Approach to genotyping errors caused by nontemplated nucleotide addition by Taq DNA polymerase.

J R Smith, J D Carpten, M J Brownstein, et al.

Genome Res. 1995 5: 312-317
Access the most recent version at doi:10.1101/gr.5.3.312

Approach to Genotyping Errors Caused by Nontemplated Nucleotide Addition by *Taq* DNA Polymerase

Jeffrey R. Smith,[1,2,5,6] John D. Carpten,[1,5] Michael J. Brownstein,[3] Soumitra Ghosh,[1] Victoria L. Magnuson,[1] Dennis A. Gilbert,[4] Jeffrey M. Trent,[1] and Francis S. Collins[1]

[1]National Center for Human Genome Research, National Institutes of Health, Bethesda, Maryland 20892; [2]Department of Internal Medicine, University of Michigan Medical Center, Ann Arbor, Michigan 48109; [3]National Institute of Mental Health, National Institutes of Health, Bethesda, Maryland 20892; [4]Applied Biosystems Division, Perkin Elmer, Foster City, California 94404

Thermostable DNA polymerases can catalyze nontemplated addition of a nucleotide to the 3′ end of amplification products. This presents a potential source of error in genotyping studies employing *Taq* DNA polymerase to amplify microsatellite loci. Although the activity is marker specific, experimental variation is often seen in the degree of modification. Consequently, for a given microsatellite marker, an allele may be inconsistently identified as either the unmodified or modified amplification product. Full automation of high-throughput genotyping has been hampered by the need for manual editing of data because of this source of allele misidentification. In this study we estimate a 1% to 3% error rate attributable to nontemplated nucleotide addition in the ABI PRISM genotyping system. We present a PCR-based strategy to minimize this source of error.

The genotyping community has been striving to develop fully automated allele sizing of short tandem repeat (STR) sequences distributed throughout the human genome to facilitate linkage studies. Adaptation of automated DNA sequencers to genotyping has greatly increased throughput by allowing fluorescently labeled PCR products of multiple markers to be resolved in each gel lane.[1,2] For the Applied Biosystems DNA sequencer, data collection and allele sizing are facilitated by GENESCAN and GENOTYPER software. However, several sources of error in allele identification necessitate manual editing of data and thereby limit the potential throughput of this system.

Errors in accurate allele identification by GENOTYPER are often caused by the incorrect labeling of either a spurious noise peak or a peak 1 nucleotide greater in size than the true allele. The latter is commonly the result of the nontemplated addition of a single nucleotide, predominantly adenosine, to the 3′ end of the fluorescently labeled strand by *Taq* DNA polymerase.[3,4] The degree to which a marker is subject to "+A" modification is relatively marker specific, though the variables contributing to this specificity have not been defined.

Markers that are consistently amplified without the +A addition are sized as the true allele (T) by GENOTYPER. Those consistently modified as the +A product are sized 1 nucleotide greater than the true allele (T + 1) by GENOTYPER. In either case, no difficulties are introduced into the genotyping analysis. However, markers that are only partially modified may be sized as either the T or T + 1 product, depending on their relative peak heights for a particular reaction. In the most complex case for a dinucleotide repeat, a ladder of bands spaced 1 nucleotide apart is generated by partial +A modification of the two true alleles as well as each of the 2-bp "stutter" products. Despite clear single-base resolution, GENOTYPER's filtering algorithm often fails to consistently identify the T or the T + 1 band. Genotyping error is introduced as a consequence; an allele may be assigned as the true allele in some family members and as the +A-modified allele in other family members. Even an allele of a single individual, amplified or electrophoresed repeatedly, may be sized inconsistently. The existence of actual alleles separated by a single base pair for a minority of "dinucleotide" repeat markers further complicates flagging of these errors for manual editing.

The study presented here addresses the contribution of nontemplated nucleotide addition by *Taq* DNA polymerase to a fraction of amplified allele product as a source of genotyping error. We assess the scope of this problem for genotyping and discuss PCR methods to minimize this source of allele misidentification.

[5]The first two authors contributed equally to this work.
[6]Corresponding author.
E-MAIL jsmith@nchgr.nih.gov and jdc@nchgr.nih.gov; FAX (301) 480-0828.

5:312–317 ©1995 by Cold Spring Harbor Laboratory Press ISSN 1054-9803/95 $5.00

MATERIALS AND METHODS

PCR Conditions

All PCR reactions were carried out using 60 ng of template DNA corresponding to a CEPH control individual 884-01, -02, -03, -04, -05, -06, -07, -08, -15, -16, -17, or -18 (BIOS Laboratories, New Haven, CT). Fifteen-microliter reaction volumes contained 50 mM potassium chloride, 10 mM Tris-HCl (pH 8.3), 333 nM each forward and reverse primer, 0.6 units of AmpliTaq DNA polymerase (Perkin-Elmer, Norwalk, CT), and 250 μM each dNTP (dATP, dCTP, dGTP, dTTP). Magnesium chloride concentration was optimized for each primer set and ranged from 1.0 to 3.0 mM.

Dinucleotide Repeat Markers

Fluorescently labeled primer panels of human dinucleotide repeat markers were from the ABI PRISM Linkage Mapping Set (Applied Biosystems Division/Perkin-Elmer, Foster City, CA). The test version of panel 2 included D1S199, D1S207, D1S424, D1S413, D1S238*, D1S252, D1S209, D1S468*, D1S216, D1S244, D1S214, D1S423, D1S498, D1S218, and D1S425*. The test version of panel 6 included D3S1262, D3S1300*, D4S413*, D4S398, D3S1304, D3S1297*, D4S415, D4S419, D4S394*, D3S1232, D4S1566, D3S1238, and D3S196. The test version of panel 9 included D6S264, D6S276, D6S308*, D6S261, D5S426*, D5S392, D6S257, D6S286*, and D5S429*. The test version of panel 20 included D14S80, D14S81*, D14S283, D14S63*, D14S72, D14S78, D14S74, D14S288, D14S68*, and D14S285*. Panel 21 was the marketed version and included D16S405*, D16S401, D16S411*, D15S130*, D16S515*, D16S520*, D15S165, D15S131, D16S503*, D15S127, D16S511, D15S153*, and D15S117*. Additional markers used were D3S1259, D3S1278, D3S1293, D3S1311, D4S428, D5S433*, D4S1565*, D5S406, D5S407*, D5S644, D6S260, D6S262, D6S281*, D6S305*, D6S309, D12S352, DXS986, DXS987, DXS990*, DXS992, DXS1001, DXS1060*, DXS1106*, DXS1193, and DXS1227*. Certain primer pairs of each panel were modified from the published sequences by ABI to prevent overlapping allele size ranges among markers of a given dye label and are indicated by asterisks, above. One primer from each pair was labeled with a fluorescent dye phosphoramidite, either 6-FAM (blue), HEX (yellow), or TET (green). PCR products for each marker of a given panel may be pooled together and coelectrophoresed unambiguously.

PCR Thermocycling Conditions

Three PCR thermocycling protocols were employed using Perkin-Elmer model 9600 thermocyclers (Perkin-Elmer, Norwalk, CT):

1. A two-step protocol: 95°C for 5 min followed by 10 cycles of 94°C for 15 sec, 55 °C for 15 sec, followed by an additional 23 cycles of 89°C for 15 sec, 55°C for 15 sec.
2. A three-step/10-min final extention (ABI PRISM) protocol: 95°C for 5 min followed by 10 cycles of 94°C for 15 sec, 55°C for 15 sec, 72°C for 30 sec, followed by an additional 20 cycles of 89°C for 15 sec, 55°C for 15 sec, 72°C for 30 sec, followed by a final extension at 72°C for 10 min.
3. A three-step/variable final extension protocol: 95°C for 5 min followed by 10 cycles of 94°C for 15 sec, 55°C for 15 sec, 72°C for 30 sec, followed by an additional 20 cycles of 89°C for 15 sec, 55°C for 15 sec, 72°C for 30 sec, followed by a final extension at 72°C for 30, 60, or 90 min.

Analysis of PCR Products

The markers of a given panel were independently amplified from a single DNA template, and the reaction products were then pooled. The volume of each reaction product that was pooled varied from 1.5 to 15 μl (total pool volume of 100 μl), as required to equalize marker signal. The volume of the pool (1.5 μl) was subsequently mixed with 2.5 μl of formamide, 0.5 μl of blue dextran loading dye, and 0.5 μl of internal size standard GS-500 (Applied Biosystems Division/Perkin-Elmer, Foster City, CA). The size standard contained DNA fragments fluorescently labeled with the dye phosphoramidite TAMRA (red), and they range in size from 50 to 500 bp. After heat denaturation, 3.7 μl of the pool/size standard mix was electrophoresed in one lane of a 7% denaturing polyacrylamide gel (Bio-Rad, Hercules, CA) at 15 W of constant power using the ABI model 373A automated sequencer (12-cm well-to-read, filter set B). Each gel lane was loaded with a pooled panel of markers for a unique genomic DNA template. Fluorescently labeled DNA fragments were analyzed, and genotype data were generated using ABI GENESCAN 672 (v. 1.2.2-1) software and ABI GENOTYPER (v. 1.1r8) DNA fragment analysis software. The default label-filtering algorithm of GENOTYPER was used to assign allele peaks. Filter-labels options included the following: (1) Remove labels from peaks whose height is <32% of the highest peak in a category's range; (2) remove labels from peaks preceded by a higher, labeled peak within 1.6 bp; and (3) remove labels from peaks followed by a higher labeled peak within 3 bp.

RESULTS AND DISCUSSION

To estimate the genotyping error rate attributable to nontemplated nucleotide addition, four lines of experiments were undertaken. In the first, a small number of DNA samples were repeatedly amplified with a single group of markers to determine reproducibility. Thirteen dinucleotide repeat markers (a test version of the ABI PRISM panel 6) were used to amplify DNA from three to eight members of CEPH family 884 by the ABI PRISM thermocycling protocol. The products were analyzed on an ABI 373A automated DNA sequencer. This experiment was repeated 12 times. The fraction of + A product for each marker was estimated by taking the mean ratio of peak heights $(h_{T+1})/(h_T + h_{T+1})$ for each allele of eight CEPH family 884 subjects.

Figure 1 provides an example of inconsistently identified alleles of marker D3S196. Figure 1, A and B, represents separate analyses of a nuclear family with this marker. GENOTYPER's default filtering algorithm identified true alleles in experiment A but identified + A-modified alleles in all but one case in experiment B. This demonstrates the variability that can be generated both between experiments and within a single experiment as a result of the + A problem. In total, marker D3S196 (0.31 mean + A fraction) was labeled as the true product at 157 of 170 alleles and as the + A product at 13 of 170 alleles. D3S1238 (0.60 mean + A fraction) was labeled as the true product at 12 of 168 alleles and as the + A product at 156 of 168 alleles. Of the remaining 11 markers of the panel, none of them had a mean + A fraction within the range bounded by 0.31 to

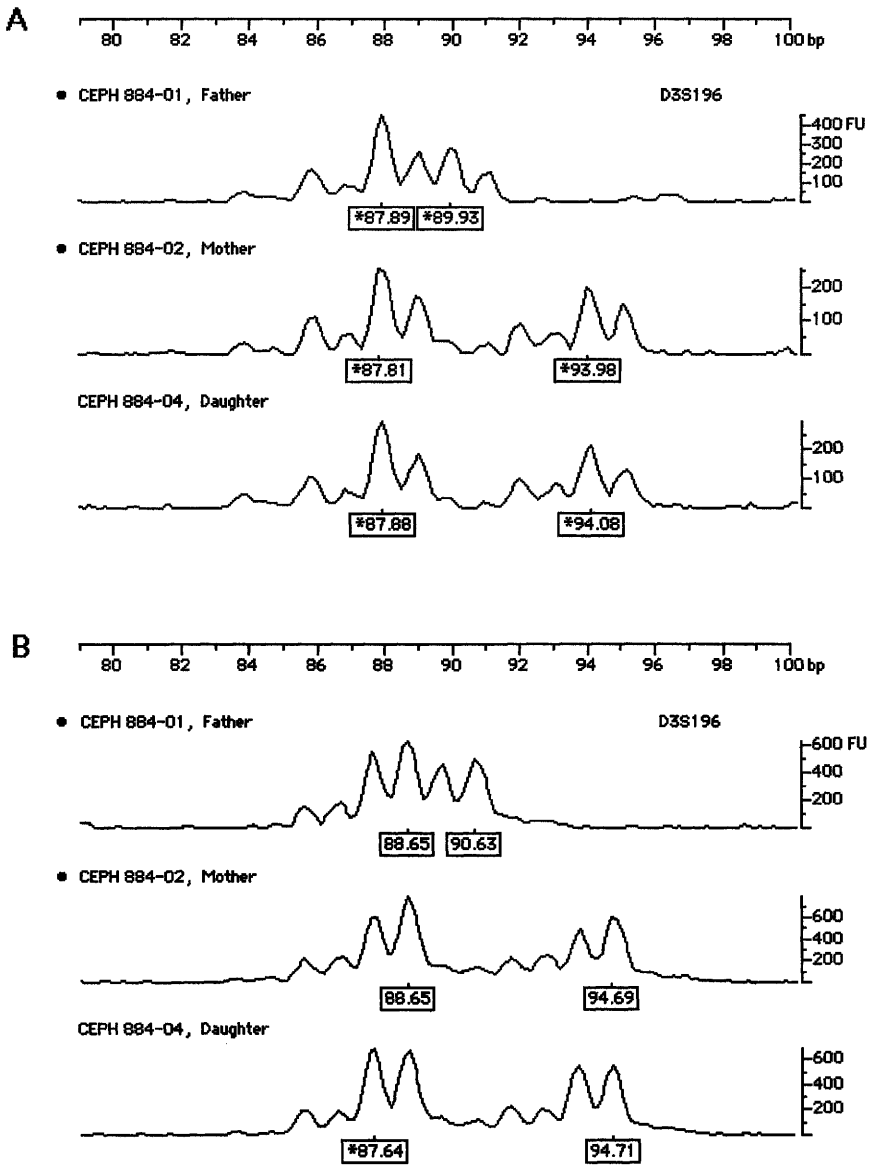

FIGURE 1 Electrophoretogram results generated by GENOTYPER representing alleles from three individuals of a nuclear family (CEPH 884) amplified with the marker D3S196. For Figs. 1–3, the *x*-axis represents size in base pairs and the *y*-axis represents peak heights in fluorescence units. Allele sizes assigned by GENOTYPER are boxed and shown below their respective peaks. Sizes marked by an asterisk (*) indicate true (T) alleles; sizes not marked by an asterisk (*) indicate alleles modified by nontemplated nucleotide addition (T + 1). (*A*,*B*) Replicate, but independent, experiments employing separate lots of AmpliTaq, $MgCl_2$, Perkin-Elmer PCR buffer II, DNA template preparations, and dNTPs.

0.60. Those with a mean + A fraction < 0.31 were consistently labeled as the true product, whereas those with a mean + A fraction > 0.60 were consistently labeled as the + A product. The cumulative error rate resulting from inconsistent identification of either the true or + A-modified allele for this panel of 13 markers was 3%.

In the second line of experiments to evaluate error rate resulting from + A modification, a small number of DNA samples were amplified once with a larger group of markers. We evaluated four panels of dinucleotide repeat markers using eight members of CEPH family 884 as genomic DNA sources, amplified with the ABI PRISM thermocycling protocol. These panels corresponded to test versions of ABI PRISM panels 6, 9, 20, and the marketed version of 21. The genotyping error rate was 1.2% (9 of 736 alleles) because of variability in the degree of + A modification of 5 of the 46 markers. Again, the markers for which errors were attributable to + A modification fell between 0.31 and 0.60 mean + A fraction.

We inferred that a dinucleotide repeat marker amplified with *Taq* DNA polymerase and modified to a mean + A fraction near 0.5 would present the greatest potential for error. Two experiments were conducted to assess the scope of error for such a worst-case marker. We first evaluated the error rate for D16S520 (ABI PRISM panel 21), a marker with a mean + A fraction of ~0.5, amplified by each of eight PCR reagent mixes (different lots of AmpliTaq, dNTP, $MgCl_2$, Perkin-Elmer buffer II, and sterile water) using 12 members of CEPH family 884 as genomic DNA sources. These 96 reactions were amplified using the ABI PRISM thermocycling protocol in eight separate Perkin-Elmer 9600 thermocyclers and electrophoresed on three gels. Our intention was to mimic experimental variability that might be encountered in a large-scale genotyping project. Markers D16S401 and D4S398 were analyzed in parallel as controls. As shown in the upper portion of each frame in Figure 2, D16S401 (0.84 mean + A fraction) was uniformly called as the + A-modified product. D4S398 (0.32 mean + A fraction) was uniformly called as the true allele. However, GENOTYPER's default filtering algorithm labeled the + A-modified peak of D16S520 (0.46 mean + A fraction, S.D. of 0.104) at 37% of the alleles and the true allele peak at 63% of the alleles.

These three markers were then reassessed in another experiment designed to minimize PCR reagent variability as a factor in the degree of + A modification. Ninety-six reaction replicates were aliquoted from a single PCR mix for each marker that included DNA from CEPH 884-01. The reactions were again amplified using the ABI PRISM thermocycling protocol on eight Perkin-Elmer 9600s and electrophoresed on three gels. D16S401 (0.86 mean + A fraction) was consistently called as the + A-modified product, whereas D4S398 (0.28 mean + A fraction) was consistently called as the true allele. Despite the uniformity of PCR reaction setup, D16S520 (0.53 mean + A fraction, S.D. 0.016) remained inconsistently labeled, with 5% true allele calls and 95% + A-modified calls.

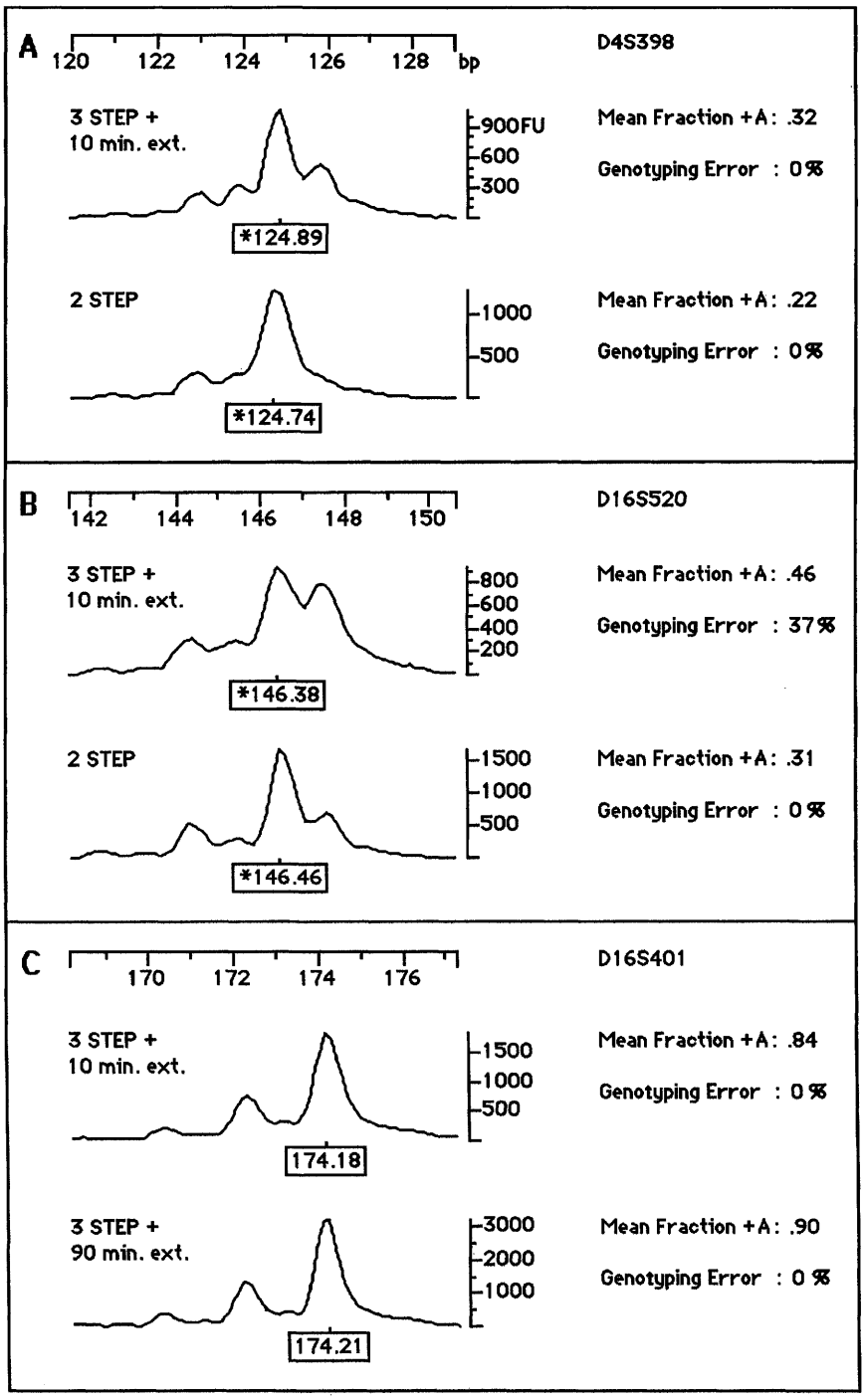

FIGURE 2 Alleles of three different markers (2A: D4S398, 2B: D16S520, and 2C: D16S401) amplified with the ABI PCR thermocycling protocol (three-step/10-min final extension) and either the two-step protocol or the three-step/90-min final extension protocol. Each marker was analyzed 96 times per thermocycling protocol. These employed multiple lots of AmpliTaq, $MgCl_2$, Perkin-Elmer PCR buffer II, and dNTPs to assay each of 12 members of CEPH family 884. Information summarizing the mean + A fraction and the percent error in allele identification by GENOTYPER is denoted at *right* of each respective electrophoretogram.

Because genotyping errors were associated with partial + A modification, we sought PCR methods to manipulate this activity. Several factors were identified that affected the degree of + A modification: time at 72°C with each cycle or during a final extension, time at room temperature after amplification, and reaction magnesium concentration. The ABI PRISM panels utilize *Taq* DNA polymerase in a single PCR thermocycling protocol and vary magnesium concentration to optimize amplification for each primer pair. Although increasing magnesium tends to increase the fraction of + A-modified product, nonspecific priming limits its usefulness in manipulating + A. Instead, we have focused on development of two thermocycling protocols that may be applied to dinucleotide repeat markers to avoid problematic partial modification.

The first thermocycling protocol cycles between denaturing and annealing without a dedicated extension step, and amplification is not followed by a final extension at 72°C. This two-step protocol minimizes the degree of nontemplated nucleotide addition by *Taq* DNA polymerase and typically generates somewhat more product than the three-step protocols. The second thermocycling protocol lengthens the postamplification final extension period of the ABI protocol to 90 min, which maximizes *Taq* DNA polymerase's ability to catalyze nontemplated nucleotide addition. Figure 3 presents comparative data for D16S520 amplified from CEPH 884-01 using the two-step protocol, the standard three-step ABI protocol, and a series of three-step protocols of increasing final extension times. These samples were aliquoted from a single PCR reagent mix that included the DNA template and were amplified concurrently in six separate Perkin-Elmer 9600s using these protocols. Samples were maintained at 4°C until electrophoresis on a single gel for analysis. Although the degree of + A modification for marker D16S520 appears complete after a 60-min final extension period, an additional 30 min is required by other markers (such as D4S398) for completion. Some markers (such as D16S401) are nearly completely modified to + A products under the standard ABI protocol and even remain + A modified under the two-step protocol.

We performed a survey to estimate the mean + A fraction for each of 71 dinucleotide repeat markers amplified from three to eight DNAs of CEPH family 884. The survey was done using the ABI thermocycling protocol and repeated for the two- and three-step/90-min final extension protocols. Figure 4A is a frequency histogram illustrating the

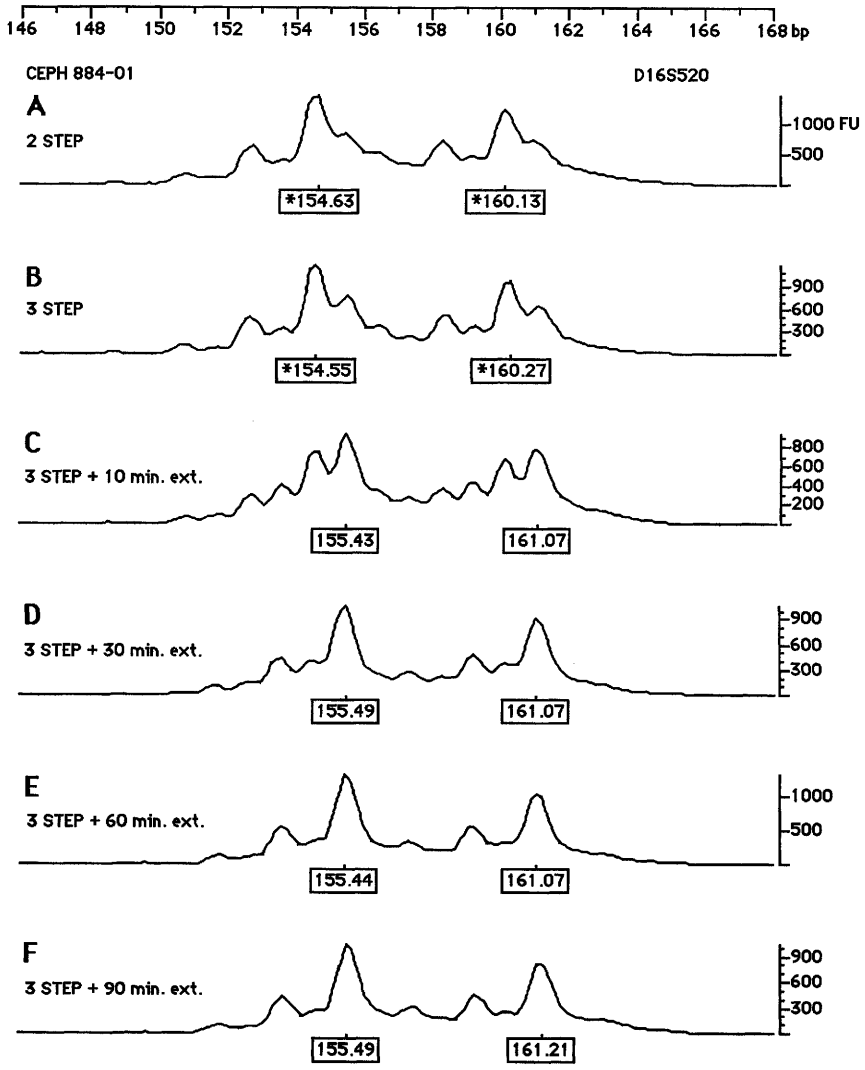

FIGURE 3 Electrophoretogram results generated by GENOTYPER displaying alleles from a single CEPH control individual (884-01) amplified with the marker D16S520 using different PCR cycling programs. Electrophoretograms in *A–F* correspond to different PCR thermocycling protocols as described in Materials and Methods.

distribution of mean + A fractions under the ABI protocol for these markers. Sequence alignment of the unlabeled primers for markers either strongly or weakly modified by nontemplated nucleotide addition failed to reveal any simple homology. Figure 4B shows the distributions under the two-step and three-step/90-min final extension protocols.

When each marker is individually optimized for the fraction of + A product using either of these thermocycling protocols, the error rate of inconsistent allele sizing resulting from + A modification can be significantly reduced. We repeated the test of D16S520 as a worst-case marker to estimate error rate attributable to the + A modification under the two-step thermocycling protocol. D4S398 and D16S401 were again tested in parallel as controls, the former under the two-step protocol and the latter under the three-step/90-min final extension protocol. Each marker was amplified by eight separate PCR reagent mixes (different lots of AmpliTaq, dNTP, $MgCl_2$, Perkin-Elmer buffer II, and sterile water) across 12 CEPH family 884 individuals. These reactions were set up in parallel with the test described above using the ABI thermocycling protocol. The 96 reactions for each marker were amplified in eight Perkin-Elmer 9600 thermocyclers and electrophoresed on three gels. Using dual thermocycling protocols, GENOTYPER was able to label alleles of each marker consistently. As illustrated in Figure 2B, alleles of D16S520 were uniformly labeled as the true allele. D4S398 (Fig. 2A) and D16S401 (Fig. 2C) were also uniformly labeled, the former as the true allele and the latter as the + A-modified allele.

In our experience, markers with a mean + A fraction between 30% and 65% present a potential source of genotyping error. Approximately a quarter of all dinucleotide repeat markers amplified with *Taq* DNA polymerase under the standard ABI PRISM thermocycling protocol fall within this range. The error frequency seems worst for the subset of these markers modified to a mean + A fraction near 0.5. However, if such a marker were consistently modified to the same fraction of + A, no genotyping error would occur. Experimental variability contributes greatly to the introduction of error when such a marker is used.

The actual genotype error rate of the small number of samples examined here by the ABI PRISM protocol was quantitated as 1%–3%. By dividing each panel of dinucleotide repeat markers judiciously into two sets, one amplified by the two-step protocol and one amplified by the three-step/90-min final extension protocol, the potential for genotyping error attributable to the + A problem can be greatly reduced.

An alternative to this dual PCR protocol strategy might be the use of a thermostable DNA polymerase lacking + A activity. A mutant version of *Taq* DNA polymerase lacking such activity is currently unavailable. *Pfu* DNA polymerase has very little + A activity and might be a preferable alternative.[4] However, for some markers we have encountered examples of a stutter peak (T-2) of higher signal than the true allele peak (T) after amplification by *Pfu* DNA polymerase (data not shown); GENOTYPER identifies the stutter peak as the allele in these cases. The potential of this property of *Pfu* DNA polymerase for error in large-scale genotyping studies remains to be investigated. Significant effort is required to construct a panel of markers that may be amplified and coelectrophoresed unambiguously. In our experience, panels of markers optimized with *Taq* DNA polymerase may not easily be converted for use with *Pfu* DNA polymerase. Methods employing an addi-

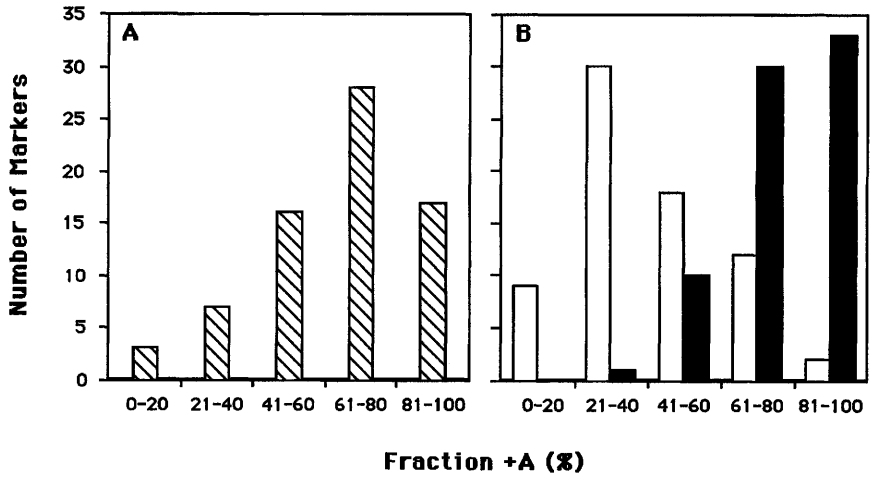

FIGURE 4 Frequency histogram representing distribution of mean + A fractions of 71 markers. Each marker was studied for three to eight CEPH 884 family members. The mean ratio of + A peak height (h_{T+1}) to total product ($h_T + h_{T+1}$) peak height for each marker was plotted. Number of markers (frequency of occurrence) is represented on the *y*-axis and the fraction of + A by percent interval is depicted along the *x*-axis. (*A*) The distribution of markers by mean fraction + A using the ABI PCR thermocycling protocol (hatched bars). (*B*) The distribution of markers by mean fraction + A using either the two-step protocol (open bars) or the three-step/90-min final extension protocol (solid bars).

tional enzymatic step using T4 DNA polymerase to remove the + A modification have also been proposed;[5] the additional sample manipulations required make this less amenable to high throughput genotyping.

Another strategy could involve improvement of software used for allele identification, enabling reliable distinction of true and modified alleles. Pattern recognition algorithms have been proposed,[6] but they must be able to cope with the difficult problem of pattern variation with changes in the degree of + A addition in two samples of the same allele. Allele identification algorithms must also differentiate the + A modification from infrequent but real single-base differences in alleles seen with imperfect dinucleotide repeat markers (e.g., D15S127).

Substitution of tri- or tetranucleotide STRs for dinucleotide STRs in genetic linkage studies will probably lessen error rate attributable to nontemplated nucleotide addition by allowing widened bin use. However, the larger allele size ranges of these markers allow fewer markers to be multiplexed per gel. Moreover, the prominent stutter pattern of dinucleotide repeats, minimal for tetranucleotide repeats, is an aid in distinguishing alleles from noise peaks. The higher density of dinucleotide repeats identified throughout the genome and greater number scorable per gel thus provide an impetus to develop further methods for their optimized and automated use.

ACKNOWLEDGMENTS

We would like to thank Adam Lowe for helpful discussions and Delphine Ally, Zarir Karanjawala, and Joseph Rayman for excellent technical support.

The publication costs of this article were defrayed in part by payment of page charges. This article must therefore be hereby marked "advertisement" in accordance with 18 USC section 1734 solely to indicate this fact.

REFERENCES

1. Ziegle, J.S., Y. Su, K.P. Corcoran, L. Nie, P.E. Mayrand, L.B. Hoff, L.J. McBride, M.N. Kronick, and S.R. Diehl. 1992. Application of automated DNA sizing technology for genotyping microsatellite loci. *Genomics* **14:** 1026–1031.
2. Reed, P.W., J.L. Davies, J.B. Copeman, S.T. Bennett, S.M. Palmer, L.E. Pritchard, S.C.L. Gough, Y. Kawaguchi, H.J. Cordell, K.M. Balfour, S.C. Jenkins, E.E. Powell, A. Vignal, and J.A. Todd. 1994. Chromosome-specific microsatellite sets for fluorescence-based, semi-automated genome mapping. *Nature Genet.* **7:** 390–395.
3. Clark, J.M. 1988. Novel non-templated nucleotide addition reactions catalyzed by procaryotic and eucaryotic DNA polymerases. *Nucleic Acids Res.* **16:** 9677–9686.
4. Hu, G. 1993. DNA polymerase-catalyzed addition of nontemplated extra nucleotides to the 3′ end of a DNA fragment. *DNA Cell Biol.* **12:** 763–770.
5. Kimpton, C.P., P. Gill, A. Walton, A. Urquhart, E.S. Millican, and M. Adams. 1993. Automated DNA profiling employing multiplex amplification of short tandem repeat loci. *PCR Methods Applic.* **3:** 13–22.
6. Perlin, M.W., M.B. Burks, R.C. Hoop, and E.P. Hoffman. 1994. Toward fully automated genotyping: Allele assignment, pedigree construction, phase determination, and recombination detection in Duchenne muscular dystrophy. *Am. J. Hum. Genet.* **55:** 777–787.

Received July 18, 1995; accepted in revised form September 27, 1995.