Encoding PCR Products with Batch-stamps and Barcodes

Megan L. McCloskey[1]* • Reinhard Stöger[1] • R. Scott Hansen[2] • Charles D. Laird[1,3,4]*

Received: 31 January 2007 / Accepted: 29 May 2007

Polymerase Chain Reaction (PCR) has become the mainstay of DNA sequence analysis. Yet there is always uncertainty concerning the source of the template DNA that gave rise to a particular PCR product. The risks of contamination, biased amplification, and product redundancy are especially high when limited amounts of template DNA are used. We have developed and applied molecular encoding principles to solve this source-uncertainty problem for DNA sequences generated by standard PCR. Batch-stamps specify the date and sample identity, and barcodes detect template redundancy. Our approach thus enables classification of each PCR-derived sequence as valid, contaminant, or redundant, and provides a measure of sequence diversity. We recommend that batch-stamps and barcodes be used when amplifying irreplaceable DNAs and cDNAs available for forensic, clinical, single cell, and ancient DNA analyses.

KEY WORDS: PCR; Contamination; Redundancy; Batch-stamp; Barcode; Source DNA; populations of molecules

Introduction

One essential condition of DNA-based technologies is that sequence information be reliable (Lewontin, 1994). Sequence information obtained by PCR, however, always carries an element of uncertainty because of risks posed by contamination, biased amplification, and product redundancy. The potential clinical impact of this uncertainty was recently illustrated when "thousands (of people) were given antibiotics and a vaccine…(and) hospital beds were taken out of commission" after a PCR-based assay incorrectly indicated that 142 individuals were infected with the bacterium that causes whooping cough (Kolata 2007). It is thus apparent that the reliability of PCR-derived assays would be significantly improved if methods were available to distinguish valid from data-corrupting sequences.

We recently addressed this problem by applying molecular encoding principles to hairpin-bisulfite PCR (Miner et al. 2004). This approach is highly successful in identifying valid, double-stranded sequences and their methylation patterns, but requires bisulfite conversion and more template material than is usually available from irreplaceable or single-cell samples (Miner et al. 2004; Laird et al. 2004; Li et al. 1988). We now report a method by which molecular codes are added to single-stranded DNA or cDNA templates, thus providing a solution to the source-uncertainty problem for the standard form of PCR.

[1] Department of Biology, University of Washington, Seattle, WA 98195, USA
[2] Department of Medicine, University of Washington, Seattle, WA 98195, USA
[3] Department of Genome Sciences, University of Washington, Seattle, WA 98195, USA
[4] To whom correspondence should be addressed; email: cdlaird@u.washington.edu

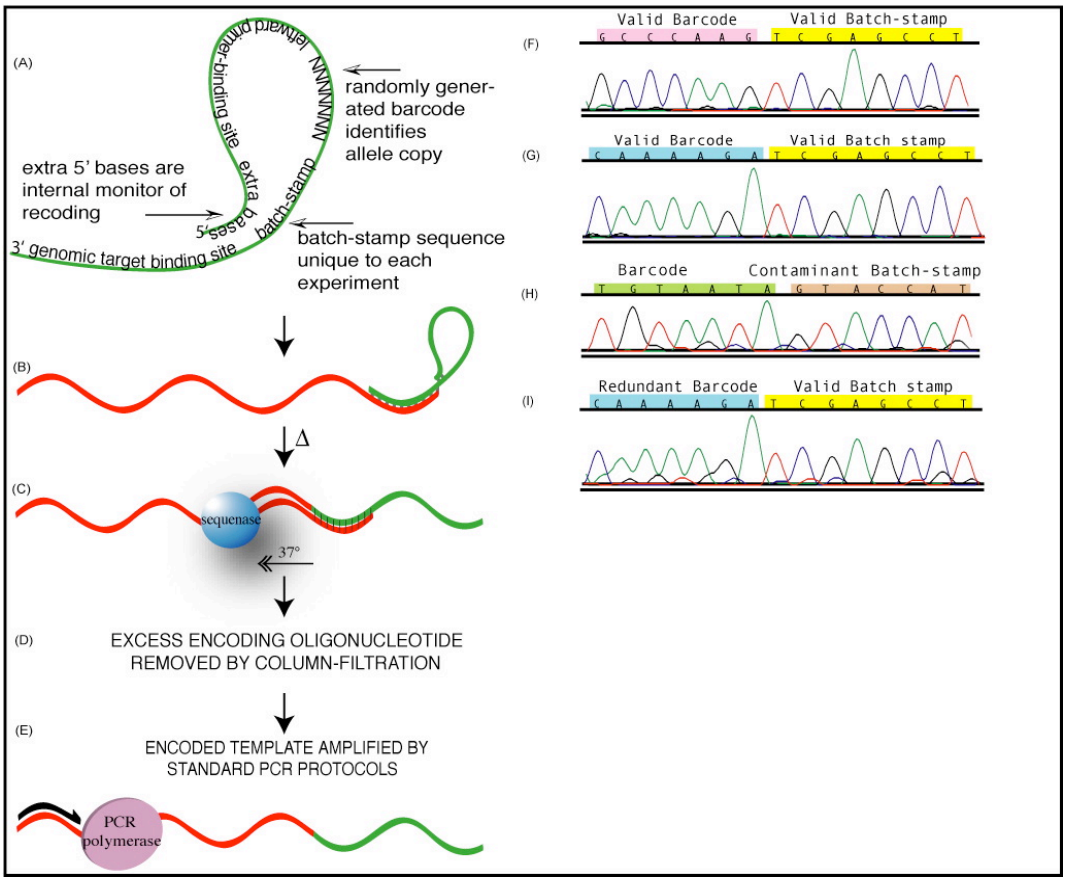

Fig. 1. Encoding oligonucleotide (A) has five distinct regions: a genomic target-binding site, a batch-stamp, a barcode, a primer-binding site for PCR, and extra bases that serve as an internal sentinel of recoding as well as tether the oligonucleotide to itself, thus minimizing non-specific annealing. Encoding oligonucleotides are annealed to denatured genomic DNA for 15 minutes at 37°C (B), and extended at 37°C by Sequenase (C). The resulting encoded DNA strands are column-purified with Qiagen QIAquick columns (D), to remove unincorporated encoding oligonucleotides prior to PCR amplification (E). PCR products were sequenced and their encoded information was evaluated. Three classes of encoded sequences recovered from a single PCR amplification are illustrated here. Of 320 sequences analyzed, 299 were valid, each having the correct batch-stamp and a distinct barcode (F and G), 13 were contaminants, having batch-stamps that were used in previous reactions (H), and eight were redundant, having barcodes and the batch-stamp identical to sequences previously analyzed from the same PCR (I).

Materials and Methods

General Design of the Encoding Oligonucleotide

Our method utilizes an encoding oligonucleotide with five distinct informational regions (Figure 1A): (i) the "batch-stamp", unique to each experiment, specifies the DNA source such as the patient or sample identification, and the date of reaction; (ii) the "random-barcode" distinguishes among sequences arising from different cell or allele copies; (iii) the "primer-binding site" facilitates specific PCR amplification; (iv) the genomic target-binding site targets the locus to be encoded prior to amplification; and (v) the "extra 5' bases" serve as an internal sentinel of recoding. This encoding oligonucleotide is annealed to denatured DNA, and extended by Sequenase (USB) (Figure 1B-C). Sequenase-extension products are column-filtered to remove any

unincorporated, encoding oligonucleotides (Figure 1D), and amplified using standard PCR protocols (Figure 1E).

Each encoding oligonucleotide is unique because of distinct batch-stamp, and random barcode, regions. There are several features, however, that are common to all encoding oligonucleotides. Each encoding oligonucleotide is designed to have only the genomic target-binding site available for annealing at 37˚C, while the remaining bases are unavailable due to secondary structure. At higher temperatures (≥ 55˚C), following Sequenase extension and column purification (see below), this secondary structure melts and makes available a primer-binding site for PCR amplification. Encoding oligonucleotides are between 65 and 75 bases in length, and were ordered with PAGE purification from Invitrogen. Upon receipt, encoding oligonucleotides were diluted to 50 μM and stored at -20˚C.

Encoding oligonucleotides were usually designed with the batch-stamp located adjacent to the 3' genomic target-binding site (Fig.1); this is the current standard in the lab. We have also designed and used oligonucleotides for which the positions of the barcode and batch-stamp were reversed. An encoding oligonucleotide of this form was used to encode the sequences presented here. This encoding oligonucleotide was ordered as:

5’-TCGAGCACATGCATGTCTTCAAAGTGGAGGCTCGANNNNNNNTCTCTTCAAGTGGCCTGGGAG-3’

Note that Fig. 1 illustrates sequences that are the reverse-complement of the ordered, encoding oligonucleotide.

Genomic Target-binding Site

The genomic target-binding site is designed to be available for annealing at 37˚C, as indicated by a $\Delta G \geq 0$ for this region (http://www.bioinfo.rpi.edu/applications/mfold/dna/form1.cgi). For our assays on the *FMR1* locus, the sequence of the genomic target-binding site was (5’-TCTCTTCAAGTGGCCTGGGAGC-3’).

Batch-stamp

We have used batch-stamps of different lengths in different experiments. We typically use between six and eight base pairs to encode sample identity and date of reaction. The valid batch-stamp shown in Fig. 1 is (5’-AGGCTCGA-3’). Each different nucleotide sequence yields a unique batch-stamp. The assigned batch-stamp is registered for a specific sample run on a specific date, and is used only once.

Barcode

The barcode is a random sequence of the four deoxyribonucleotides. The number of distinguishable barcodes in a population of oligonucleotides used in a reaction is determined by the number of random bases, *n*. This enables one to distinguish among 4^n allele copies per reaction. We currently use seven nucleotides for the barcode, giving 16,384 possible barcodes for each encoding oligonucleotide.

Primer-binding Site

The primer-binding site of the encoding oligonucleotide is the target of one of the primers used for PCR amplification. It is designed to have no complementarity with competing genomic fragments. The primer-binding site for the oligonucleotide shown in Fig 1 is 5’-ACATGCATGTCTTCAAAGTGG-3’. The other primer-binding site for PCR amplification is contained in genomic DNA that is copied during the initial Sequenase extension.

Internal Sentinel of Recoding

Extra bases 5’ to the primer binding site serve two functions: (i) these bases would be present in sequences if the oligonucleotide were used inappropriately as a primer during the final round of PCR amplification, and thus serve as a sentinel for recoding; (ii) these bases contribute to the secondary structure of the oligonucleotide at 37°C. The sentinel bases are chosen to tether the 5’ end of the oligonucleotide to a complementary portion of the batch-stamp, thus forming a hairpin loop that exposes only the genomic target-binding site for annealing to

template molecules. This design is intended to protect against preferential annealing of encoding oligonucleotides whose random barcodes have some complementarity to, and thus higher stability of hybridization with, the target locus. Non-random use of barcodes has been observed when hairpin-loop structures are used to encode double-stranded DNA molecules (Miner *et al.* 2004). At temperatures appropriate for PCR (≥55˚C), the secondary structure of the oligonucleotide will melt and expose the primer-binding site. Each oligonucleotide is designed to have a slightly negative ΔG of folding at 37˚C and a positive ΔG at higher temperatures. For the valid oligonucleotide shown in Fig. 1, the extra sentinel bases are (5'-TCGAGC-3').

Sequenase v2.0 (USB) Extension

To reduce the occurrence of secondary structure in the *FMR1* target region, the 5' CGG repeat was removed by restriction enzyme digestion. Human genomic DNA (2.71 μg) was cleaved with Hinf I (Boehringer Mannheim GmbH) and Alu I (NEB BioLabs), 2μl each (10,000 units/ml), at 37˚C for 1 h, followed by enzyme inactivation at 80˚C for 30 min. Digested genomic DNA (0.27 μg) was mixed with 5x Sequenase Buffer (1.65 μl) and diluted with water to a total volume of 13 μl, denatured at 95˚C for 2 min, cooled on ice, and centrifuged for 10 s to avoid condensation. Encoding oligonucleotide, stored at 50μM at -20˚C, was diluted in 1x Sequenase Buffer to 450nM, heated to 95˚C for 2-5 min, and allowed to self-anneal by gradual cooling to room temperature. Self-annealed oligonucleotides may be prepared prior to use and stored at -20˚C. Self-annealed encoding oligonucleotide (2 μl) was added to the digested, denatured genomic DNA and annealed at 37˚C for 15 min. During the annealing period, a fresh dNTP mixture was prepared as follows: 3.7 μl each of 1M MgCl and 100mM DTT, 4.6μl of 100mM dNTPs, and 4.0 μl each of 5x Sequenase Buffer and water were mixed by pipetting and kept on ice. After the 15 min annealing period, the oligonucleotide-annealed DNA was placed on ice, and then microfuged for 10 s. The fresh dNTP mixture (2 μl) was added to the oligonucleotide-annealed DNA. Sequenase v2.0 polymerase (1.5 μl of 13 units/μl, diluted 1:5 in cold 1x TE, pH8 to 2.6 units/μl) was then added, resulting in a total volume of 18.5 μl. Sequenase extension was carried out at 37˚C for at least 30 min, followed by inactivation at 67°C for 30 min.

Contamination of virgin DNA or RNA prior to addition of barcodes and batch-stamps will not be detected with our methods. To minimize the possibility of such contamination, we recommend that the steps of extension and heat-inactivation of Sequenase be carried out in heat blocks without lids, rather than in closed PCR machines, coupled with standard pre-PCR precautions such as using separate rooms and dedicated filter-tip pipets and solutions.

Column Purification

Sequenase extension products were purified to remove any unincorporated encoding oligonucleotides using the Qiagen QIAquick PCR purification kit. Columns were used per manufacturer's instructions.

PCR Conditions

Amplification

In an effort to reduce contamination with PCR products from previous experiments, the underside of the heated lid of the PCR thermal-cycler was cleaned with DNAZap (Ambion) prior to each thermal-cycler use. This treatment of our thermal-cyclers reduced, but did not always prevent, contamination from previous runs.

HotStarTaq Master Mix was used per manufacturer's instructions (Qiagen) with activation at 95˚C for 15 min, followed by 35 cycles of denaturation at 95˚C for 35 s, annealing at 63˚C for 35 s, and extension at 72˚C for 40 s; final extension at 72˚C was for 4 min. Primers used were 5'-ACATGCATGTCTTCAAAGTGG-3' and 5'-GGATGCATTTGATTTCCCACGCC-3'.

Sequence Recovery

All PCR products were visualized with ethidium bromide using 1.8% agarose gel-electrophoresis; PCR products of the expected size were subcloned and analyzed with TOPO-TA Cloning Kits (Invitrogen). Sequencing reactions were carried out with fluorescent dideoxynucleotides (BigDye Terminator v3.1 Applied Biosystems) in the Comparative Genomics Center, Department of Biology, University of Washington. Each base-call was verified using the electropherogram trace.

Results and Discussion

We applied molecular encoding and standard PCR amplification to the human locus *FMR1*. Valid sequences were identified by their date- and sample-specific batch-stamp, and their unique cell- or allele-copy-specific barcodes (Fig. 1F-G). Contaminant sequences were identified by batch-stamps assigned to separate reactions (Fig. 1H). The appearance of sequences bearing inappropriate batch-stamps enabled us to trace contaminating sequences to PCR reactions run either in parallel or previously in our laboratory, even though standard precautions were used to limit contamination. With this sensitive method, contaminant sequences were detected even when control PCR reactions lacking template DNAs appeared negative on agarose gels. This result confirms our previous observations with batch-stamps and barcode sequences amplified by hairpin-bisulfite PCR (Miner et al. 2004).

In addition to distinguishing valid from contaminant sequences, our methods detect redundant sequences arising from the same cellular DNA template (Fig. 1I). Redundancy is frequent when template quantities are limited, or when large numbers of product sequences are analyzed. Redundant sequences recovered from the same PCR reaction are readily identified because they are identical in both barcode and batch-stamp. Identifying redundant sequences and removing them from data sets is especially important when analyzing mosaic samples, which occur in both genetic and epigenetic diseases (Cohn et al. 1990; Stöger et al. 1997; Wong et al. 1997). An additional benefit of our method for mosaic samples is that counting sequences with non-redundant barcodes provides a quantitative measure of template diversity.

The accurate validation of sequences requires efficient removal of unincorporated, encoding oligonucleotides, which we ensure by column purification prior to PCR. Without column purification, recoding of PCR products was readily detected. In our method, incorporation of a free oligonucleotide during PCR could replace an original barcode, producing erroneous and misleading results. We have designed two assays to test for efficient removal of unincorporated oligonucleotides. The first assay is intrinsic to each experiment. Extra bases 5' to the primer-binding site on the encoding oligonucleotide (Fig. 1A) mark any product that resulted from use of an unincorporated, encoding oligonucleotide in lieu of the bonafide PCR primer, during the last round of PCR.

Our second assay can detect and quantify recoding in any round of PCR. In control experiments, a "spiking" oligonucleotide with a distinct batch-stamp was added after Sequenase extension and before column purification. After column purification, recoding was significantly less than 1% ($P<0.05$): among 320 analyzed sequences, none was observed to have the extra 5' bases or to have the batch-stamp of the spiking oligonucleotide. In a control experiment, we verified that the spiking oligonucleotide is efficiently incorporated as a primer in PCR amplification when not removed by column purification.

The encoding concepts described here for standard PCR can markedly increase the reliability of PCR-based sequence information. Although our method adds extra costs to PCR amplification, these costs are modest when compared with those that arise from incorrect diagnoses and decisions such as those reported by Kolata (2007). We recommend that batch-stamps and barcodes be used when amplifying irreplaceable DNAs and cDNAs for forensic, clinical, single cell, and ancient DNA analyses.

Acknowledgements This work was supported by the National Institutes of Health Grants GM 53805, HD 02274, and by the Washington Research Foundation. The method presented here is included in United States Patent Application 20070020640, filed by the University of Washington. We thank Carl Bergstrom, Alice Burden, Diane Genereux, Brooks Miner, and Jessica Sneeden for helpful suggestions and discussion.

References

Cohn, D.H., Starman, B.J., Blumberg, B., and Byers, P.H. 1990. Recurrence of lethal osteogenesis imperfecta due to parental mosaicism for a dominant mutation in a human type I collagen gene (COL1A1). *Am J Hum Genet* **46**(3): 591-601.

Kolata, G. 2007. Faith in Quick Test Leads to Epidemic that Wasn't. In *The New York Times*, pp. 1.

Laird, C.D., Pleasant, N.D., Clark, A.D., Sneeden, J.L., Hassan, K.M., Manley, N.C., Vary, J.C., Jr., Morgan, T., Hansen, R.S., and Stöger, R. 2004. Hairpin-bisulfite PCR: assessing epigenetic methylation patterns on complementary strands of individual DNA molecules. *Proc Natl Acad Sci U S A* **101**(1): 204-209.

Lewontin, R.C. 1994. The Use of DNA Profiles in Forensic Contexts; DNA Fingerprinting: A Review of the Controversy Commen. *Statistical Science* **9**(2): 259-262.

Li, H.H., Gyllensten, U.B., Cui, X.F., Saiki, R.K., Erlich, H.A., and Arnheim, N. 1988. Amplification and analysis of DNA sequences in single human sperm and diploid cells. *Nature* **335**(6189): 414-417.

Miner, B.E., Stöger, R., Burden, A.F., Laird, C.D., and Hansen, R.S. 2004. Molecular barcodes detect redundancy and contamination in hairpin-bisulfite PCR. *Nucleic Acids Res* **32**(17): e135.

Stöger, R., Kajimura, T.M., Brown, W.T., and Laird, C.D. 1997. Epigenetic variation illustrated by DNA methylation patterns of the fragile-X gene FMR1. *Hum Mol Genet* **6**(11): 1791-1801.

Wong, D.J., Barrett, M.T., Stöger, R., Emond, M.J., and Reid, B.J. 1997. p16INK4a promoter is hypermethylated at a high frequency in esophageal adenocarcinomas. *Cancer Res* **57**(13): 2619-2622.