SNPCEQer: Detecting SNPs in Sequences Generated by the Beckman CEQ™2000 DNA Analysis System

BioTechniques 33:814-822 (October 2002)

ABSTRACT

SNPCEQer identifies and reports SNPs in sequences obtained from the Beckman CEQ™2000 DNA Analysis System. SNPCEQer aligns sequences obtained using CEQ2000 heterozygote detection analysis and reports discrepancies between individual sequences and the consensus sequence it generates from this set as SNPs when the individual base calls have high-quality values. SNPCEQer reported comparable numbers of SNPs to the UNIX-based PolyPhred (148 vs. 165, respectively) in regions amplified from eight genes. A total of 21 different SNPs was discovered. Each gene region was analyzed in 96–306 samples. SNPCEQer was designed to operate from Windows NT®, making SNP detection more accessible to users without UNIX systems. SNPCEQer is available free of charge at http://innovation.swmed.edu.

INTRODUCTION

SNPs are the most frequently found form of DNA sequence variation in the human genome (2). These variations can be used in charting the evolutionary history of populations and in studying the relationship between genome structure and function. SNPs are also important, as some of them have a role in human disease. Much of the sequenced human genome must be further characterized with respect to SNPs (11,13).

The application of fluorescence-based sequencing technologies to the analysis of the human genome inspired the development of computer programs to facilitate the automation of base calling, sequence assembly, and assembly editing (4–6,14). Several workers developed programs to automate the detection of single nucleotide substitutions from sequence data derived using fluorescence-based Sanger sequencing technologies—PolyPhred, RelComp, and PolyBayes (8,9,12). At the time of this writing, these programs operated in UNIX operating systems. Commercially available packages that perform sequence alignment and discrepancy detection in Microsoft® Windows™ environments are Sequencher™ (GeneCodes, Ann Arbor, MI, USA) and Lasergene™ (DNASTar, Madison, WI, USA). However, neither of these programs grades the SNPs or “conflicts” detected according to quality value.

The CEQ™2000 system (Beckman Coulter, Fullerton, CA, USA) uses fluorescent dye terminator chemistry. The fluorescent data are automatically processed to produce high-quality base sequences. This is only possible because the Windows-based CEQ2000 software, unlike the software of other sequencing systems, computes a quality value (QV) for each base called.

\[ QV = -10\log_{10}(p) \]

where \( p \) is the estimated probability of error for the called base (1).

Version 4.0 and later versions of the CEQ2000 analysis software tag putative heterozygous base positions, after base calling is complete, using operator-selected parameters for percent of average peak spacing, height ratio, and sensitivity (3). Putative SNPs are reported using International Union of Biochemistry ambiguity codes. This software detects putative heterozygous positions within sequences but cannot determine whether a particular sample is homozygous at a position for a base other than that most commonly found at that position. Through comparison of individual sequences with a SNPCEQer-generated consensus sequence, SNPCEQer can detect such SNPs and grade them according to their CEQ2000 quality values. SNPCEQer also reports heterozygous positions tagged by the CEQ2000 software, the consensus sequence, and other relevant summaries and statistics.

MATERIALS AND METHODS

SNPCEQer

SNPCEQer was developed to detect SNPs in sequences produced using the CEQ2000 Analysis System with heterozygote detection analysis. SNPCEQer computes a multiple sequence alignment using an algorithm based on tree-based consistency objective function for alignment evaluation (T-Coffee) (10). T-Coffee is based on the widely used progressive approach to multiple sequence alignment. T-Coffee processes a data set of all pairwise alignments between the sequences and uses the resulting library of alignment information to guide the progressive alignment.

SNPCEQer analyzes up to 500 sequences input in *.phd format. The user may include a GenBank® sequence or other test sequence in text file format in the analysis to which SNPCEQer assigns a uniform quality value of 50. The sequences contained in the *.phd files are trimmed using a sliding window approach and quality value information supplied from the CEQ2000 software. Sequences shorter than an adjustable threshold length after trimming are eliminated before multiple alignment.

The program generates a consensus sequence and tags high-quality discrepancies between individual sequences and the consensus as putative SNPs. The quality values for the individual sequences are those assigned by the CEQ2000 software. The quality value for any position in the consensus sequence is the average of all quality values from sequences contributing to the consensus at this position. SNPs with quality values below a defined threshold are not reported.

The output from the program includes a text file (snp.txt) presenting the putative SNPs in dbSNP submission format and graded according to quality, number of SNPs at each polymorphic position, and number of SNPs in each sequence analyzed. The alignment of all the sequences input and the consensus generated from these sequences (myalign.aln) is presented in BioEdit format (7). BioEdit can be obtained from http://jwbrrown.mbio.ncsu.edu/BioEdit/bioedit.html. Information relating to the computation of the consensus sequence (pm.txt), sequences input for analysis (allSeq.txt), discarded sequences (badSeq.txt), and sequences analyzed (analyzedSeq.txt) are reported.
## Table 1. Gene Regions Sequenced, Primers Used for PCR and Sequencing, and Experimental Details

<table>
<thead>
<tr>
<th>Gene</th>
<th>GenBank Accession and Subsequence Amplified</th>
<th>Product Size</th>
<th>PCR Primers (Forward/Reverse)</th>
<th>PCR, T&lt;sub&gt;ann&lt;/sub&gt;, PCR (°C), Premix Used, or Additions to PCR</th>
<th>Sequencing Primers (Forward/Reverse)</th>
<th>T&lt;sub&gt;ann&lt;/sub&gt; Sequencing</th>
<th>No. of Samples</th>
</tr>
</thead>
<tbody>
<tr>
<td>EDNRA</td>
<td>D11145, 79-525</td>
<td>447</td>
<td>AATTGCCTCAAGATGAAACC TGGTACTTCTC TACCTATAATACATTG</td>
<td>Epicentre, 60.0, Premix D</td>
<td>CTCAAGATGAAACCCTTTG</td>
<td>60.0</td>
<td>188</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>AGAGATCTACCTGGGGAACCTG AGGAAGGTGCTGATCTGGAAG</td>
<td>CCGACGACACTGATCCT</td>
<td>TGGTACTTCTACCTATAATACATTG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>BDKRB2</td>
<td>S45489, 342-941</td>
<td>600</td>
<td>ATGCCTATTCCTGATAGATCCAGAG ACGGCAGGGTCTTGTG</td>
<td>Epicentre, 60.0, Premix K</td>
<td>ATGCCTATTCCTGATAGATCAGAG</td>
<td>61.0</td>
<td>186</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Epicentre, 64.0, Premix J</td>
<td>CCGACGACACTGATCCT</td>
<td>AGGAAGGTGCTGATCTGGAAG</td>
<td></td>
<td></td>
</tr>
<tr>
<td>RUNX2</td>
<td>NM_004348.1, 1-371</td>
<td>371</td>
<td>ATGCCTATTCCTGATAGATCCAGAG ACGGCAGGGTCTTGTG</td>
<td>Epicentre, 64.0, Premix J</td>
<td>ATGCCTATTCCTGATAGATCAGAG</td>
<td>61.0</td>
<td>96</td>
</tr>
<tr>
<td>CREB1</td>
<td>10716632, 1333373-133867</td>
<td>495</td>
<td>GCCAAAGAGTGTCTGGGATG GGTACTTCTCAGTCAGTTTCTCC</td>
<td>Epicentre, 59.5, Premix G</td>
<td>GGTACTTCTCAGTCAGTTTCTCC</td>
<td>59.5</td>
<td>192</td>
</tr>
<tr>
<td>ADRB1</td>
<td>12329314, 361117-36750</td>
<td>633</td>
<td>GAGCCCCGGTAACTGTCG CCGTTGAGGAAGCT</td>
<td>Epicentre, 60.5, Premix J</td>
<td>GAGCCCCGGTAACTGTCG</td>
<td>60.5</td>
<td>190</td>
</tr>
<tr>
<td>ADRB2</td>
<td>29370, 1287-1868</td>
<td>582</td>
<td>CTTCTTGCACATCTCAGATG TGCCGTTGCTGAGATG</td>
<td>Epicentre, 59.3, Premix G</td>
<td>—</td>
<td>59.3</td>
<td>192</td>
</tr>
<tr>
<td>MCIP1</td>
<td>14780300, 109773-110435</td>
<td>663</td>
<td>ACAGTCACAAGTCTCTTGGT CTGGACTCCCAGAATGTCT</td>
<td>Epicentre, 59.5, Premix L</td>
<td>ACAGTCACAAGTCTCTTGGT</td>
<td>59.5</td>
<td>192</td>
</tr>
<tr>
<td>KPNB2</td>
<td>NT_006713, 1260683-1261146</td>
<td>463</td>
<td>CATGGGCTGTGCTTACTAAT TGGGTAATGGGAGACGCT</td>
<td>Epicentre, 61.3, Premix D</td>
<td>CATGGGCTGTGCTTACTAAT</td>
<td>61.0</td>
<td>96</td>
</tr>
<tr>
<td>MLYCD</td>
<td>NT_010428.8, 436405-436828</td>
<td>424</td>
<td>TAATCTTGGATCTGTTGCTGGTAC GAGAGTTGCTTGGACAATTGTC</td>
<td>AmpliTaq Gold, 55.0, 1.25 M betaine</td>
<td>TAATCTTGGATCTGTTGCTGGTAC</td>
<td>57.0</td>
<td>306</td>
</tr>
<tr>
<td>MAP4K3</td>
<td>AC007684, 171854-172396</td>
<td>543</td>
<td>TGTCTGGGGTGAAGACTACA ATGAAAGAATGAGAGG</td>
<td>AmpliTaq Gold, 62.7</td>
<td>CGATAAAAAAGCCTCACCTCAA</td>
<td>57.0</td>
<td>96</td>
</tr>
<tr>
<td>REV1L</td>
<td>NT_022171.8, 2102049-2102501</td>
<td>453</td>
<td>ATGAGCAACATGGCCAAAGGT AAGGAATCCAGGGGTTTCT</td>
<td>AmpliTaq Gold, 54.2, 5% DMSO</td>
<td>AAGGAATTACCCAGGGTTTCT</td>
<td>57.0</td>
<td>166</td>
</tr>
<tr>
<td>TRF4</td>
<td>NM_006999.2, 293254-293763</td>
<td>510</td>
<td>CACAACGGCAGTAAACTGTG CACACGGCTCCTCAACC</td>
<td>AmpliTaq Gold, 64.0, 5% DMSO</td>
<td>CTCGTTGGTACGGAGGT</td>
<td>57.0</td>
<td>166</td>
</tr>
<tr>
<td>Gene</td>
<td>SNP Position (GenBank accession and Position where Available)</td>
<td>Sequence Context (5′→3′)</td>
<td>SNPCEQer SNP No.</td>
<td>CEQ Software and SNPCEQer</td>
<td>PolyPhred</td>
<td>Both Programs</td>
<td></td>
</tr>
<tr>
<td>--------</td>
<td>---------------------------------------------------------------</td>
<td>--------------------------</td>
<td>------------------</td>
<td>----------------------------</td>
<td>------------</td>
<td>---------------</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Sequences Input into SNPCEQer and dbSNP Reference (where Available)</td>
<td></td>
<td>SNP No.</td>
<td>Heterozygous</td>
<td>Homozygous</td>
<td>Heterozygous</td>
<td>Homozygous</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S</td>
<td>R</td>
<td>C</td>
<td>S</td>
</tr>
<tr>
<td>EDNRA</td>
<td></td>
<td></td>
<td>188</td>
<td>1</td>
<td>D11145</td>
<td>475</td>
<td>CTTGCCCTTTGAGACCCCTAT (C/M)TATGTGTCATCGATCTCCTCC</td>
</tr>
<tr>
<td>BDKRB2</td>
<td></td>
<td></td>
<td>186</td>
<td>1</td>
<td>S45489</td>
<td>444</td>
<td>AACTTGACTGGCTTTTGG (G/A)GAGAGGCCTGGCCGCTGAGTGT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>624</td>
<td>2</td>
<td>576</td>
<td>GGTGATCTGGGCTGGTGTCGCTGCT (G/R)TCCTGAGCTCAACCATGCT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>692</td>
<td>3</td>
<td>624</td>
<td>GCTGAGGGCCACACGGTCA (C/S)CGCTTGTGTCATCGATGCTACC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>858</td>
<td>rs5224</td>
<td>424</td>
<td>AAGTTCAAGGAGATCCAGAC (G/R)GAGAGGAGGGCCACACGGTCT</td>
</tr>
<tr>
<td>RUNX2</td>
<td></td>
<td></td>
<td>96</td>
<td>1</td>
<td>NM_004348.1</td>
<td>198</td>
<td>GCCGCCCGCGGTCGGCGCGGC (A/R)GCCGCCGGGTCCGGCGCGCGGCGC</td>
</tr>
<tr>
<td>ADRB1</td>
<td></td>
<td></td>
<td>190</td>
<td>1</td>
<td>NM_004348.1</td>
<td>231</td>
<td>CGTACAGACGGCCGCTGGCGGC (T/Y)TTCTCGGAGGAGGGACGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>293</td>
<td>2</td>
<td>293</td>
<td>TTGCCCGCCACACGATGAGCAG (C/Y)ACGATGACCCAGCAGCAG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>323</td>
<td>3</td>
<td>323</td>
<td>GTCTGCGATGCGCCACGAT (C/T)ACACGACATTGCACCCAC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>401</td>
<td>4</td>
<td>401</td>
<td>ACCACCGACGGCCCACATGAC (C/M)AGGTCGGCGCTGGCCAGGGA</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>5</td>
<td>626</td>
<td>5</td>
<td>626</td>
<td>GGGAGATGGGCCCAACGGTG (G/R)CACACGAGGGCCGGGCCGCGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>6</td>
<td>657</td>
<td>6</td>
<td>657</td>
<td>GTGATGAGATGGGGGACGA (A/R)GGACACCCAGGGCCGAGATGG</td>
</tr>
<tr>
<td>ADRB2</td>
<td></td>
<td></td>
<td>192</td>
<td>1</td>
<td>NM_000024</td>
<td>742</td>
<td>GGCTCTCTGCTGGTGTCGCCC (G/T)GACCAGTCTGCTCAGTGG</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>2</td>
<td>1059</td>
<td>2</td>
<td>1059</td>
<td>AGCCAGCGAGGGTGAAAGGT (G/R)CCATGATGCTGATCGGCTGATT</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>3</td>
<td>1075</td>
<td>3</td>
<td>1075</td>
<td>AAGGTGAAGAGGAGGCAGGC (A/R)GCAGGGGCTGGAATGGCC</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>4</td>
<td>1103</td>
<td>4</td>
<td>1103</td>
<td>GTTATCTGTGATACATGTG (A/R)CAATGTAAAGCGATGGAGAAG</td>
</tr>
<tr>
<td>Gene</td>
<td>SNP No.</td>
<td>SNP Position (GenBank accession) and Position and dbSNP Reference (5′→3′)</td>
<td>Sequence Context (5′→3′)</td>
<td>SNPs Reported (Heterozygous Positions/Homozygous Positions)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>---------</td>
<td>---------</td>
<td>--------------------------------------------------------------------------</td>
<td>--------------------------</td>
<td>-------------------------------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>CEQ Software and SNPCEQer PolyPhred Both Programs</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Heterozygous Homozygous Heterozygous Homozygous Heterozygous Homozygous</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>S R C S R C S R C S R C S R C S R C</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MCIP1</td>
<td>184</td>
<td>NM_004414.2 533 GGATCGGCGTGTACTCCGGC (C/Y)TCCTGGCTCTGGAATAATT    1 1 1B 0</td>
<td>1 1 1B 0 1 1 1P 0 1 1 1B 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>537 AGGTGAGCGGCGGTACTCCGGC (C/Y)GCGCTCCTGTGCTGGAATAATT 2 7 2S,3B 1 1 1B</td>
<td>7 5 2S,3B 1 1 1B 3 3 3P 1 1 1P 3 3 3B 1 1 1B</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>KPNB2</td>
<td>96</td>
<td>NT_006713 1260811 ATGAATGCCACCAAAAATTT (C/Y)TCTAAAACCAAAACAGGT</td>
<td>1 1 0P 0 0 0 0 0 0 0 0 0</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>MLYCD</td>
<td>306</td>
<td>NT_010428.8 436706 GGCCTGGGATTTATCTTCTC (C/Y)TTTCAGGAAATGAATGGGG</td>
<td>19 3 24 3 18 3</td>
<td>19 3 24 3 18 3</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>2</td>
<td>436796 TGCACAATTGTCTTATGTAT (C/Y)GTGCTTGAGTTACCATTCATT</td>
<td>28 11 25 7 22 7</td>
<td>28 11 25 7 22 7</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

S SNPs discovered in *forward sequences*
R Number of *reverse sequences* corresponding to primary sequences containing SNPs available for analysis
C SNPs confirmed in *reverse sequences*
P Confirmed in *reverse sequence* by PolyPhred analysis
S Confirmed in *reverse sequence* by SNPCEQer analysis
B Confirmed *reverse sequence* by both programs
T Trimmed by PhredPhrap

*For the purposes of this table, the phrase ‘forward sequences’ refers to the sequences analyzed initially and ‘reverse sequences’ refers to samples sequenced on the opposite strand to confirm the presence of an SNP.*
Short Technical Reports

Test Sequences and PCR Primers

The efficacy of SNPCEQer was demonstrated on sequences obtained for 12 sections of the human genome. These sections were PCR-amplified from genomic DNA samples obtained from collaborators in University of Texas Southwestern Medical Center and sequenced using Beckman Coulter dye terminator cycle sequencing. The samples were from patients with conditions ranging from appendicitis to cancer to dilated cardiomyopathy. Table 1 details gene regions sequenced, sample numbers, PCR and sequencing primers, and PCR conditions.

DNA Amplification and Sequencing

PCRs were set up using approximately 100 ng template DNA, 1 µM each primer (Operon Technologies, Alameda, CA, USA), 0.625 U enzyme, and either the AmpliTaq Gold® (Applied Biosystems, Foster City, CA, USA) system with 200 µM each dNTP (Invitrogen, Carlsbad, CA, USA) or the FasilSafe™ PCR system (EPICENTRE, Madison, WI, USA).

Thermal cycling was conducted in a PTC-100™ thermal cycler (MJ Research, Watertown, MA, USA). Reactions were subjected to 96°C for 3 min; 40 cycles of 96°C for 45 s, annealing temperature (T\textsubscript{a}) for 45 s (Table 1), 72°C for 2.5 min; and 72°C for 4 min. PCR products were purified using a PCR Purification Kit (Qiagen, Valencia, CA, USA). PCR product concentration was estimated from electrophoresis of sample against GeneChoice® Ladder II (PGC Scientifics, MD, USA).

Cycle sequencing reactions employed CEQ DTCS kits (Beckman Coulter) and used 1.25 µM primer, approximately 100 ng template DNA, and 50 fmol PCR product. Sequencing reactions were subjected to 96°C for 1 min; 40 cycles of 96°C for 40 s, T\textsubscript{a} for 40 s (Table 1), 61°C for 4 min; and 61°C for 4 min. Sequencing was conducted on CEQ2000 sequencers using the Long Fast Read method.

SNPCEQer and PolyPhred Analysis

The sequences were analyzed using SNPCEQer on Windows NT and 2000 and using PolyPhred on Redhat Linux 7.0 to compare SNPCEQer with a widely used, highly regarded SNP detection program. SNPCEQer was fed Beckman CEQ2000-generated sequences produced using heterozygote detection analysis with 50% for percentage of average peak spacing and for height ratio, 0.25 for the sensitivity parameter, and 50–550 nucleotides or 10–50 nucleotides before the last called base as the range parameter. Bases were trimmed from both ends of the sequence until less than 4 of 25 consecutive bases had quality values less than 20. The minimum threshold length for inclusion in SNPCEQer analysis after trimming was set at 200 bases. The SNPCEQer quality value threshold was set at 20 because this value approximates the single pass error rate for gel-based sequencing (approximately 1%) but minimizes false-negative calls. Together with the CEQ2000 heterozygote analysis detection parameters selected, this threshold quality value gave equivalent levels of SNP reporting with PolyPhred for most of the sequences analyzed. PolyPhred analysis was conducted on the same sequences using PhredPhrap for base calling and sequence assembly and Consed for assembly display (6). The PolyPhred program was tuned (trained) on sequences obtained using Applied Biosystems chemistry and sequencing and may not be optimized for the CEQ data at this time (9). Chromatograms containing putative SNPs were visually examined to confirm the SNP. In some instances, samples containing SNPs were sequenced on the opposite strand to confirm the SNP.

RESULTS AND DISCUSSION

SNPCEQer reported comparable numbers of SNPs to PolyPhred (148 vs. 165, respectively) in regions amplified from eight genes: EDNRA, BDKRB2, RUNX2, ADRB1, ADRB2, MCIP1, KPNB2, and MLYCD. All of these putative SNPs were confirmed by visual inspection of the chromatograms. Twenty-one polymorphic positions were identified. Both programs reported SNPs in the same positions, except in four cases (Table 2). SNPCEQer reported ADRB1 SNP 4, ADRB2 SNP 1, and KPNB2 SNP 1, but PolyPhred did not, whereas the reverse was true for ADRB2 SNP 2. The SNPs reported are detailed in Table 2.

In most cases, both programs reported SNPs at the same positions but did not necessarily tag the SNP in the same sequences: BDKRB2 SNP 4, RUNX2 SNP 1, ADRB1 SNP 4, ADRB2 SNPs 1–4, MCIP1 SNP 2, KPNB2 SNP 1, and MLYCD SNPs 1 and 2. PolyPhred reported BDKRB2 SNP 4, RUNX2 SNP 1, ADRB1 SNP 1, and MLYCD SNP 1 in more sequences than did the combination of CEQ2000 heterozygote detection software and SNPCEQer. The reverse was true for the other SNPs.

For example, PolyPhred reported ADRB1 SNP 1 in 46 of 190 sequences, whereas SNPCEQer tagged the SNP in 14 sequences. SNPCEQer did not report the SNP in these sequences because of the proximity of the SNP to one end of the sequence. The SNP was not reported because of trimming in 14 cases, because the quality value was less than 20 in 16 cases and the heterozygote detection sensitivity parameter was too stringent in four cases. It is recommended that, when SNPCEQer detects a SNP close to one end of the sequence, the sequences be reanalyzed using a reduced sensitivity threshold, more “relaxed” trimming criteria (e.g., trimming until less than 4 of 25 consecutive bases have quality values less than 15), and a lower minimum quality value threshold (14). Following reanalysis of these sequences, the total number of SNPs discovered by SNPCEQer was 179 versus 165 for PolyPhred. The adjustment of these parameters controls the balance between false-positive and false-negative rates and enables the investigator to tailor the analysis to individual sets of sequences. PolyPhred also failed to call SNPs because of PhredPhrap sequence trimming. Four sequences were reported to contain ADRB2 SNP 1 by SNPCEQer but not by PolyPhred.

BDKRB2 SNP 3 and KPNB2 SNP 1 were reported by SNPCEQer but not found in sequences on the opposite strand for the same samples. PolyPhred also reported KPNB2 SNP 1. ADRB1 SNP 4 was reported by SNPCEQer.
only and ADRB2 SNP 2 by PolyPhred only. These findings emphasize the importance of sequencing in both directions and that an alternative method of SNP confirmation may be required.

Initially, SNPs were reported for all sets of sequences analyzed, but many of these were false-positive SNPs located in stretches of consecutive Ts. A filter was included in the SNPCEQer program to flag these SNPs as suspect, except where the SNP was reported at the first T. No genuine SNPs were reported in the regions sequenced from REV1L, CREB1, TRF4, and MAP4K3.

SNPCEQer offers a convenient method of compiling a list of potential polymorphic positions for further examination. SNPCEQer provides a functionality not currently available as part of the CEQ2000 software analysis package; since it can run in a Windows NT environment, it offers convenience and familiarity to the typical user.

REFERENCES


BioTechniques 33:822-830 (October 2002)

ABSTRACT

There is widespread interest in devising genotyping methods for SNPs that are robust, inexpensive, and simple to perform. Although several high-throughput SNP genotyping technologies have been developed, including the oligonucleotide ligation assay, real-time PCR, and mass spectrometry, the issues of simplicity and cost-effectiveness have not been adequately addressed. Here we describe the application of a novel computer software package, SNPkit, which designs SNP genotyping assays based on a classical approach for discriminating alleles, restriction enzyme digestion. SNPkit can be used in genotyping assays for almost any SNPs including those that do not alter “natural” restriction sites. Using this method, 164 SNPs have been evaluated in DNA samples from 48 immortalized cell lines of randomly selected Chinese subjects. Sixty-two (37.8%) of the SNPs appeared to be common (frequencies of the minor alleles ≥5%) and were subsequently applied to a larger population-based sample. Overall, by using SNPkit, we have been able to validate and genotype accurately a large fraction of publicly available SNPs without sophisticated instrumentation.

INTRODUCTION

SNPs are the most abundant type of DNA variation, with an estimated frequency of 1 in every 300–500 bp (7), and they provide important physical landmarks for identifying genes important in human disease and drug responses (13). Because of rigorous efforts in SNP discovery, more than 2 million SNPs have been identified and cataloged (19).

Several publicly available SNP data-

For reprints of this or any other article, contact
Reprints@BioTechniques.com