Nucleic Acids Research, 1997, Vol. 25, No. 14 2745–2751

PolyPhred: automating the detection and genotyping of single nucleotide substitutions using fluorescence-based resequencing

Deborah A. Nickerson*, Vincent O. Tobe and Scott L. Taylor

Department of Molecular Biotechnology, Box 357730, University of Washington, Seattle, WA 98195-7730, USA

Received April 22, 1997; Revised and Accepted May 27, 1997

ABSTRACT

Fluorescence-based sequencing is playing an increasingly important role in efforts to identify DNA polymorphisms and mutations of biological and medical interest. The application of this technology in generating the reference sequence of simple and complex genomes is also driving the development of new computer programs to automate base calling (Phred), sequence assembly (Phrap) and sequence assembly editing (Consed) in high throughput settings. In this report we describe a new computer program known as PolyPhred that automatically detects the presence of heterozygous single nucleotide substitutions by fluorescence-based sequencing of PCR products. Its operations are integrated with the use of the Phred, Phrap and Consed programs and together these tools generate a high throughput system for detecting DNA polymorphisms and mutations by large scale fluorescence-based resequencing. Analysis of sequences containing known DNA variants demonstrates that the accuracy of PolyPhred with single pass data is >99% when the sequences are generated with fluorescent dye-labeled primers and ~90% for those prepared with dye-labeled terminators.

INTRODUCTION

Single base substitutions are the most frequent form of DNA sequence variation in the human genome (1,2). Identification of these variations plays an important role in detailing the evolutionary history of human populations (3,4) and in exploring the relationships between genome structure and function (genotype–phenotype correlations) through genetic and disequilibrium mapping (5,6). Furthermore, many diagnostic applications depend on accurate identification of single nucleotide substitutions for finding mutated genes (7,8), matching tissues prior to transplantation (9) and analyzing samples in forensic situations (10).

Amplification of specific genomic regions using PCR has greatly simplified the process of comparing sequences to identify DNA variations by eliminating the need for genomic cloning from multiple individuals (11). Once a region has been amplified, a number of techniques can be employed to comparatively scan it for sequence variants. These include denaturing gradient gel electrophoresis (12), chemical or enzymatic cleavage (13–15), heteroduplex analysis (16), the analysis of single-stranded DNA conformations (17), hybridization to oligonucleotide arrays (18,19) and DNA sequencing (20–22). Among these approaches, DNA sequencing offers several advantages, including its ease of application (use of a single set of reagents and assay conditions), its automation with fluorescence-based methods and its ability to provide complete information about the location and nature of the sequence variant(s) in a single pass.

Despite the advantages of detecting DNA variations by sequence analysis, it is difficult to accurately identify heterozygous sites (two bases at the same location in a sequence) because of the variability in fluorescence signals and the inconsistency of base calling at these sites. Recently, several approaches have been taken to improve identification of heterozygous sites using automated sequence analysis (22–25). In one approach, heterozygotes are found by comparing the pattern of fluorescence dye incorporation between the sequence traces (22). Since this pattern is faithfully reproduced every time the same sequence is generated, heterozygous positions in a trace can be accurately identified based on a predictable reduction (~50%) in normalized peak area when compared with homozygous positions. In this report we present a computer program known as PolyPhred that automatically finds potential heterozygotes in a sequence using this comparative approach. We also compare program performance with sequencing chemistries that produce highly variable patterns of dye incorporation (sequences generated with dye-labeled terminators; 26) and those that produce more uniform fluorescence incorporation (dye-labeled primer sequencing) in terms of their accuracy and efficiency in heterozgyote detection. Lastly, we report the discovery of new DNA variations using comparative sequencing and PolyPhred.

MATERIALS AND METHODS

PCR primers

Primers for PCR amplification of genomic DNA were assembled using standard phosphoramidite chemistry on an Applied Biosystems 394 DNA synthesizer (Foster City, CA). Primers were prepared and used to amplify 11 genomic regions containing single nucleotide substitutions representing all potential nucleotide

*To whom correspondence should be addressed. Tel: +1 206 685 7387; Fax: +1 206 685 7301; Email: debnick@u.washington.edu

changes (A↔T, A↔C, G↔T, C↔G, A↔G, C↔T). The regions examined were: (i) exon 2 of the human steroid 5α-reductase gene (SRD5A1, A↔G, GDB:193189, CCCAAATCATTTAAGATAGGATTAC, ATGATGTGAACAAGGCGGAGTTCAC, 60°C); (ii) intron 8 of the human lipoprotein lipase gene (LPL, A↔C, GDB:191079, TACACTAGCAATGTCTAGCTGA, TCAGCTTTAGCCCAGAATGC, 60°C); (iii) exon 28 of the von Willebrand factor pseudogene (VWFP, A↔T, GDB:194282, TGTAAAACGACGGCCAGT(–21M13)AGCCGTCGTGGTACTCCACCACA, CAGGAAACAGCTATGACC(M13Rev)AGATTCTGTGGGAATATGGAAGTAGTCA, 55°C); (iv) exon 5 of the guanine nucleotide binding protein (GNAS, A↔G, GDB:203981, TCTTGTAGCGCCCTCCCA, TGCCCATGTGCAGGGCTGTCACTCATGTT, 60°C); (v) a segment from the 3′-untranslated region of β2-integrin (ITGB2, C↔T, GDB:185175, GAGCACTTGGTGAAGACAAG, GGATGTCATTTTATACCCTG, 51°C); (vi) intron 3 of adenine nucleotide translocator 1 (ANT1, C↔T, GDB:201792, ACAGGGCTCCTTTCAGTCTTCC, CAAATGCTGGTGAGGGCTCCG, 57°C); (vii) exon 4a of solute carrier family 2, member 4 (SLC2A4, C↔T, GDB:180271, CAGGAAGGGAGCCACTGCTG, ATCTGAAAGCCCAGGCATGG, 63°C); (viii) a segment from the 3′-untranslated region of the tyrosinase-related protein 1 gene (TYRP1, A↔C, GDB:555709, GTCGGGAGTTTAGTGTACCT, TCTGAAAGGGTCTTCCCAGC, 60°C); (ix) intron 4 of the constant region of the human T cell receptor (TCR) α locus (TCRCA, C↔T, G↔T and C↔G, TGTAAAACGACGGCCAGT(–21M13)GAGCTAAGAGAGCCGTACTGG, CAGGAAACAGCTATGACC(M13Rev)CTTGAAGCTGGGAGTGG, 55°C) (27); (x) a variable gene segment from the human TCR α locus (TCRVA23, C↔T and C↔G, TGTAAAACGACGGCCAGT(–21M13)GTCTAAGTGACAGAAGGAATG, AATGTATAAAGTACTACGTCCTGA, 55°C) (28); (xi) a variable gene segment from the human TCR β locus (TCRVB23, A↔G and G↔T, GenBank accession no. U96844, TGTAAAACGACGGCCAGT(–21M13)GGAAAGCCTGAGTTAGCTGAGC, CAGGAAACAGCTATGACC(M13Rev)AGAATAGAAGCATCTCTGGG, 55°C).

DNA amplification

DNA samples from the parents of the 40 families available through the Centre d'Etude du Polymorphisme Humaine (CEPH) were used for PCR amplification of the target loci. All amplification reactions were performed in a 96-well microtiter plate thermal cycler (PTC 100; MJ Research, Watertown, MA). The reactions were assembled (20 µl total volume) and contained a standard PCR buffer (10 mM Tris–HCl, pH 8.3, 50 mM KCl, 1.5 mM $MgCl_2$ and 0.001% gelatin), the four deoxynucleotide triphosphates at 40 µM each, 0.5 µM each primer, 0.5 U Taq polymerase (Perkin Elmer Cetus, Norwalk, CT) and 20 ng genomic DNA. Following assembly, the reactions were covered with 50 µl mineral oil. Thermal cycling was performed with an initial denaturation at 94°C for 1 min followed by 35 cycles of denaturation at 95°C for 20 s, primer annealing for 30 s (temperatures specified above with primer sequences) and primer extension at 72°C for 2 min. After 35 cycles, a final extension was carried out at 72°C for 5 min. Individuals were selected for DNA amplification based on genotypes previously established in these loci using PCR combined with an oligonucleotide ligation assay (OLA; 29).

DNA sequencing

Following DNA amplification, unincorporated PCR primers and deoxynucleotide triphosphates in the samples were inactivated prior to sequencing by enzymatic treatment. This was accomplished by mixing 6 µl PCR product with 1 µl exonuclease I (10 U/µl; Amersham Life Science Inc., Arlington Heights, IL) and 1 µl shrimp alkaline phosphatase (2 U/µl; Amersham) and incubating at 37°C for 15 min followed by 80°C for 15 min to inactivate the exonuclease and alkaline phosphatase enzymes prior to sequencing. In our hands PCR products treated with these enzymes sequence as well in terms of quality and read length as those isolated by agarose gel electrophoresis coupled with column purification (26). Cycle sequencing was performed according to the manufacturer's instructions using ABI PRISM Dye Terminator or Dye Primer Sequencing Kits with Amplitaq DNA polymerase FS (Perkin Elmer Corp., Foster City, CA). For dye-terminator cycle sequencing the entire enzyme-treated PCR sample (8 µl total following treatment) was used as the sequencing template. The sequencing primer (3.2 pmol, same as PCR primer) and 8 µl Dye Terminator Ready-Reaction sequencing premix were added to the template. Following a denaturation step at 96°C for 2 min, dye-terminator reactions were incubated at 96°C for 15 s, 50°C for 1 s and 60°C for 4 min for 25 cycles. Excess dye-terminators were removed by ethanol precipitation. In the case of dye-primer sequencing, PCR products were generated using locus-specific primers containing either the –21M13 or M13Rev primer sequences at their 5′-end. For sequencing the enzyme-treated PCR sample was subdivided into four separate reactions as follows: 1 µl each of the PCR sample mixed with 4 µl PRISM ready premix for the A and C reactions and 2 µl each of the PCR sample mixed with 8 µl PRISM ready premix for the G and T reactions. Sequencing reactions were denatured for 1 min at 96°C and subjected to 15 cycles at 96°C for 10 s, 55°C for 5 s and 70°C for 1 min and 15 cyles at 96°C for 10 s and 70°C for 1 min. Then, the A, C, G and T reactions were pooled and subjected to ethanol precipitation. The extension products obtained with either chemistry were evaporated to dryness under pressure (Savant Instruments, Farmingdale, NY), resuspended in 3 µl 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 373 sequencer according to the manufacturer's directions.

Sequence analysis

The ABI sequence software (version 2.1.2) was used for lane tracking and first pass base calling (Perkin Elmer). Chromatograms were transferred to a Unix workstation (Sun Microsystems Inc., Mountain View, CA), base called with Phred (version 0.961028), assembled with Phrap (version 0.960731), scanned by PolyPhred (version 0.970312) and the results viewed with the Consed program (version 4.0). Specific descriptions and documentation on Phred, Phrap and Consed are available at http://www.genome.washington.edu (P.Green, personal communication). PolyPhred has been designed to parse information from Phred and Phrap output files and via a flat file provide input to Consed to aid in identification of heterozygous single nucleotide substitutions by color coding potential sites. All data presented in this report were generated using command line parameters requiring a peak drop ratio of 0.55, a second peak ratio of 0.15 and, unless noted otherwise, an average sequence quality setting of 30. PolyPhred

Figure 1. Two windows from the 'color means quality and tags view' of the Consed program. Sequences from two TCR β sites (in TCRVB23) obtained by dye-primer sequencing of PCR products from 10 different individuals are shown in each window. In Consed, sequence quality measured by Phred is depicted using a gray scale with a white background indicating the highest quality bases and increasing shades of gray indicating decreasing data quality. Potential heterozygotes identified by PolyPhred are color coded in blue in this Consed view. (**a**) When a common DNA polymorphism is identified (e.g. position 214) homozygotes for each of the alternative alleles and heterozygotes (highlighted blue) are normally detected. (**b**) Less common DNA variants usually appear as rare heterozygotes among a background of homozygotes (e.g. position 351). There are no false positives or false negatives identified in these windows and the genotypes of these individuals were completely consistent with those obtained by PCR/OLA.

is available via Email from debnick@u.washington.edu and more documentation is available at http://droog.mbt.washington.edu.

RESULTS

In comparing sequence traces of homozygotes with those for heterozygotes two changes are usually present: (i) a significant drop in normalized peak height at a polymorphic site when traces from homozygous and heterozygous individuals are compared; (ii) a second underlying peak at the position in question (22,26). To automate identification of substitution variations using these criteria, we created a program known as PolyPhred. Its functions are fully integrated with three software packages currently applied in large scale sequence analysis: Phred, Phrap and Consed (P.Green, B.Ewing and D.Gordon, personal communication). PolyPhred reads the normalized peak areas and quality values obtained by Phred for each position in a sequence. It then searches for reductions in peak areas at each position across the sequence alignment obtained from the Phrap assembly program. If the required drop in peak area is found at a position and a second base is detected by Phred, PolyPhred calls the site a potential heterozygote and information on this position is stored in the program's output file.

By interfacing the information obtained by PolyPhred with the 'quality means color and tags view' in the Consed program, potential heterozygotes become color coded, as shown in Figure 1 (position 214, Fig. 1a, and position 351, Fig. 1b). In these examples Consed views of sequences from 10 individuals are shown. Heterozygotes at two positions in the coding region of a TCR gene were automatically identified by PolyPhred (91 bp of assembly sequence are shown in each window and, altogether for the 10 individuals, 1820 bp of sequence are displayed). When a common polymorphism is identified (position 214 in Fig. 1a, His→Arg substitution) homozygotes for each of the alternative alleles are detected (e.g. in this instance four individuals homozygous G and one individual homozygous A), in addition to heterozygotes containing the two alternative alleles (e.g. sequences from the five heterozygous individuals color coded blue). However, less common alleles will typically be identified just as heterozygotes among the homozygotes (e.g. the three heterozygotes color coded blue at position 351 in Fig. 1b). It is worth noting that the variant at position 351 (Val→Gly) would have been missed if identification was based solely on the results of sequence alignment, since neither the ABI nor the Phred program called these positions as Ns: in all cases the G peak was sufficiently dominant even in a heterozygote to meet the ABI and Phred criteria for a G.

Table 1. Sequence quality and PolyPhred performance

Chemistry	PolyPhred quality threshold	Average quality of data analyzed	True[a] positives (TP)	False positives (FP)	Total bases scanned	Ratio[b] TP/FP
Dye-primer	20	36	147	1545	61 947	1:11
	25	37	147	811	57 915	1:6
	30	38	146	381	53 472	1:3
	35	40	143	179	47 517	1:1
	40	42	107	54	34 419	2:1
Dye-terminator	20	29	202	2637	81 479	1:13
	25	31	197	1662	73 750	1:8
	30	32	173	768	62 772	1:4
	35	35	104	245	42 091	1:2
	40	39	28	31	16 711	1:1

[a]Includes variations previously identified in target loci and typed by PCR/OLA in addition to new variations detected and confirmed by sample resequencing.
[b]The ratio of true positives to false positives rounded to the nearest whole number.

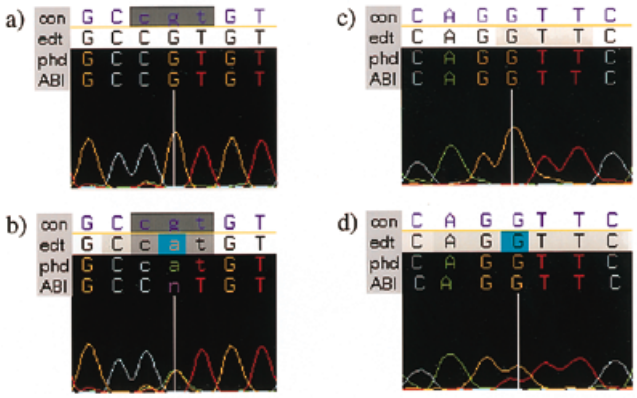

Figure 2. Examples of two polymorphic sites identified by PolyPhred in TCRVB23; position 214 (**a** and **b**) and position 351 (**c** and **d**). An example of a homozygote (a and c) and heterozygote (b and d) for each of these sites has been isolated from a trace editor window opened by the Consed program. Base calls for the consensus (con), edited (edt, where PolyPhred color codes heterozygotes), the Phred (phd) and ABI (ABI) programs are shown. When traces are opened in Consed, the sequence position used to open the file is indicated by the vertical white line. Note that in the heterozygotes (b and d) there is a significant drop in peak height in the Gs and a second peak is found in the heterozygotes (b, G/A; d, G/T), but not in the homozgyotes (a and c).

Once potential heterozygotes are tagged in Consed, the traces can be viewed by the analyst for editing or evaluation purposes. Examples of homozygous and heterozygous sequencing traces taken from Consed are shown in Figure 2. When comparing homozygous and heterozygous sequencing traces (as in Fig. 2), a characteristic drop in peak height (area) is obvious and a significant signal for a second base is present in the heterozygous sequences. With sequencing chemistries that produce more uniform fluorescence peaks (i.e. dye-labeled primer sequencing) the two peaks in a heterozygote are usually similar in size and are frequently called Ns. However, many heterozygous sites (~35%) will still be missed and called as homozygotes because of the peak disparity between the bases (position 351, Fig. 2d). This problem is even greater for chemistries that give rise to more uneven fluorescence incorporation (26). For example in dye-terminator sequencing >70% of the heterozygotes are called as homozygotes.

One of the key features of the Phred/Phrap/Consed environment is that data quality is monitored and displayed as an integral part of the system. By operating in this environment we have found that there is a clear relationship between sequence quality (determined by Phred) and PolyPhred performance (Table 1). Three factors are used to generate quality measures in Phred; these include peak spacing, the relative size of the uncalled and called peaks and the dip in signal between called peaks (B.Ewing and P.Green, personal communication). As expected when sequence quality is low (Phred quality = 20), the signal-to-noise ratio as measured by the ratio of PolyPhred true positives (confirmed by PCR/OLA, 92%, or by sample resequencing, 8%) to false positives is low (Table 1). With PCR products <700 bp in length this quality setting allows nearly complete scanning of all the bases in each trace, including the low quality bases at the start and end of the sequence. However, even at this modest quality level, the total positives identified by PolyPhred are <3% of the total base pairs examined.

The signal-to-noise ratio (true positive to false positives) improves greatly as the scanning window for PolyPhred is set to analyze data at increasing quality thresholds. The total number of base pairs that can be scanned at higher quality settings differs significantly for the two chemistries examined. With dye-primer sequencing and a quality setting of 40, an average of 250 bp are scanned in 94% of the available sequence traces (143 traces altogether) and the signal-to-noise ratio is 2:1 (Table 1). At higher qualities, false positives are easily distinguished by an analyst and are usually caused by a fluctuation in peak height from a homozgyote combined with some increase in the background noise that led PolyPhred to detect a second base. Examples of these types of false positives are shown in Figure 3 for sequences produced with dye-labeled primer (Fig. 3a and b) or dye-labeled terminator (Fig. 3c and d) sequencing.

In terms of PolyPhred's performance, sequencing with dye-labeled primers has many advantages over sequencing with dye-labeled terminators (Table 1). Across the range of sequence quality the ratios of signal-to-noise for sequences generated with dye-labeled primers are nearly twice those observed with dye-labeled terminator sequencing (Table 1). The incorporation of dye-labeled terminators is known to produce uneven peaks which will impact sequence quality by their effects on peak areas (26).

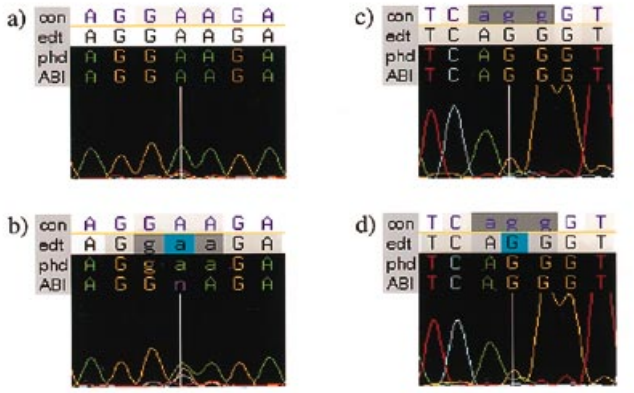

Figure 3. Examples of false positive calls made by Polyphred. In dye-primer sequencing, a correctly called non-variant site (**a**) compared with a false positive (**b**). This false positive is triggered by premature termination in the sequencing products and generates a low quality base in a high quality region. In dye-terminator sequencing, random peak fluctuations associated with small peak contexts such as Gs after As (**c**) and background noise will often trigger a false positive call (**d**) by PolyPhred.

Furthermore, the predictable sequence patterns that are produced also greatly influence the number of false positives produced by PolyPhred. In fact, >50% of the false positives identified by PolyPhred were associated with sequence contexts that produce small peaks with dye-labeled terminators (26). A single sequence context, the small G peaks that follow A peaks, was associated with >30% of all the false positives identified with this sequencing chemistry (Fig. 3d).

To measure accuracy in calling heterozygotes, we compared the genotypes determined by PolyPhred using a moderate quality level (30) with those previously determined by PCR/OLA. These results are shown in Table 2 and are broken down by sequencing chemistry and substitution composition. With single pass sequence data PolyPhred was able to achieve >99% accuracy in calling heterozyotes when sequences were generated using the dye-primer chemistry, i.e. 134 of 135 PolyPhred genotypes were correctly identified when compared with genotypes obtained by PCR/OLA. Upon examination of the data associated with the missed C/T heterozygote, we found that although there was an appropriate peak drop at this position, a second peak was not detected at the position by Phred. In our experience this is unusual for sequences generated with dye-labeled primers. With sequences generated with dye-labeled terminators, the overall accuracy of PolyPhred was nearly 90% (131 of 146 heterozygotes), but the errors were not evenly distributed among the different substitution types. Changes involving Cs and Gs were clearly more difficult to call. Approximately 35% of the heterozygotes of this composition were incorrectly called as homozygotes by PolyPhred. The other three substitutions (all C/T) were missed because the peak drop was not sufficient (reflecting fluctuations in dye incorporation).

In addition to the DNA variations typed and compared with OLA genotypes, nine new single nucleotide substitutions were identified by PolyPhred. These variants are summarized in Table 3 and all have been confirmed by resequencing. Although several of the variations (SRD5A1 and VWFP) were frequent, these results also document the program's sensitivity in terms of identifying single heterozygotes among homozygotes, e.g. three non-coding variants were detected in the TCRVA23 gene as single heterozygotes among 22 homozygotes (Table 3).

Table 2. Efficiency of heterozygote detection by PolyPhred

Chemistry	Base substitutions						Totals
	A/C	A/G	A/T	C/G	C/T	G/T	
Dye-primer							
PCR/OLA genotypes	18	14	17	26	36	24	135
PolyPhred calls							
True positive	18	14	17	26	35	24	134
True negative	0	0	0	0	0	0	0
False positive	0	0	0	0	0	0	0
False negative	0	0	0	0	1	0	1
Dye-terminator							
PCR/OLA genotypes	17	14	17	34	41	23	146
PolyPhred calls							
True positive	17	14	17	22	38	23	131
True negative	0	0	0	0	0	0	0
False positive	0	0	0	0	0	0	0
False negative	0	0	0	12	3	0	15

DISCUSSION

Resequencing of genes to identify DNA variations will play a major role in the post-genomics analysis of human biology and medicine (2–10). In this regard, high density oligonucleotide arrays are currently under testing for many types of resequencing projects (18,19). However, four-color fluorescence-based sequencing is currently a more mature technology capable of high throughput and with its increasing availability, decreasing cost and improving accuracy will likely be applied in many areas of genome resequencing. This is particularly true in those situations where little is known about the types and/or the distribution of the DNA variations within a sequence. Indeed, fluorescence-based sequencing is already frequently applied for diagnostic purposes (7,8,10,20, 21,24).

Programs that automate analysis of high throughput fluorescence-based sequencing, such as Phred, Phrap and Consed, are rapidly emerging and we have integrated the use of PolyPhred with these large scale software tools. With regard to base calling, one of the significant advantages of Phred is its ability to provide quality measures for each base in a sequence. For some applications, such as shotgun sequencing of M13 templates with dye-labeled primers, quality can be related to an objective estimate of base calling error rate (P.Green, unpublished observation). Similar correlations between quality and base calling error are not yet available for PCR product sequencing and it is important to consider developing these standards, particularly for diagnostic resequencing (10,19). In these applications quality control is essential, as is assay standardization. Impartial measures of quality from programs such as Phred can help in further automating data analysis for large scale resequencing, just as it does in generating the reference sequence [e.g. setting base (or even whole sequence trace) inclusion or exclusion standards and providing an estimate of error and accuracy for every base call in a sequence]. Furthermore, these standards could be used to monitor the efficiency or effectiveness of new protocols and in the development of new applications, such as PolyPhred, where sequence quality can be related to performance standards.

Table 3. New DNA variations detected by PolyPhred

Gene	Sequence context[a]	Reference[b] allele	Frequency[c]
TCRCA	TTAGGGACG(C/T)GGGTCTCTG	M94081	C 9%/T 91% (58)
LPL	CTGAACACC(A/G)GGTTAGGCT	M76722	A 9%/G 91% (54)
VWFP	CCTGGTGGT(A/G)CCTCCCACA	M60676	A 50%/G 50% (82)
SRD5A1	AATTTACCC(G/A)TTTCTGATG	M68883	G 50%/A 50% (32)
ITGB2	AGCCATGGC(C/A)GGCCGGGTG	X63926	C 80%/A 20% (40)
ANT1	TGAACCATA(T/C)GAAATTGCC	J04982	T 42%/C 58% (14)
TCRVA23	AATTTAAAG(G/A)TAATTTCTA	U32531	G 98%/A 2% (46)
TCRVA23	AAATGGAAA(T/A)GAGCAAAGA	U32531	T 98%/A 2% (46)
TCRVA23	CTGCAATGT(G/T)AGTTAGAGG	U32531	G 98%/T 2% (46)

[a]The alternative alleles are presented in parentheses. The first allele is the one reported in the reference sequence.
[b]GenBank and EMBL accession nos for the reference allele sequence.
[c]Allele frequency among the test panel with the number of chromosomes typed by PolyPhred in parentheses.

One of the major challenges in identifying DNA variations by fluorescence-based approaches is related to the detection of two bases at a single position within a sequence. This is a difficult problem since the signal levels in a heterozygous sample will be 50% of that obtained with a homozygous sample. However, this drop in signal intensity can be used to comparatively and accurately detect heterozygotes among homozygotes (22). We found that this feature alone, although sensitive, is not specific enough in heterozygote detection because of fluctations in the sequencing traces. To increase specificity we also require the presence of a second underlying peak. This additional requirement does increase the specificity of heterozygote detection, but slightly decreases its sensitivity. In this report the absence of a detectable second peak by Phred clearly led to some errors (false negatives) in heterozygote detection by PolyPhred using single pass sequences. Although new approaches to trace normalization may help with the initial identification of second peaks by Phred, we are also evaluating whether the use of additional information associated with changes in the local sequence context due to the presence of a sequence variation, i.e. 3′-base effect (22), can help in improving the accuracy of heterozygote identification.

Several different fluorescence-based chemistries can now be applied in diagnostic resequencing and new ones are under development (30,31). In principle, any fluorescence-based chemistry that can be analyzed using the Phred, Phrap and Consed programs could also be scanned for DNA variations using PolyPhred. In terms of program performance, however, more uniform incorporation of the fluorescence dyes does yield higher accuracies in identifying heterozygotes with single pass data, i.e. comparing dye-primer performance (more uniform peaks with >99% accuracy) versus dye-terminator performance (uneven peaks with ~90% accuracy). In diagnostic situations where accuracy is essential, e.g. mutation scanning for genetic diseases or sequencing major histocompatibility genes for tissue typing, it is clear that chemistries that give rise to more uniform dye incorporation should be applied to the genes of interest and sequencing of the opposite strand should also be considered. However, in cases where one is not interested in finding every variant but needs to quickly scan a region for additional genetic markers (32), then dye-terminator-based chemistries combined with PolyPhred can identify a large number of the variants present among individuals scanned (~90% at quality 30).

The current version of PolyPhred is capable of identifying single nucleotide substitutions and although these variations are the most frequent basis for disease-causing mutations, automating the identification and typing of insertions and deletions of one or more base pairs in a sequence will also be important (8,18). A trained analyst can easily detect and resolve these variations. However, the development of a computationally efficient and accurate system for calling insertion/deletion variations will be challenging and this is a focus of our current work with PolyPhred.

In summary, there are many approaches being applied to detect DNA polymorphisms and mutations (33). Among these, direct sequencing serves as the gold standard in terms of sensitivity and accuracy (33). This, combined with new tools like PolyPhred, is rapidly making automated fluorescence-based resequencing a sensible and cost effective approach for identifying DNA polymorphisms and mutations in many biological and medical applications.

ACKNOWLEDGEMENTS

We thank Drs Phil Green and Brent Ewing and Mr David Gordon for sharing their insights and programs (Phred, Phrap and Consed) with us. We also thank Drs Maynard Olson and Mark Rieder and Ms Ursula Petralia for their helpful comments. This work was supported in part by the National Science Foundation (DIR 8809710), the National Institute for Human Genome Research (HG01436) and the Department of Energy (DE-FG03-97ER-62385).

REFERENCES

1 Cooper,D.N., Smith,B.A., Cooke,H.J., Niemann,S. and Schmidtke,J. (1985) *Hum. Genet.*, **69**, 201–205.
2 Cooper,D.N. and Krawczak,M (1993) *Human Gene Mutation*. Bios Scientific, Oxford, UK.
3 Wallace,D.C. (1994) *Proc. Natl. Acad. Sci. USA*, **91**, 8739–8746.
4 Erlich,H.A., Bergstrom,T.F., Stoneking,M. and Gyllensten,U. (1996) *Science*, **274**, 1552–1554.

5 Trivier,E., DeCesare,D., Jacquot,S., Pannetier,S., Zackai,E., Young,I., Mandel,J.L., Sassone-Corsi,P. and Hanauer,A. (1996) *Nature*, **384**, 567–570.
6 DeKok,Y.J., van der Maarel,S.M., Bitner-Glindzicz,M., Huber,I., Monaco,A.P., Malcolm,S., Pembrey,M.E., Ropers,H.H. and Cremers,F.P. (1995) *Science*, **267**, 685–688.
7 Hedrum,A., Pont'en,F., Ren,Z., Lundeberg,J., Pont'en,J. and Uhl'en,M. (1994) *BioTechniques*, **17**, 118–119, 122–124, 126–129.
8 Shattuck-Eidens,D., McClure,M. and Simard,J. (1995) *J. Am. Med. Ass.*, **273**, 535–541.
9 Santamaria,P., Boyce-Jacino,M.T. and Lindstrom,A.L. (1992) *Hum. Immunol.*, **33**, 69–81.
10 Wilson,M.R., DiZinno,J,A., Polanskey,D., Replogle,J. and Budowle,B. (1995) *Int. J. Legal Med.*, **108**, 68–74.
11 Saiki,R.K., Gelfand,D., Stoffel,S., Scharf,S.J., Higuchi,R., Horn,G.T., Mullis,K.B. and Erlich,H.A. (1988) *Science*, **239**, 487–491.
12 Sheffield,V.C., Cox,D.R., Lerman,L.S. and Myers,R.M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 232–236.
13 Cotton,R.G., Rodrigues,N.R. and Campbell,R.D. (1988*) Proc. Natl. Acad. Sci. USA*, **85**, 4397–4401.
14 Myers,R.M., Larin,Z. and Maniatis,T. (1985) *Science*, **230**, 1242–1246.
15 Youil,R., Kemper,B.W. and Cotton,R.G. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 87–91.
16 Glavac,D. and Dean,M. (1995) *Hum. Mutat.*, **6**, 281–287.
17 Orita,M., Suzuki,Y., Sekiya,T. and Hayashi,K. (1989) *Genomics*, **5**, 874–879.
18 Hacia,J.G., Brody,L.C., Chee,M.S., Fodor,S.P.A. and Collins,F.S. (1996) *Nature Genet.*, **14**, 441–447.
19 Chee,M.S., Yang,R, Hubbell,E., Berno,A., Huang,X.C., Stern,D., Winkler,J., Lockhart,D.J., Morris,M.S. and Fodor,S.P. (1996) *Science*, **274**, 610–614.
20 Gibbs,R.A., Nguyen,P.N., McBride,L.J., Koepf,S.M. and Caskey,T.M. (1989) *Proc. Natl. Acad. Sci. USA*, **86**, 1919–1923.
21 Leren,T.P., Rodningen,O.K., Rosby,O., Solberg,K. and Berg,K. (1993) *BioTechniques*, **14**, 618–623.
22 Kwok,P.-Y., Carlson,C., Yager,T.D., Ankener,W. and Nickerson,D.A. (1994) *Genomics*, **23**, 138–144.
23 Phelps,R.S., Chadwick,R.B., Conrad,M.P., Kronick,M.N. and Kamb,A. (1995) *BioTechniques*, **19**, 984–989.
24 Versluis,L.F., Rozemuller,E., Tonks,S., Marsh,S.G., Bouwens,A.G.M., Bodmer,J.G. and Tilanus,M.G.J. (1993) *Hum. Immunol.*, **38**, 277–283.
25 Hattori,M., Shibata,A., Yoshioka,K. and Sakaki,Y. (1993) *Genomics*, **15**, 415–417.
26 Parker,L.T., Zakeri,H., Deng,Q., Kwok,P.-Y. and Nickerson,D.A. (1996) *BioTechniques*, **21**, 694–699.
27 Nickerson,D.A., Whitehurst,C., Boysen,C., Charmley,P., Kaiser,R. and Hood,L. (1992) *Genomics*, **12**, 377–387.
28 Boysen,C., Carlson,C., Hood,E., Hood,L. and Nickerson,D.A. (1996) *Immunogenetics*, **44**, 121–127.
29 Tobe,V.O., Taylor,S.L. and Nickerson,D.A. (1996) *Nucleic Acids Res.*, **24**, 3728–3732.
30 Ju,J., Ruan,C., Fuller,C.W., Glazer,A.N. and Mathies,R.A. (1995) *Proc. Natl. Acad. Sci. USA*, **92**, 4347–4351.
31 Metzker,M.L., Lu,J. and Gibbs,R.A. (1996) *Science*, **271**, 1420–1402.
32 Kwok,P.-Y., Deng,Q., Zakeri,H., Taylor,S.L. and Nickerson,D.A. (1996) *Genomics*, **31**, 123–126.
33 Grompe,M. (1993) *Nature Genet.*, **5**, 111–117.