniform peak heights in the sequencing trace. This is not the case when dye-labeled ddNTPs are used in Taq-FS cycle sequencing reactions. While the rate of dye-terminator incorporation is more efficient with Taq-FS, the peak pattern is still highly variable and different from that produced by the wild-type enzyme. We have systematically examined pairs of sequence-tagged sites that vary at only a single nucleotide to determine how base changes influence the peak heights of neighboring bases in sequencing traces generated by the Taq-FS dye-terminator chemistry. In 31 of 64 possible 3-base windows (48%), we find that the peak height of a particular base can be predicted by knowing just one or two bases 5ℓ to the base in question. We have also compared and contrasted the peak patterns produced by the Taq-FS enzyme with those previously identified for the wild-type enzyme. Establishing the patterns in peak heights within local sequence contexts can improve the accuracy of base-calling and the identification of polymorphisms/mutations when using the Taq-FS dye-terminator cycle-sequencing chemistry.*

INTRODUCTION

Direct sequencing of DNA amplified by the polymerase chain reaction (PCR) is being used increasingly in large-scale sequencing projects and in scanning specific regions for polymorphisms and mutations (1–3,8). Among the available sequencing chemistries, cycle-sequencing with thermostable DNA polymerases and dye-labeled dideoxy chain terminators is the most versatile because it eliminates the need for specially modified sequencing primers, since the PCR primers can also be used as sequencing primers (6,8). It is also the simplest sequencing chemistry because only one sequencing reaction is needed for each DNA sample.

Recently, a major improvement in this sequencing chemistry has come from the development of mutant *Taq* DNA polymerases where a phenylalanine residue in the active site has been replaced by a tyrosine (F667Y) (7). Like bacteriophage T7 DNA polymerase, these mutant *Taq* DNA polymerases incorporate chain-terminating dideoxyribonucleoside triphosphates (ddNTPs) into DNA much more efficiently than wild-type *Taq* DNA polymerase (7). This property greatly improves the quality of the sequencing data obtained with this chemistry and reduces the amount of ddNTPs required for each sequencing reaction (5,7). Although improved incorporation of ddNTPs into DNA during cycle sequencing using dye-labeled sequencing primers results in nearly uniform peak heights in a sequencing trace (5), this is not the case when dye-labeled ddNTPs are used in the cycle sequencing reactions. While the rate of dye-terminator incorporation is more efficient when a mutant *Taq* DNA polymerase is used, the peak pattern is still uneven and different from that produced by the wild-type enzyme. Uneven peak heights can decrease the accuracy of base-calling and make mutation and polymorphism detection more difficult. The ability to predict peak heights in a given local sequence context can help to compensate for these drawbacks.

Analysis of the sequencing traces in cycle sequencing using the wild-type enzyme and dye-terminators has revealed sequence context-dependent trends in patterns of dye-terminator incorporation as well as a number of reproducible artifacts (4). Knowledge of these trends and artifacts has increased the accuracy and confidence in identifying polymorphisms and mutations in a sequencing trace and in editing sequences obtained using this chemistry. To increase the accuracy of sequence analysis with the new mutant *Taq* DNA polymerase (F667Y), we have analyzed the trends and reproducibility of dye-terminator incorporation in sequencing traces produced by one member of this

Table 1. Effect of 5′ Bases on Peak Heights of the 3′ Base

Base String[a]	Mean Peak Height (mm)[b]	Range (mm)	Percent <10 mm (small)[c]	Percent 10–20 mm (average)[d]	Percent >20 mm (large)[e]
1. CA**A**	7.0 ± 2.1	3–10	39/39 (100%)		
2. TA**A**	6.9 ± 2.1	3–13	60/65 (92%)		
3. AC**A**	7.3 ± 2.0	5–13	38/41 (93%)		
4. NG**A**	29.1 ± 3.0	16–30			117/120 (98%)
5. AA**C**	14.2 ± 2.3	11–20		38/38 (100%)	
6. GA**C**	13.7 ± 4.4	8–29		26/30 (87%)	
7. TA**C**	15.1 ± 4.0	10–24		27/30 (90%)	
8. TG**C**	16.8 ± 3.5	10–24		26/30 (87%)	
9. CT**C**	25.0 ± 4.5	17–30			30/35 (86%)
10. TT**C**	27.1 ± 4.6	13–30			29/31 (94%)
11. AA**G**	7.6 ± 3.2	2–14	28/30 (93%)		
12. CA**G**	6.1 ± 2.8	3–15	49/51 (96%)		
13. TA**G**	5.5 ± 2.4	2–11	29/30 (97%)		
14. AC**G**	6.0 ± 2.7	2–13	27/30 (90%)		
15. GC**G**	7.4 ± 2.7	3–17	36/41 (88%)		
16. TC**G**	7.1 ± 4.2	3–23	28/32 (88%)		
17. NG**G**	28.1 ± 3.6	16–30			117/120 (98%)
18. AC**T**	12.8 ± 4.0	6–25		28/33 (85%)	
19. GC**T**	11.9 ± 2.3	7–17		27/30 (90%)	
20. TC**T**	12.3 ± 2.7	9–20		29/31 (94%)	
21. TG**T**	26.5 ± 5.5	10–30			35/41 (85%)
22. NT**T**	8.3 ± 2.0	3–13	143/155(92%)		

[a]The 3′ base is in bold type.
[b]Mean peak height followed by standard deviation.
[c]The proportion of cases for a particular window where the 3′ base has a peak height of ≤1/3 of full scale (≤10 mm, designated as "small").
[d]The proportion of cases for a particular window where the 3′ base has a peak height of 1/3–2/3 of full scale (10–20 mm, designated as "average").
[e]The proportion of cases for a particular window where the 3′ base has a peak height of ≥2/3 of full scale (≥20 mm, designated as "large").

new family of polymerases, AmpliTaq® DNA Polymerase, FS. These results are compared to the patterns previously established for dye-terminator incorporation with the wild-type enzyme (4).

MATERIALS AND METHODS

PCR Amplification and Purification of PCR Products

Human genomic DNA (16 ng) from individuals with known genotype were amplified in 40-µL reactions, gel-purified and eluted into 50 µL of water as previously described in detail (2,4). The purified DNA was used as sequencing template without further characterization.

Taq-FS Cycle-Sequencing Using Dye-Labeled Terminators

Cycle sequencing was performed on the GeneAmp® PCR System 9600 (Perkin-Elmer, Norwalk, CT, USA) utilizing the PRISM™ Ready Dye Terminator Cycle Sequencing kit with AmpliTaq DNA polymerase, FS (Taq-FS; Perkin-Elmer/Applied Biosystems Division [PE/ABI], Foster City, CA, USA), according to the manufacturer's directions. Briefly, the gel-purified DNA (10.4 µL) was added to a MicroAmp™ reaction tube (Perkin-Elmer) containing 3.2 pmol of sequencing primer and 8.0 µL of premix (containing buffer, dNTPs, dye-labeled ddNTPs and Taq-FS/pyrophosphatase). After the initial denaturation at 96°C for 2 min, the reaction mixture was incubated for 25 cycles at 96°C for 15 s, 50°C for 1 s and 60°C for 4 min. Excess dye-labeled terminators were removed from the extension products by spin

column purification (CentriSep™ spin columns; Princeton Separations, Adelphia, NJ, USA) according to the manufacturer's directions. [Note: The excess dyes can also be removed successfully by ethanol precipitation, which avoids the labor and expense associated with spin-column purification.] Once separated, the extension products were evaporated to dryness under reduced pressure (SpeedVac®; Savant Instruments, Farmingdale, NY, USA). Each sample was resuspended in 3 µL of loading buffer (5:1, 1% deionized formamide/50 mM EDTA, pH 8.0), heated for 2 min at 90°C and loaded onto an Applied Biosystems 373A sequencer (PE/ABI) according to the manufacturer's instructions.

Analysis of Sequencing Traces

Homozygous individuals containing each of the alternative alleles for 33 different polymorphic sequence-tagged sites (pSTSs: 66 sequencing traces altogether) were sequenced using the Taq-FS enzyme. These traces represent all of the possible three-base windows (64 in total) where the middle base in the window is the only variant base in each pair of sequences. In addition, 20 control sequencing traces, produced using the Taq-FS dye-terminator chemistry, were also analyzed to extend and confirm these trends observed in the allelic pairs of sequencing traces to provide a significantly larger sample of each of the three-base windows. Control sequencing traces were normalized by reanalyzing 3600 data points of the sequencing read using the ABI 373A Analysis Software to produce identical spacing between the traces. To ensure that there was as much uniformity as possible, measurements were taken only from the middle portion (bases 100–200) of the trace. This strategy avoids potential problems that could arise by comparing the weaker peaks toward the end of the sequencing run to the stronger peaks at the beginning of the sequencing run. For each of the 64 possible three-base windows, the base most 3′ in each window was measured in the context of the two bases 5′ to it. When a 3′-base peak was found to be equal to or less than 1/3 of full scale, it was defined as a "small" peak. Similarly, when a 3′-base peak was found to be equal to or greater than 2/3 of full scale, it was defined as a "large" peak. When a 3′-base peak was found to be between 1/3 to 2/3 of full scale, it was defined as an "average" peak. For example, if the full height of the analyzed sequencing traces was 30 mm, the "small" peaks measured 10 mm or less in height, "large" peaks measured 20 mm or more in height and average peaks ranged between 10 and 20 mm.

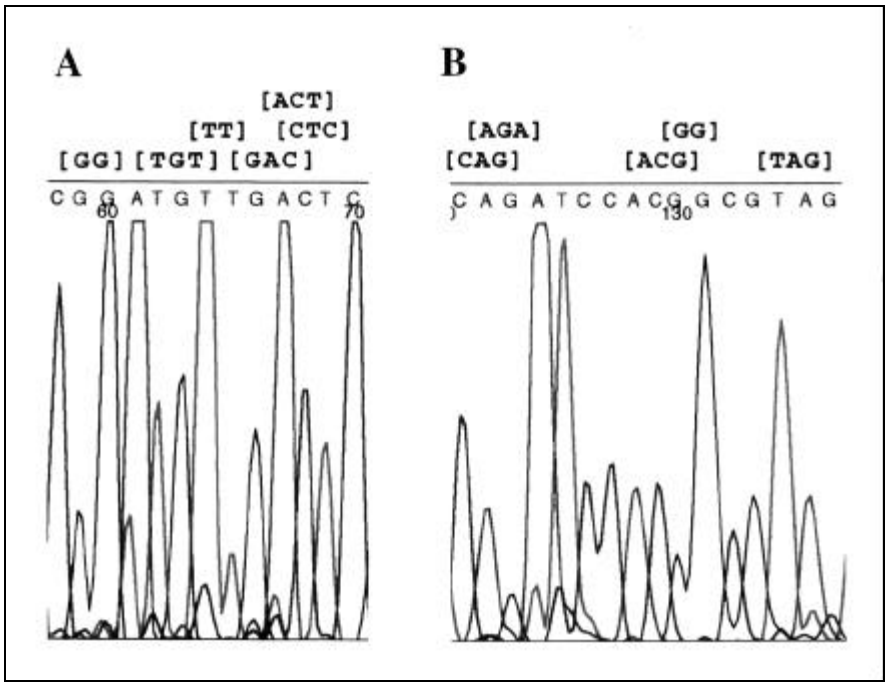

Figure 1. Effects of 5′ base(s) on peak heights. Panels A and B show the patterns of peak height variation in Taq-FS cycle sequencing using dye-labeled dideoxy chain terminators. (A) The G peak following G is large, the T peak following TG is large, the T peak following T is small, the C peak following GA is average, the T peak following AC is average and the C peak following CT is large. (B) The G peak following CA is small, the A peak following AG is large, the G peak following AC is small, the G peak following G is large and the G peak after TA is small.

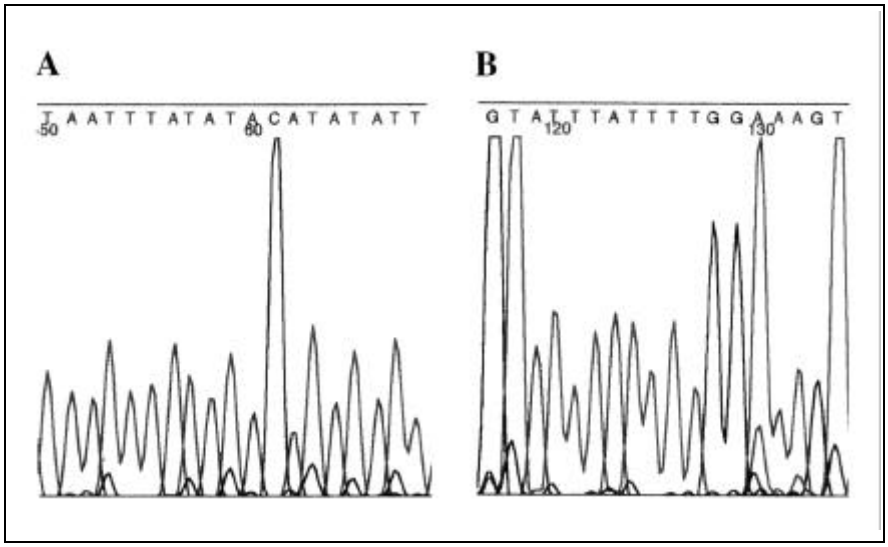

Figure 2. Sequencing artifacts. (A) T peaks after A's contain artifactual C signals. (B) T peaks after G's contain artifactual C signals and A peaks after G's contain artifactual T signals. These artifacts occur frequently, but are not always present.

Table 2. Comparison of Peak Patterns in Taq-FS and AmpliTaq Sequencing Traces

	Base string[a]	Peak Height of 3′ Base	
		Taq-FS	AmpliTaq
1.	NT**T**	small	small
2.	CA**A**	small	small
3.	TA**A**	small	small
4.	CT**C**	large	large
5.	TT**C**	large	large
6.	AC**A**	small	large
7.	TC**T**	average	large
8.	NG**A**	large	small/average
9.	NG**G**	large	variable
10.	AA**G**	small	average/large
11.	CA**G**	small	average/large
12.	TA**G**	small	average/large
13.	AC**G**	small	average/large
14.	GC**G**	small	average/large
15.	TC**G**	small	average/large
16.	AC**T**	average	variable
17.	GC**T**	average	variable
18.	AA**C**	average	variable
19.	TG**T**	large	small/average
20.	NG**C**	average/large	small
21.	AA**A**	small/average	small
22.	GA**A**	small/average	small
23.	AT**A**	small/average	small
24.	GT**A**	small/average	small
25.	GC**A**	small/average	large
26.	TC**A**	small/average	large
27.	AC**C**	average/large	large
28.	GC**C**	average/large	large
29.	AT**C**	average/large	large
30.	TA**T**	small/average	large

[a]The 3′ base is in bold type.

RESULTS AND DISCUSSION

Trends in Dye-Terminator Incorporation by Taq-FS

Analysis of all three-base windows within the polymorphic STSs (64 windows altogether) revealed several trends that were confirmed by analyzing similar three-base windows (>30 independent observations of each window) in 20 control sequencing traces. First, the peak patterns produced with Taq-FS were highly reproducible and predictable in specific sequence contexts. In cases where the two homozygotes of a substitution polymorphism were analyzed (see Table 1 for specific examples), we found that the only affected peak height in these three-base windows was the base immediately 3′ to the substituted nucleotide. Second, only a small number of bases 5′ to a particular base significantly influenced its peak height. For example, among the 64 three-base windows, 31 (48%) consistently showed small, average or large peak heights for the 3′ base. In fact, 12 of these 31 cases were represented by NGA, NGG or NTT (see Table 1). In cases where base-strings of NGA or NGG were found, the 3′ base was consistently "large" (>2/3 full scale in height, see Figure 1). In the base-string of NTT, the 3′ base was

found to be consistently "small" (<1/3 full scale in height, see Figure 1A). In each of these instances, knowing only one base immediately 5′ to a particular base was sufficient to predict its peak height.

Several other reproducible trends in the pattern of dye-terminator incorporation by Taq-FS were also apparent. In three-base strings of AAG, CAG, TAG, CAA, TAA, ACG, GCG, TCG or ACA, the 3′ base was found to be consistently "small". In strings of ACT, GCT, TCT, AAC, GAC, TAC or TGC, the 3′ base was consistently "average" in size (1/3 to 2/3 of full scale). For base-strings of CTC, TTC and TGT, the 3′ base was consistently "large". See Figure 1 for selected examples of these trends.

It is interesting to note that we did not find any consistent sequencing artifacts in sequencing traces produced by Taq-FS. In a significant number of cases, however, small C peaks were found under the T peaks in AT strings (90%, Figure 2A), small T peaks were found under the A peak in GA strings (75%, Figure 2B) and small C peaks were found under the T peak in GT strings (90%, Figure 2B) even in high-quality sequencing traces.

When the peak patterns produced by Taq-FS were compared to those obtained using the wild-type enzyme, AmpliTaq, the majority (ca. 80%) of the three-base windows showed patterns (explained below) that differed considerably depending on which enzyme was used in the sequence analysis.

I. Peak patterns that are similar using Taq-FS and AmpliTaq. There are eight 3-base strings that give similar peak patterns with both enzymes. The NTT, CAA and TAA windows all result in small peaks at the 3′ base with either enzyme, while the windows CTC and TTC result in large 3′ C peaks (Table 2).

II. Peak patterns that are different but predictable with both enzymes. Two 3-base strings, namely, ACA and TCT, give different but predictable 3′-peak heights for each enzyme. In sequencing traces produced with AmpliTaq, one finds large 3′ A and 3′ T peaks, whereas in Taq-FS sequencing traces one finds small 3′ A and average 3′ T peaks for ACA and TCT, respectively (Table 2).

III. Peak patterns predictable with Taq-FS but not with AmpliTaq. Eighteen 3-base strings are predictable when Taq-FS is used, but are variable when the wild-type enzyme is used. For example, in Taq-FS sequencing traces, the 3′ base peak is consistently small with base strings AAG, CAG, TAG, ACG, GCG and TCG; large with NGA, NGG and TGT; and average with AAC, ACT and GCT. These 3-base strings produce no consistent patterns in sequencing traces generated by AmpliTaq (Table 2).

IV. Peak patterns variable with Taq-FS but predictable with AmpliTaq. Fourteen 3-base strings have predictable peak heights for the 3′ base in sequencing traces generated with AmpliTaq. These include NGC, AAA, GAA, ATA and GTA (always small); GCA, TCA, ACC, GCC, ATC and TAT (always large). However, there were no consistent trends when Taq-FS was used (Table 2).

Polymorphism/Mutation Detection by Taq-FS Sequencing

In sequencing pSTSs with the wild-type enzyme, we previously found that changes in the peak heights of bases immediately 3′ to polymorphic bases were useful in providing additional criteria for identifying heterozygous bases in a sequencing trace (2,4). In the current study, we have found that the trends in 3′ bases can be used to predict which heterozygous base combinations may be missed by existing base-calling software when Taq-FS sequencing chemistry is used. An example of this

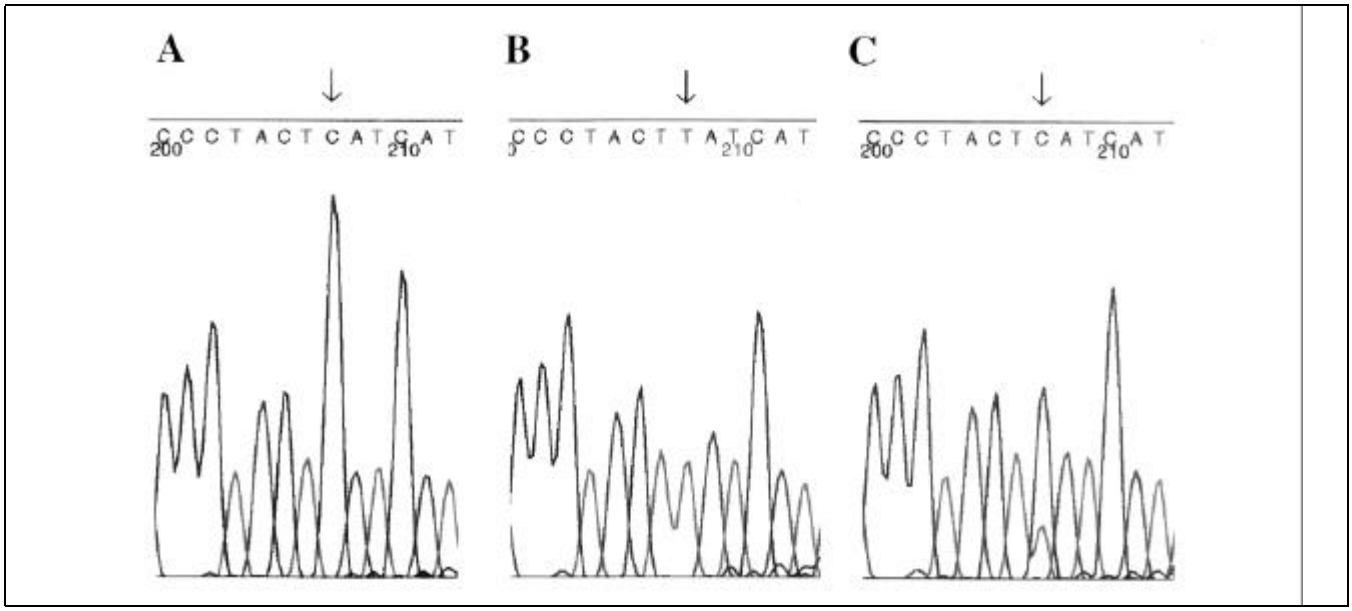

Figure 3. Disparity in peak heights at the polymorphic site in a heterozygote. (A) Sequencing trace of the homozygote with the allele CT**C**. The 3′ C peak is large and is about twice the size of the C peak in a heterozygote. (B) Sequencing trace of the homozygote with the allele CT**T**. The 3′ T peak is small but is still twice the size of the T peak in a heterozygote. (C) Sequencing trace of the heterozygote with CT(**C/T**). The base-calling program called this heterozygous base a "C" rather than an "N".

situation is shown in Figure 3. In Figure 3A, the window CTC shows the characteristically large 3′ C peak. In Figure 3B, the allelic CTT window shows the characteristically small 3′ T peak. When a heterozygote C/T is present at the 3′ base (Figure 3C), the C peak at the polymorphic site is about one-half the size of the C peak in the homozygous CTC window and the T-peak at this location is about one-half the size of that in the homozygous CTT. Because of the large disparity in peak heights for these two bases, one could predict that this would be a difficult combination for the base-calling program to recognize in the heterozygote. Indeed, the base-calling software assigned a "C" for the polymorphic base in the heterozygote instead of an "N" (Figure 3C).

Although Taq-FS incorporates dye-terminators much more efficiently than does the wild-type enzyme, the patterns produced by this improved DNA polymerase are still quite variable. As our results show, the pattern of dye-terminator incorporation in many instances is highly reproducible and dependent on the local sequence context. For example, ddGTP is incorporated much slower than average by Taq-FS when it is following an A base. This leads to the production of a very small G peak following an A peak. Conversely, ddGTP is incorporated much faster than average when it follows another G peak, therefore producing a very large G peak.

By examining sequences of PCR products (STSs) containing more than 80 single base-pair substitutions, we have identified several general trends among the variable peak heights produced by Taq-FS cycle sequencing using dye-labeled terminators. Our analysis reveals that, in 12 out of the possible 64 three-base windows, the peak height of the 3′ base could be accurately predicted by knowing only the base immediately 5′ to the base in question (NGA, NGG and NTT, see Table 1). In addition, in 19 out of the remaining 52 three-base combinations, the peak height of the 3′ base can be predicted by knowing the 2 bases immediately 5′ to it. In short, in almost half of the cases (48%), one can predict the peak height of a base just by knowing the 2 bases immediately 5′ to it. In the other half of the cases where the pattern of the 3′ base does not display a consistent trend, the height of the 3′ base is usually in the average or higher range (>1/3 full scale). Therefore, these windows are rarely problematic in base assignment. It is possible that the peak heights of 3′ bases in these combinations might be predictable if larger windows of bases were considered. Further elucidation of trends involving more than a string of 3 bases requires a much larger data set and a more rigorous algorithm, which is currently under development.

Some of the patterns found for Taq-FS sequencing are similar to those identified with the wild-type enzyme (AmpliTaq). In this regard, both DNA polymerases give sequencing traces with some troublesome small peaks and these would be excellent targets to address when developing improvements in the sequencing chemistry, i.e., new mutant enzymes or improved reaction conditions. Unlike the extremely small peaks (usually much less than 1/6 of full scale) found in sequencing traces produced by the wild-type enzyme (such as the T peak following another T, the A peak following another A or the C peak following a G), the small peaks found in sequencing traces generated by Taq-FS are usually between 1/6 to 1/3 of full scale, making it relatively easy to assign the bases accurately. The one exception in sequencing with Taq-FS and dye-labeled terminators is the G peak following an A. Here, the G peaks can be extremely small, with a significant proportion of them at <1/6 of full scale.

The ability to predict the expected peak height in sequencing traces can improve base-calling accuracy and help one to recognize heterozygotes with greater confidence. Furthermore, it is possible to predict which overlapping heterozygous bases will be difficult to recognize because of the potential disparity of peak heights for those windows with reproducible trends (Table 1 and Figure 3). Incorporating these trends into base-calling programs will enhance their accuracy and allow for greater automation in the identification of polymorphisms/mutations by DNA sequence analysis.

ACKNOWLEDGMENTS

We thank I. Bauer-Sardina for her technical assistance. This work was supported by grants from the Department of Energy to D.A.N. and P.Y.K. (D.E.-FG06-94ER-61909), the National Science Foundation to D.A.N. (DIR 8809710) and a National Institutes of Health Training Grant to L.T.P. (5T32AR07284).

REFERENCES

1. **Gibbs, R.A., P.N. Nguyen, L.J. McBride, S.M. Koepf and C.T. Caskey.** 1989. Identification of mutations leading to the Lesch-Nyhan syndrome by automated direct DNA sequencing of in vitro amplified cDNA. Proc. Natl. Acad. Sci. USA *86*:1919-1923.
2. **Kwok, P.Y., C. Carlson, T.D. Yager, W. Ankener and D.A. Nickerson.** 1994. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. Genomics *23*:138-144.
3. **Lee, L.G., C.R. Connell, S.L. Woo, R.D. Cheng, B.F. McArdle, C.W. Fuller, N.D. Halloran and R.K. Wilson.** 1992. DNA sequencing with dye-labeled terminators and T7 DNA polymerase: effect of dyes and dNTPs on incorporation of dye-terminators and probability analysis of termination fragments. Nucleic Acids Res. *20*:2471-2483.
4. **Parker, L.T., Q. Deng., H. Zakeri, C. Carlson, D.A. Nickerson and P.-Y. Kwok.** 1995. Peak height variations in automated sequencing of PCR products using *Taq* dye-terminator chemistry. BioTechniques *19*:116-121.
5. **Reeve, M.A. and C.W. Fuller.** 1995. A novel thermostable polymerase for DNA sequencing. Nature *376*:796-797.
6. **Rosenthal, A. and D.S. Charnock-Jones.** 1992. New protocols for DNA sequencing with dye terminators. DNA Seq. *3*:61-64.
7. **Tabor, S. and C.C. Richardson.** 1995. A single residue in DNA polymerases of the Escherichia coli DNA polymerase I family is critical for distinguishing between deoxy- and dideoxyribonucleotides. Proc. Natl. Acad. Sci. USA 92:6339-6343.
8. **Tracy, T.E. and L.S. Mulcahy.** 1991. A simple method for direct automated sequencing of PCR fragments. BioTechniques *11*:68-75.

Received 8 April 1996; accepted 20 June 1996.

Address correspondence to:

Linda T. Parker
Division of Dermatology
Washington University School of Medicine
660 S. Euclid Ave, Box 8123
St. Louis, MO, USA
Internet: linda@psts.wustl.edu