2003

Efficient high-throughput resequencing of genomic DNA

Raymond D. Miller
Washington University School of Medicine in St. Louis
Shenghui Duan
Washington University School of Medicine in St. Louis
Elizabeth G. Lovins
Washington University School of Medicine in St. Louis
Ellen F. Kloss
Washington University School of Medicine in St. Louis
Pui-Yan Kwok
Washington University School of Medicine in St. Louis

Efficient High-Throughput Resequencing of Genomic DNA

Raymond D. Miller, Shenghui Duan, Elizabeth G. Lovins, et al.

Genome Res. 2003 13: 717-720
Access the most recent version at doi:10.1101/gr.886203

Efficient High-Throughput Resequencing of Genomic DNA

Raymond D. Miller, Shenghui Duan, Elizabeth G. Lovins,[1] Ellen F. Kloss, and Pui-Yan Kwok[1,2]

Washington University, Division of Dermatology, St. Louis, Missouri 63110, USA

Targeted resequencing of genomic DNA from organisms such as humans is an important tool enabling experimental access to variation within the species and between similar species. Taking full advantage of the reference genome sequences in designing robust, specific PCR assays and using stringent conditions, resequencing can be done efficiently without purification of the PCR product. By using a 10-fold greater amount of one primer when setting up the PCR initially in a new version of asymmetric PCR, one simply adds the rest of the sequencing reagents at the end of PCR and allows the sequencing reaction to proceed, with the excess PCR primer serving as the sequencing primer. We demonstrated that this streamlined protocol can be used with PCR products up to 1300 bp and had up to a 97% success rate in high-throughput analysis of allele frequencies for >30,000 single-nucleotide polymorphisms (SNPs). SNP primers and characterization results are provided at http://snp.wustl.edu.

The detailed study of genetic variation within populations and among related studies has been greatly aided by techniques of resequencing coupled with the availability of genomic draft sequences from a number of organisms including humans (International Human Genome Sequencing Consortium 2001). Genetic variations including single-nucleotide polymorphisms (SNPs) have been detected by several kinds of resequencing protocols. First, comparison of overlapping (i.e., resequenced) draft sequence derived from two chromosomes has allowed efficient detection of human SNPs (Taillon-Miller et al. 1998; Marth et al. 1999). Second, comparison of newly determined sequences from random locations with draft sequence has also allowed detection of SNPs (Altshuler et al. 2000; The International SNP Map Working Group 2001; Holden 2002). Third, targeted resequencing of genomic PCR products, either from multiple individuals of the same species or between species including orangutan compared with human DNA, has provided an additional means to scan for variation (Taillon-Miller et al. 1999; Miller et al. 2001).

Our group has developed a high-throughput pipeline using targeted resequencing of pooled human DNAs as a means to characterize the allele frequencies of SNP candidates (Marth et al. 2001; Vieux et al. 2002). The process includes design of stringent PCR assays performed under uniform conditions, proceeding without a cleanup step to cycle sequencing using chain-termination fluorescent dyes (Fig. 1). Although we have used this process for high-throughput resequencing, it could easily be adapted for smaller numbers of samples.

A time-consuming and costly step in our original targeted resequencing was the use of size-exclusion columns to remove the PCR primers and dNTPs at the end of PCR, followed by the readdition of one primer for cycle sequencing, for example (Taillon-Miller et al. 1999). We reasoned that this process could be made more efficient if the PCR reaction is very specific (with only one product formed), if one (and only one) of the PCR primers is left at the end of PCR to serve as primer in the subsequent sequencing reaction, and if the residual dNTPs do not interfere with the sequencing reaction. We report here that by designing robust and specific PCR assays based on whole-genome sequence data, reducing significantly the amount of dNTPs used in PCR, and using a 10-to-1 excess of one of the PCR primers in the protocol, one can obtain excellent DNA sequence data by simply adding the PCR product to the sequencing master mix (containing the DNA polymerase, fluorescent dye-terminators, and buffer). We have applied this streamlined protocol to characterize >30,000 SNP candidates by pooled DNA resequencing with excellent results.

[1]Present address: University of California, San Francisco, Cardiovascular Research Institute, San Francisco, California 94143-0130, USA.
[2]Corresponding author.
E-MAIL kwok@cvrimail.ucsf.edu; FAX (415) 476-2283.
Article and publication are at http://www.genome.org/cgi/doi/10.1101/gr.886203. Article published online before print in March 2003.

RESULTS AND DISCUSSION

To examine the range of effective lengths one can use with the streamlined resequencing approach, we designed a series of PCR primers approximately every 100 or 200 bases in an 800-bp interval in the human β-*glucuronidase* (*GUSB*) gene on Chromosome 7. We used high-throughput primer design methods to avoid locating any primer in repetitive DNA, and to yield uniform but stringent (55°C optimal T_m with reaction annealing at 58°C and a hot start *Taq* DNA polymerase) PCR reactions, in which any left primer could be paired with any right primer (see Methods). Using different combinations of PCR primers, we were able to generate PCR products ranging from 100–790 bp in length.

We analyzed various size PCR products from the grid by agarose gel electrophoresis using DNA from each of three individuals to see the effects of 1:1 primer ratios compared with 10:1 ratios and also dNTP concentration using 0.25, 0.5, 1.0, and 2.0 nmole of each dNTP per reaction in 10-μL reaction volumes. Theoretically, with 1.0 pmole of primer used in the reaction, an 800-bp product would maximally require ~0.40 nmole of each dNTP. The two lower levels of dNTP (0.25 and 0.5 nmole/10 μL) concentration were limiting to the PCR re-

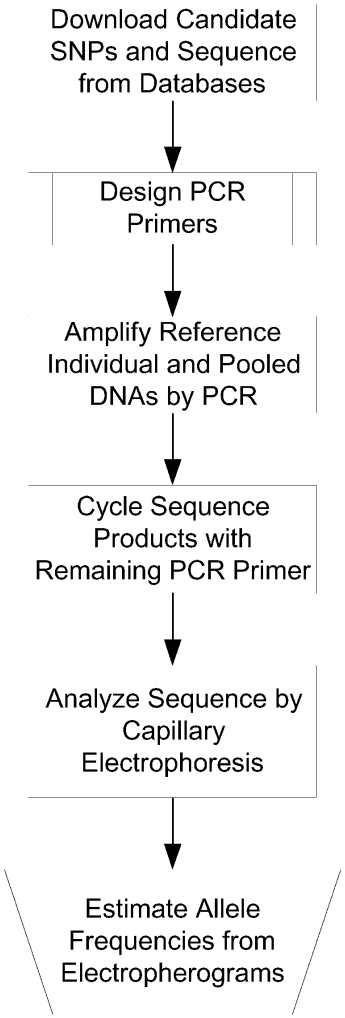

Figure 1 Resequencing flow diagram for SNPs. Databases (e.g., http://www.ncbi.nlm.nih.gov/SNP/; http://snp.cshl.org/) provided the surrounding sequence necessary for primer design (see Methods). With adjustment of input DNA, a similar flow diagram could be used for other resequencing projects, such as for exons.

action, even for shorter products. Using the two higher concentrations of dNTP (1.0 and 2.0 nmole/10 µL), strong and unique PCR products were produced from all primer combinations with either the 1:1 or 10:1 primer ratios, except those involving primer 1left. PCR reactions using the primer 1left contained an additional band, and sequencing results from these reactions always yielded poor results. We searched this 20-bp primer against the genomic sequence using the BLAST algorithm (http://www.ncbi.nlm.nih.gov/blast/Blast.cgi) and found only one exact match (expected location), but there were several exact matches for the most 3′ 17 bp. Fortunately, the failure with this one primer was a rare case, and for all others, the assay design conditions, even with excess of one primer, produced abundant and apparently unique products (data not shown).

Given a PCR success of up to 790 bp with the streamlined protocol, we sought to increase this range. Because of repetitive sequences, we could not increase the primer grid to the right, but we were able to increase the range to the left, yielding products up to 1310 bp with primers of the opposite orientation. Additionally, we chose a new primer, 40left, to replace the problematic primer 1left (see Methods). Appropriately sized, unique bands were produced using these primers combined with right primers. For PCR products up to the maximum size we tested, we observed little difference in the amount of product produced with PCR extension times of 30 or 60 seconds (data not shown).

We performed DNA sequencing with a range of PCR products using the streamlined protocol with a 10:1 primer ratio and reduced dNTP concentration. One concern about the protocol was that without a cleanup step, the *Taq* DNA polymerase and unincorporated dNTPs would be carried over from the PCR to the sequencing reaction and could perhaps unbalance the dNTP/ddNTP ratios established by the manufacturer for sequencing, leading to faint signals for shorter products. This effect did not prove to be a practical problem (data not shown). With longer PCR products, we obtained good sequence up to >500 bp. The practical extent of sequencing was limited by the quality of information from the capillary sequencer, not by the protocol.

A method for resequencing called direct amplification and sequencing (DEXAS) is nominally simpler than our protocol because the PCR and sequencing steps are performed in the same tube at the same time using two kinds of DNA polymerases (Kilger and Paabo 1997). However, with competing polymerases, single optimized conditions that yield consistent results for high-throughput reaction using a variety of GC content DNAs may not be possible for DEXAS. Also, initial versions of DEXAS required much DNA (60 ng vs. 4 ng for our method) and a dye-labeled sequencing primer, which adds expense and limits the flexibility of the approach. Our protocol, although novel, uses straightforward principles. Any kind of genomic DNA for which sequence is available could be reliably resequenced using these methods. The various steps are each robust, and one can easily apply high-throughput techniques.

Asymmetric PCR is a well-known approach to produce single-stranded DNA for a variety of applications. Because the PCR primers used are in different concentrations, the method presented here is, by definition, a version of asymmetric PCR. However, the intent of various forms of early asymmetric PCR differed from the intent in this protocol. The initial presentation of asymmetric PCR some 15 years ago (before cycle sequencing was possible) was in response to the problems of trying to sequence double-stranded PCR products that rapidly reannealed to block priming. The first protocol used a 100:1 ratio of PCR primers to produce a mixture of double- and single-strand DNA, which, after purification, was sequenced by adding back the initially rare primer and other reagents (Gyllensten and Erlich 1989). Another early technique used a temporally asymmetric PCR: After regular PCR and cleanup, one primer was readded for more thermocycling before further cleanup and sequencing (Bartek et al. 1991; Kaltenboeck et al. 1992). A third technique, thermal asymmetric PCR, was accomplished by the use of primers with different melting temperatures for PCR, readdition of the higher melting temperature primer without purification, and cycle sequencing with radiolabel (e.g., Liu et al. 1993). In our method, production of single-stranded DNA is not the intent, and the single-stranded DNA formed is not the template for the subsequent sequencing step. The main goal is to streamline the sequencing protocol without compromising sequencing quality while reducing the cost of sequencing.

Although we used the protocol for high-throughput resequencing, it could be applied to projects of any size. We have used our methods to design primers for all human SNPs in dbSNP possible, and the sequences are freely available (http://snp.wustl.edu). For other design needs, one can use the parameters listed in the Methods section below with the Web-based version of the Primer3 program. Our collaborators have successfully used the protocol to resequence various exons in human and mouse DNA.

The streamlined protocol has been used as the workhorse for our project to characterize the allele frequencies of human SNP candidates (e.g., Fig. 2). More than 30,000 SNPs in three populations have been successfully characterized and made publicly available (http://snp.wustl.edu; also dbSNP and The SNP Consortium Web sites). In general, the rate of technical success with the protocol has been high; for example, one of us (E.G.L.) has had a success rate of 97% with 6099 SNPs. When DNA sequence is available, the streamlined protocol provides a successful and efficient means to study genetic variation as long as the PCR product is unique and one of the PCR primers is exhausted at the end of PCR.

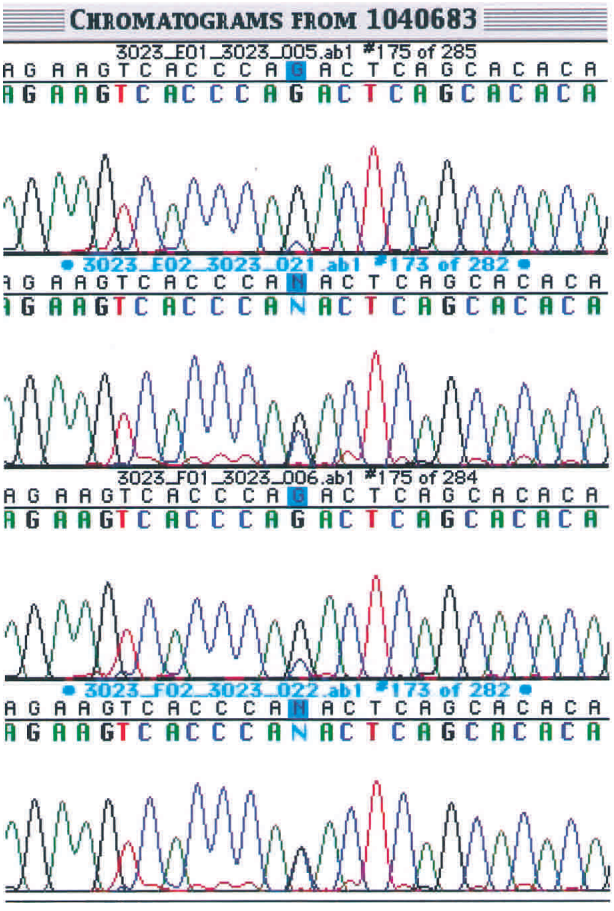

Figure 2 Aligned resequencing electropherograms for the region containing SNP rs1040683. From the *top*, the traces are, respectively, from African Americans, Asians, European Americans, and a reference DNA. The SNP, which maps to the X-chromosome, is highlighted. The reference individual is a C/G heterozygote, and the allele frequencies differ among the populations.

METHODS

Primer Design

Briefly, to pick primers for high-throughput targeted resequencing, we used several steps: targets were chosen, the DNA sequences were obtained from databases, repetitive sequences were determined and marked using RepeatMasker software (http://ftp.genome.washington.edu/RM), primers were chosen using set, stringent parameters and Primer3 software (0.9, Unix version, http://www-genome.wi.mit.edu/genome_software/; Rozen and Skaletsky 2000), and the results underwent a final analysis using a Perl script to screen out certain situations prone to sequencing failure such as long homopolymers (Vieux et al. 2002).

To provide a grid of left and right primers for testing the 10-to-1 protocol, we obtained genomic sequence surrounding SNP rs2008188 in intron 1 of *GUSB* from the Golden Path (April 2001 freeze, June 2002 chr7:64064170–64073837, http://genome.ucsc.edu/). The sequence was RepeatMasked and formatted with a Perl script in which the parameters for Primer3 were set including an optimal product size of 800 bp. After this pair was designed, the left primer (designated 1left) was fixed as a parameter for Primer3, and the product lengths were successively adjusted to create a series of right primers producing additional product lengths up to 790 bp at 100-bp or 200-bp increments. Similarly, the farthest right primer (designated 800right) was fixed as a parameter for Primer3, and the product lengths were adjusted to choose a series of left primers. Later, to extend the product range up to 1300 bp and to replace one problematic primer, additional primers were chosen using similar methods. For the primer name, the number refers to the approximate location (in base pairs) on an arbitrary grid, and any left primer could be paired with any right primer for PCR. For the first group the primers were 1left, 5′-ACTTGTAAATGCTGCCAAAT-3′; 100left, 5′-TTTCGCAAGTAATATACAACAGA-3′; 200left, 5′-TCACTATAGCTGACTCTCCTGTT-3′; 400left, 5′-CCAACTTTGTTTCCAATATTCT-3′; 600left, 5′-GGTACTGCTCTAGCAGACTTTT-3′; 700left, 5′-AAAATAAAGATCCACTTGATGGT-3′; 100right, 5′-CTGTTGTATATTACTTGCGAAAAG-3′; 200right, 5′-GATTTACTTTTGGGATACACTCA-3′; 400right, 5′-GAAGCTGGTTTAATCCATGTAG-3′; 600right, 5′-GTTCACTGAAGAGTACCAGAAAA-3′; 700right, 5′-ACCATCAAGTGGATCTTTATTTT-3′; 800right, 5′-TTTTATTCTGGGTTACATCATTC-3′. For the second group, the primers were −500left, 5′-ATTCTCACTCTTACGCTTTACCT-3′; −400left, 5′-ATCTTCAGTTTATGGTAAGTCCA-3′; −300left, 5′-GTTATTCTCTTTGAAGACCAATCT-3′; and 40left, 5′-GTTGAAACTCACCTGTATTTGAT3′. The primer pairs 1left with 800right and −500left with 800right, respectively, produce 790-bp and 1310-bp PCR products.

PCR/Sequencing Protocol

For PCR and sequencing, we used three human genomic DNAs from the CEPH collection (Coriell Institute, Camden, NJ). Primers were ordered in a 96-well format (Integrated DNA Technologies or QIAGEN Operon), and the primer stock concentration was adjusted to 40 µM. PCR amplification was in a 96- or 384-well format, each standard reaction in a 10-µL volume containing 50 mM KCl, 10 mM Tris-HCl (pH 8.3), 3.5 mM $MgCl_2$, 100 nM each of the 4 dNTPs, 1.0 µM primer-1 (also for sequencing), 0.1 µM primer-2, 4 ng of DNA, and 0.15 U of hot start *Taq* DNA polymerase (JumpStart *Taq*, Sigma-Aldrich; or Platinum *Taq*, Invitrogen Life Technologies). The thermocycling program consisted of an initial polymerase activation step at 95°C for 2 min, 35 cycles of denaturation at 92°C for 10 sec, annealing at 58°C for 20 sec, and extension at 68°C for 30 sec, followed by a final extension at 68°C for 10 min.

Each sequencing reaction contained 2.5 µL of the PCR reaction, 6.5 µL of water, 2.0 µL of BigDye version 3 mix (Applied Biosystems), and 1.0 µL of 5× sequencing buffer, according to the protocol of the dye manufacturer. The thermocycling program consisted of an initial step at 96°C for 2 min, then 25 cycles of denaturation at 96°C for 15 sec, annealing at 50°C for 1 sec, and extension at 60°C for 4 min. Removal of the extra dye was performed according to the manufacturer's protocol using columns in 96- or 384-well format (Princeton Separation or Genetix). Electrophoresis and sequencing detection were performed using an ABI PRISM 3700 DNA Analyzer (Applied Biosystems). Electropherograms were aligned with Sequencher software (Gene Codes), and allele frequencies were estimated as described (Kwok et al. 1994).

ACKNOWLEDGMENTS

We thank other members of the SNP characterization team including Mathew Minton, Nicholas Addleman, Andrew J. Reinhart, Rachel Donaldson, and Nicholas Pavelka, who have commented on and extensively used the streamlined protocol. We thank one reviewer for providing references to asymmetric PCR. This work was funded in part by grants from the NIH (HG01720 and GM63340) and The SNP Consortium.

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

Altshuler, D., Pollara, V.J., Cowles, C.R., Van Etten, W.J., Baldwin, J., Linton, L., and Lander, E.S. 2000. An SNP map of the human genome generated by reduced representation shotgun sequencing. *Nature* **407:** 513–516.

Bartek, J., Iggo, R., Vojtesek, B., and Lane, D.P. 1991. Asymmetric PCR-based strategy for genetic analysis of the p53 tumor suppressor gene in cell lines and tumor tissues. *Neoplasma* **38:** 93–99.

Gyllensten, U.B. and Erlich, H.A. 1989. Ancient roots for polymorphism at the HLA-DQ α locus in primates. *Proc. Natl. Acad. Sci.* **86:** 9986–9990.

Holden, A.L. 2002. The SNP consortium: Summary of a private consortium effort to develop an applied map of the human genome. *Biotechniques* **32:** S22–S26.

International Human Genome Sequencing Consortium. 2001. Initial sequencing and analysis of the human genome. *Nature* **409:** 860–921.

The International SNP Map Working Group. 2001. A map of human genome sequence variation containing 1.42 million single nucleotide polymorphisms. *Nature* **409:** 928–933.

Kaltenboeck, B., Spatafora, J.W., Zhang, X., Kousoulas, K.G., Blackwell, M., and Storz, J. 1992. Efficient production of single-stranded DNA as long as 2 kb for sequencing of PCR-amplified DNA. *Biotechniques* **12:** 164–171.

Kilger, C. and Paabo, S. 1997. Direct DNA sequence determination from total genomic DNA. *Nucleic Acids Res.* **25:** 2032–2034.

Kwok, P.Y., Carlson, C., Yager, T.D., Ankener, W., and Nickerson, D.A. 1994. Comparative analysis of human DNA variations by fluorescence-based sequencing of PCR products. *Genomics* **23:** 138–144.

Liu, Y.G., Mitsukawa, N., and Whittier, R.F. 1993. Rapid sequencing of unpurified PCR products by thermal asymmetric PCR cycle sequencing using unlabeled sequencing primers. *Nucleic Acids Res.* **21:** 3333–3334.

Marth, G.T., Korf, I., Yandell, M.D., Yeh, R.T., Gu, Z., Zakeri, H., Stitziel, N.O., Hillier, L., Kwok, P.Y., and Gish, W.R. 1999. A general approach to single-nucleotide polymorphism discovery. *Nat. Genet.* **23:** 452–456.

Marth, G., Yeh, R., Minton, M., Donaldson, R., Li, Q., Duan, S., Davenport, R., Miller, R.D., and Kwok, P.Y. 2001. Single-nucleotide polymorphisms in the public domain: How useful are they? *Nat. Genet.* **27:** 371–372.

Miller, R.D., Taillon-Miller, P., and Kwok, P.Y. 2001. Regions of low single-nucleotide polymorphism incidence in human and orangutan Xq: Deserts and recent coalescences. *Genomics* **71:** 78–88.

Rozen, S. and Skaletsky, H. 2000. Primer3 on the WWW for general users and for biologist programmers. *Methods Mol. Biol.* **132:** 365–386.

Taillon-Miller, P., Gu, Z.J., Li, Q., Hillier, L., and Kwok, P.Y. 1998. Overlapping genomic sequences—A treasure trove of single-nucleotide polymorphisms. *Genome Res.* **8:** 748–754.

Taillon-Miller, P., Piernot, E.E., and Kwok, P.Y. 1999. Efficient approach to unique single-nucleotide polymorphism discovery. *Genome Res.* **9:** 499–505.

Vieux, E.F., Kwok, P.Y., and Miller, R.D. 2002. Primer design for PCR and sequencing in high-throughput analysis of SNPs. *Biotechniques* **32:** S28–S32.

WEB SITE REFERENCES

http://snp.wustl.edu; SNP primers and characterization results are provided at the common link homepage at Washington University.

http://ftp.genome.washington.edu/RM; The Repeat Masker Server at the University of Washington.

http://genome.ucsc.edu/; The University of California at Santa Cruz Genome Browser.

http://snp.cshl.org/; The SNP Consortium.

http://www-genome.wi.mit.edu/genome_software/; The Whitehead Institute for Biomedical Research.

http://www.ncbi.nlm.nih.gov/blast/Blast.cgi; BLAST.

http://www.ncbi.nlm.nih.gov/SNP/; dbSNP.

Received October 8, 2002; accepted in revised form January 23, 2003.