

Gene capture and random amplification for quantitative recovery of homologous genes

Laurel D. Crosby*, Craig S. Criddle

Environmental Engineering & Science Program, Stanford University, Stanford, CA 94305, USA

Received 27 June 2006; accepted 20 September 2006

Available online 30 September 2006

Abstract

The polymerase chain reaction (PCR) is instrumental in molecular analysis of microorganisms, allowing for the selective amplification of nucleic acids directly from clinical and environmental samples. However, the principles that allow for targeted amplification of DNA become a hindrance when attempting to simultaneously discriminate and quantify complex mixtures of homologous genes. Here we present a simple solution to the quantitative problem by separating the enrichment and amplification aspects of a conventional PCR reaction. In this assay, genes are enriched using a DNA oligonucleotide capture probe and subsequently amplified in a two-step random amplification protocol. In order to evaluate the quantitative aspects of the gene capture assay, we used real-time quantitative-PCR to measure initial and final concentrations of homologous genes from constructed mixtures of genomes. Upon sampling for the universal DNA-dependent RNA polymerase gene, *rpoC*, we were able to demonstrate quantitative recoveries from a mixed DNA sample despite differences in gene copy number ranging up to 4 orders of magnitude. This suggests that minority populations as low as 0.01% of the total community are represented as accurately as populations at higher abundance. These results offer new possibilities for accurately and quantitatively monitoring diverse mixtures of microorganisms.

© 2006 Elsevier Ltd. All rights reserved.

Keywords: 16S rRNA; *rpoC*; Quantitative; Random; PCR; Probes; Microarray

1. Introduction

The polymerase chain reaction (PCR) is a key step in most nucleic acid-based diagnostic techniques, allowing for selective amplification of desired sequences above the complexity of background DNA. In many cases, PCR is used to amplify homologous genes from mixtures of organisms, and downstream detection technologies offer various types of discrimination based on amplicon fragment sizes or sequence characteristics. More recently, hybridization of PCR amplicons to DNA oligonucleotide microarrays has allowed for entire communities to be surveyed in a single reaction [1,2]. However, the ability to accurately discriminate different organisms in an assay depends on the sequence resolution and characteristics of the genetic target. Likewise, the ability to estimate relative abundance is only as good as the techniques used to

extract, prepare and amplify signal from the raw samples. Regardless of DNA extraction method, the current paradigm for comprehensive microbial diagnostics and community analysis is based on PCR amplification of the ubiquitous 16S ribosomal RNA gene (rRNA). Although the PCR and the 16S rRNA gene serve as the foundation for most nucleic acid-based detection technologies, biases inherent in these tools pose a challenge for truly universal and quantitative analysis of complex samples. This paper describes a novel approach to DNA amplification that overcomes the biases associated with PCR amplification of 16S rRNA genes.

Several basic constraints must be met for PCR amplification of mixed templates: all molecules must be equally accessible; primer and template hybrids should form with equal efficiency; polymerization efficiency should be the same for all; and substrate exhaustion should affect all templates equally [3]. Of these considerations, primer and template interactions deserve much of the attention because priming sequences often differ among various

*Corresponding author. Tel.: +1 650 814 6229.

E-mail address: laurel@stanford.edu (L.D. Crosby).

groups of organisms. The efficiency of primer binding depends on nucleotide composition of the template priming site, G+C content, and various chemical and thermal parameters of the PCR reaction. In addition to variations in annealing efficiencies, the exponential increase in template copies over successive PCR cycles may contribute to error. For example, stochastic variation in amplification during the early cycles of the PCR can become exacerbated over successive cycles leading to a situation known as “drift,” although this problem can be mitigated by performing fewer cycles and combining replicate reactions [4]. Another result of exponential amplification is that primer concentrations decrease as templates increase, creating a situation where complementary DNA strands compete with primer for template binding [5]. This poses a problem for quantitative analyses in mixed samples because the amplification of different templates may saturate at different times, and the ratios of the different templates eventually converge after saturation at a common plateau. Taken as a whole, biases related to PCR amplification limit the ability to perform quantitative analyses of mixed samples unless target sequences are identified and quantified one at a time.

DNA probing technology offers a solution to the limitations of traditional two-primer PCR. By attaching a single-stranded DNA probe onto the surface of a super-paramagnetic particle or other substrate, it is possible to capture desired genes from a sample and eliminate the background of the genome [6]. This approach has the effect of enriching the gene of interest relative to background DNA, and works with a single oligonucleotide capture sequence rather than two priming sequences as required in a conventional PCR reaction. Thus, design parameters and hybridization conditions for capture probes are much less stringent than for PCR primer pairs: there are fewer constraints on the melting temperatures between oligonucleotide and template hybrids; there is no risk of forming heterodimers; and it is possible to accommodate much higher levels of degeneracy in the capture probe pool. The ability to design probes with higher degeneracy allows for comprehensive capture of protein-encoding genes, since these sequences present greater variability in the wobble positions of the nucleic acid code.

Despite the benefits of using a DNA probe to capture genes of interest, one of the limitations is that the copies of captured genes are too few to be visualized, cloned, sorted, or otherwise analyzed. This is where the PCR has an advantage for generating substantial quantities of material for further study, and to date, bead-based sequence capture with DNA probes has been used primarily as a pre-enrichment step for conventional PCR [7,8]. An alternative approach for amplifying the enriched material is to use a random PCR reaction with fully degenerate hexanucleotide primers [9]. In this case, random hexamers are used to amplify the DNA without regard to primer design or specificity of the target sequences. As a result, signal amplification may proceed with minimal bias and therefore

preserve the quantitative ratios of homologous genes in the enriched sample.

The goal of this research was to explore the quantitative aspects of gene capture and random amplification, a technique given the name *CAPRA* (Fig. 1). Quantitative PCR was used to measure ratios of homologous genes from mixed genomic DNA samples during initial, intermediate and final phases of the assay. In this case, Q-PCR served as a simple analytical tool that allowed for development and refinement of the assay conditions, independently of any specific downstream detection technology. Our intent was to develop an understanding of *CAPRA* with the longer-term goal of optimizing these steps for sequence identification and detection through DNA clone libraries and DNA oligonucleotide microarrays, respectively.

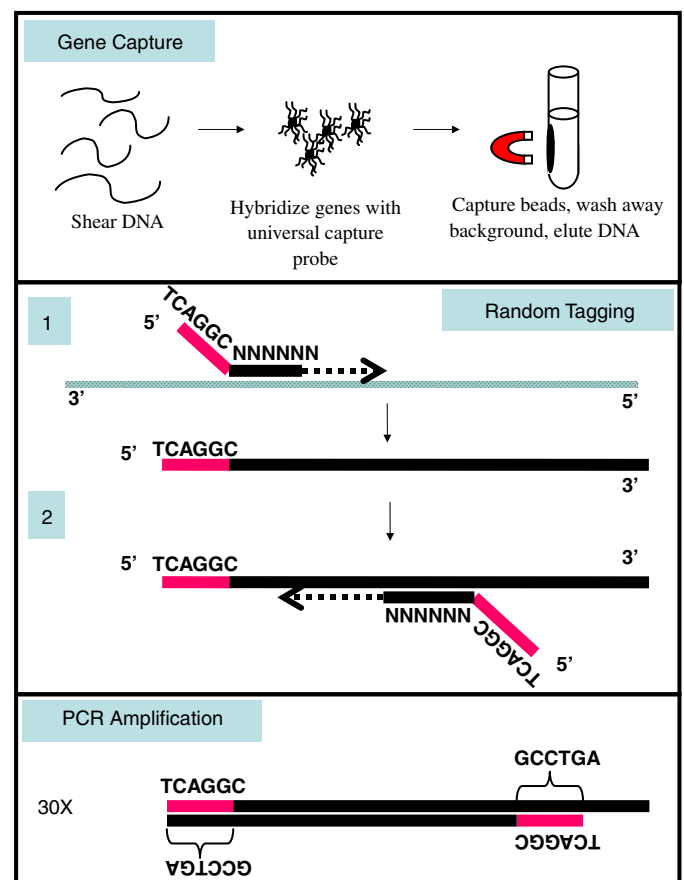


Fig. 1. Overview of gene capture and random amplification (*CAPRA*). Genomic DNA samples are sheared into fragments, hybridized with biotin-labeled oligonucleotide probes, and captured with streptavidin-coated paramagnetic particles. Captured fragments are then rinsed of background and eluted in water prior to random amplification. The random PCR protocol involves two steps: incorporation of new terminal sequences onto newly synthesized amplicons using tagged random hexamers, and a second amplification reaction with only the specific 5' portion of the tagging primer. The sequence “TCAGGC” is used here as an illustration. A minimum of two cycles are needed during the tagging step to generate the complement of the original tag sequence. This complementary sequence becomes the new binding site for the second set of amplification reactions, and subsequent amplicons have binding sites on both ends. This allows for unbiased amplification of intervening sequences, since all fragments should have identical primer binding sites.

1.1. A universal target for gene capture

Much of the current understanding of microbial diversity and phylogeny is based on the study of the 16S rRNA gene [10,11]. This gene encodes for one of the structural RNA components of the prokaryotic ribosome and has several unique features which make it a valuable target for molecular studies. For example, highly conserved regions of sequence offer universal priming sites for the polymerase chain reaction, allowing for a common set of PCR primers to be used to amplify the 16S rRNA genes from unknown organisms. Other regions of the 16S rRNA gene are more variable, and the sequence differences between organisms can be scored and calculated for determining degrees of relatedness. As a result, large databases of 16S rRNA gene sequences and a variety of molecular analytical tools have been developed that offer deeper insights into the microbial world [12].

However, other aspects of the 16S rRNA gene are less well-suited for accurate and quantitative analysis of microbial communities. For example, the discriminating power of the 16S rRNA gene is relatively coarse and limits the differentiation of more closely related species and strains. Another confounding factor is that microorganisms can have variable numbers of copies of the 16S rRNA gene, and the different copies within one organism can accumulate sequence mutations independently of each other [13,14]. When sampling 16S rRNA genes from an undefined community, the heterogeneous copies arising from one organism can lead to an overestimation of diversity. Variations in copy number also contribute to quantitative bias in DNA-based molecular assays, since organisms with higher copy number give a stronger signal relative to their population size compared to organisms with lower gene copy number [15]. A final consideration for the use of the 16S rRNA gene in community analysis is the use of so-called “universal” priming sites for the PCR. These are short stretches of conserved sequence which are important for maintaining the structural integrity of the ribosome. However, the term “universal” is a misnomer, because these priming sequences are not strictly conserved across all microbial lineages [16]. This influences the ability of the PCR to successfully amplify 16S rRNA gene sequences from organisms whose priming sites differ from the commonly used primers, and may lead to a significant underestimation of diversity in natural samples [17]. Clearly, the factors involved in over- and underestimating microbial diversity do not compensate for each other: two wrongs do not make a right.

Alternatives to the rRNA genes for molecular analysis include the various single-copy “core” genes that encode for proteins, where many have co-evolved with the ribosome [18]. These protein-encoding genes offer a finer level of sequence resolution, due to the accumulation of silent mutations in the wobble positions of the nucleic acid code. A single-copy gene also lends to the accuracy of an analysis technique, since one gene represents only one cell.

The genes encoding the β and β' subunits of the DNA-dependent RNA polymerase, *rpoB* and *rpoC*, respectively, are two examples of universal core genes that are emerging in prominence in molecular techniques. These genes have been successfully substituted for the 16S rRNA gene in assays that target specific groups of organisms, such as fingerprinting of isolates from marine ecosystems using denaturing gradient gel electrophoresis (DGGE) and the characterization of marine prokaryotes from clone libraries [19,20]. Although protein-encoding genes do not offer universal priming sites for conventional PCR, the DNA-dependent RNA polymerase genes and other conserved housekeeping genes represent valuable targets for the identification of universal DNA capture probes. For example, a comparison of *rpoC* genes available in the NCBI comprehensive microbial resource database reveals a short stretch of amino acids with strict sequence conservation. This amino acid sequence, NADFDGD, corresponds to the Mg-chelating center of the RNA polymerase enzyme [21]. Further review of the associated literature suggests that the sequence (Y/F)NADFDGD(E/Q)M(N/A) is universally conserved across all known domains of life [22]. After accounting for the degeneracy in the wobble positions of the nucleic acid code, this set of oligonucleotide capture probes has the capacity to target Eubacteria, Archaea, the Eucaryotic domain and all its kingdoms, as well as viruses containing their own DNA-dependent RNA polymerases. Among cell-based organisms, this allows for the possibility of developing a truly universal assay.

2. Materials and methods

2.1. Beads and probes

Streptavidin-coated MagneSphere paramagnetic particles were obtained from Promega in 0.6 ml aliquots and were used in a MagneSphere Technology magnetic separation stand (Promega, Madison, WI). Oligonucleotide probes were synthesized with a 5'-biotin molecule and a polynucleotide A(12) linker with a degenerate nucleic acid sequence accommodating all possible combinations of the amino acid sequence FDGDQMA (5'-TTYGAYGGN-GAYCARATGGC-3'). Probes were reconstituted to a final concentration of 10 μ M for the working stock, and 10 μ l was used per capture reaction.

2.2. Gene capture

A hybridization protocol for bead-based gene capture was modified from a method by Mangiapan et al. These authors report increased detection sensitivity by adding unbound biotin-labeled probe directly to the DNA sample, then adding streptavidin-coated particles after hybridization [7]. The method was modified as follows: DNA was first extracted from pure cultures using a Bactozyme DNA isolation kit and sheared with a Hydroshear apparatus (Genomic Solutions, Ann Arbor,

MI) to generate fragments with an average size of 4 kb (Fig. 2). In order to denature the DNA sample, 50 μ l of genomic DNA (representing 1–4 μ g of DNA) was heated to 95 °C in a heat block for 5 min then quenched in an ice slurry. Ten μ l of biotin-labeled probe was added to the cooled sample, followed by 450 μ l of DIG EasyHyb buffer (Roche Applied Science, Indianapolis, IN) and the sample was transferred to a 37 °C incubator and rotated gently for 2 h.

Streptavidin-coated paramagnetic microspheres were prepared by washing three times with 2X SSC buffer, and the final wash was removed immediately prior to the addition of sample. The samples of DNA and bound probe were added to prepared beads and allowed to incubate with gentle rotation at 37 °C for 20 min. To remove background DNA, beads were drawn aside using a magnetic separation stand and the supernatant was removed. The beads and captured material were washed a total of four times, once with 300 μ l of 2X SSC, and three times with 300 μ l of 1X SSC. The best signal-to-noise ratio for *rpoC* fragments was obtained when each wash step was allowed to incubate for 5 min with gentle rotation at room temperature. Captured DNA was eluted with three volumes of DNase-free water for a total of 400 μ l. This material was concentrated using a Montage PCR Centrifugal Concentrator (Millipore, Billerica, MA) and recovered from the membrane filter in a volume of 15–20 μ l of DNase-free water.

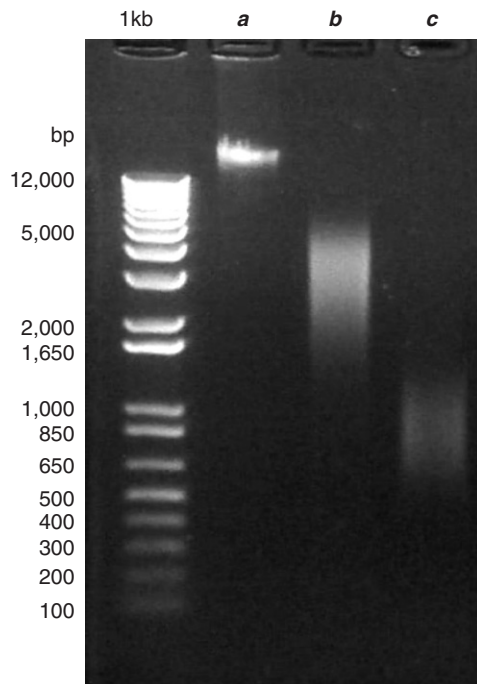


Fig. 2. DNA samples prepared for gene capture and random amplification. DNA is extracted from cells (lane a) and sheared into fragments using a Hydroshear apparatus at speed code 12 for a fragment size range of 2–5 kb (lane b). Captured fragments from this DNA pool are then randomly amplified (lane c), producing a further truncated set of fragments of approximately 500–1200 base pairs.

2.3. Random amplification

A two-step random amplification protocol was adapted from Bohlander et al., in order to amplify the captured material [9]. For the first reaction, 7 μ l of bead-captured material was mixed with 2 μ l 5X Sequenase buffer and 1 μ l of 40 pmol/ μ l Primer A (5'- GTTTCCCAGTCACGATC-3'). This mixture was heated to 94 °C for 2 min, then cooled to 10 °C and held for 5 min in order to allow the primers to anneal. At this point, 5 μ l of reaction buffer was added (1X Sequenase buffer, 300 μ M each dNTP, 15 mM dithiothreitol, 150 μ g/ml bovine serum albumin, and 0.8 U Sequenase enzyme). After adding the reagents, the temperature was ramped from 10 to 37 °C over 8 min and then held at 37 °C for an additional 8 min. This cycle was repeated a second time, except that 1.2 μ l of a 1:4 dilution of Sequenase enzyme (0.9 μ l Sequenase dilution buffer, 0.3 μ l Sequenase enzyme) was added instead of reaction buffer. After the two cycles, the reaction mixture was diluted to 60 μ l and used for Step 2.

The second step of the random amplification protocol more closely resembles a typical PCR reaction. In this case, the template from the first step was amplified with Primer B, which represents only the specific 5' portion of Primer A (5'- GTTTCCCAGTCACGATC-3'). The reaction was carried out with 5 μ l of template in a 50 μ l mixture of 1X FailSafe Premix F (Epicentre Technologies, Madison, WI), 1 μ M Primer B, and 1.25 U of Ampliqaq DNA Polymerase LD (Applied Biosystems, Foster City, CA). The cycling conditions for this reaction were 30 s at 94 °C, 30 s at 40 °C, 30 s at 50 °C and 1 min at 72 °C, for a total of 26 cycles.

2.4. Quantitative-PCR and CAPRA

Real-time Q-PCR was used to evaluate the quantitative ratios of homologous, as well as two non-homologous genes during various phases of the CAPRA assay. Initial experiments were performed with a pure culture of *Shewanella oneidensis* as a positive control, where capture efficiency was determined by measuring the signal-to-noise ratio of *rpoC* compared to the recovery of a random background gene, uridine kinase, *udk*.

Five different organisms were used to study the efficiency of gene capture for a mixture of homologous genes, including *Agrobacterium tumefaciens* (*Atu*), *Deinococcus radiodurans* (*Dra*), *Mycobacterium tuberculosis* (*Mtu*), *Shewanella oneidensis* (*Son*), and *Vibrio cholerae* (*Vch*) which served as the internal standard. Test communities were prepared with unspecified quantities of DNA from each organism, and were measured by Q-PCR to determine the initial numbers of the various *rpoC* genes. Specific Q-PCR primers were initially designed for each organism that fell within 800 bases downstream of the capture site, and this distance was based on an average genomic DNA shear fragment size of approximately 4 kb. Primers were subsequently redesigned to narrow the proximity between the capture and detection site to 400 bp (Table 1). Standard

Table 1
Primer pairs used for Q-PCR analyses

Organism (gene)	Primers
<i>Agrobacterium tumefaciens</i> (<i>rpoC</i>)	Forward: 5'-TCCAAGATCCATGAAACGACGCCT-3' Reverse: 5'-TTGGTCATTTCTGGTTGCAGGTG-3'
<i>Deinococcus radiodurans</i> (<i>rpoC</i>)	Forward: 5'-GTACTACACCAGCCGTGAGCGTAT-3' Reverse: 5'-TCTACGATACGGCGTTGTTTCGCTG-3'
<i>Mycobacterium tuberculosis</i> (<i>rpoC</i>)	Forward: 5'-GTACTACACCAGCCGTGAGCGTAT-3' Reverse: 5'-TCTACGATACGGCGTTGTTTCGCTG-3'
<i>Shewanella oneidensis</i> (<i>rpoC</i>)	Forward: 5'-GTACTACACCAGCCGTGAGCGTAT-3' Reverse: 5'-TCTACGATACGGCGTTGTTTCGCTG-3'
<i>Shewanella oneidensis</i> (<i>udk</i>)	Forward: 5'-GACCATCCCAAAGCGTTAGA-3' Reverse: 5'-ATTGCAGGAACATAGGACGG-3'
<i>Vibrio cholerae</i> (<i>rpoC</i>)	Forward: 5'-CCAACGGTCGTGTCAATCATCTTG-3' Reverse: 5'-AAGGCGAAGGTATGTACCTGACTG-3'

curves for Q-PCR were generated using a 10-fold dilution series of sheared DNA from each pure culture (speed code 12, average size 4 kb), over a dynamic range of approximately 10^2 to 10^8 copies of *rpoC*. Minimum detection sensitivity was evaluated for each primer pair, and was on the order of 10^1 copies for *V. cholerae* and *S. oneidensis*, 10^2 copies for *A. tumefaciens* and *M. tuberculosis*, and 10^4 for *D. radiodurans*. Mixtures of DNA were prepared such that initial and final concentrations fell at least one order of magnitude above these respective limits. Specificities for Q-PCR primers were determined by a factorial experiment with each pure culture and primer set, and cross-amplification was negligible among the *Dra*, *Mtu* and *Vch* primer pairs. For the *Son* primers, *A. tumefaciens* was amplified at 0.02% of the signal compared to that observed from an equivalent copy number of *rpoC* from *S. oneidensis*. The *Atu* primers also amplified *rpoC* from *S. oneidensis* at a rate of 0.003%, and with *D. radiodurans* at a much more substantial rate of 7% given an equivalent *rpoC* gene copy number. As a result, there were three instances where values for *A. tumefaciens* were confounded by cross-amplification with *D. radiodurans*, and these samples were omitted from the analysis. For combinations of *S. oneidensis* and *A. tumefaciens*, gene ratios were kept within ranges where cross-hybridization contributed to less than 2% of the total signal for a given organism.

Q-PCR reactions were prepared with 1X SYBR Green Mastermix (Applied Biosystems, Inc., CA), organism-specific primers, and 1 μ l of product from the random

amplification reaction. The Q-PCR cycling profile consisted of a 95 °C hold for 10 min, followed by 40 cycles of 95 °C for 15 s and 60 °C for 1 m. In order to evaluate the percentage recovery of each gene, the measured quantities of *rpoC* from the different organisms were normalized to the internal standard to determine an initial ratio, Q_i . For example, the initial ratio for *A. tumefaciens* relative to *V. cholerae* was expressed as Atu_i/Vch_i . After each step of CAPRA, the quantities of *rpoC* genes for each organism were again measured and normalized to the internal standard to obtain a final ratio, Q_f , with the percent recovery was expressed as Q_f/Q_i .

In addition to sampling *rpoC* genes from mixtures of pure cultures, samples were also prepared in a background of human genomic DNA (Human Female DNA, Promega, Madison, WI). A mixture of *A. tumefaciens*, *S. oneidensis*, and *V. cholerae* DNA was prepared and triplicate aliquots representing 200 ng were added to 2 μ g of human DNA sheared at speed code 12. Gene capture and random amplification was performed on each replicate, with each sample measured for *rpoC* genes before and after capture, and again after random amplification. Controls containing only human DNA were also measured with each Q-PCR primer pair, and amplification was observed with the *A. tumefaciens* primers at levels less than two orders of magnitude below the least-abundant sample.

3. Results

3.1. Optimization of the CAPRA assay

In order to develop the methodology, CAPRA was performed with sheared genomic DNA from a pure culture of *Shewanella oneidensis* as a positive control. Conditions were optimized by comparing the enrichment of the *rpoC* gene relative to a single-copy background gene selected at random, uridine kinase (*udk*). Given that these two genes are initially present in the *S. oneidensis* genome at a ratio of 1:1, the efficiency of capture was expressed as a signal-to-noise ratio of *rpoC:udk*. Under gentle wash conditions, the enrichment of *rpoC* was reproducibly observed by a factor of over 300 times compared to *udk*. After enrichment of *rpoC*, the random amplification of these two non-homologous genes was observed to occur at an equivalent rate, indicating that there was no primer preference for either gene (Fig. 3).

3.2. CAPRA for homologous genes

Measuring and discriminating homologous *rpoC* genes from a mixture of different organisms presented a challenge because of the sequence conservation within this gene and the constraints in primer design for the Q-PCR analytical technique. The results of gene capture for the mix of five organisms showed that ratios were preserved within a factor of two (Fig. 4), even though the initial

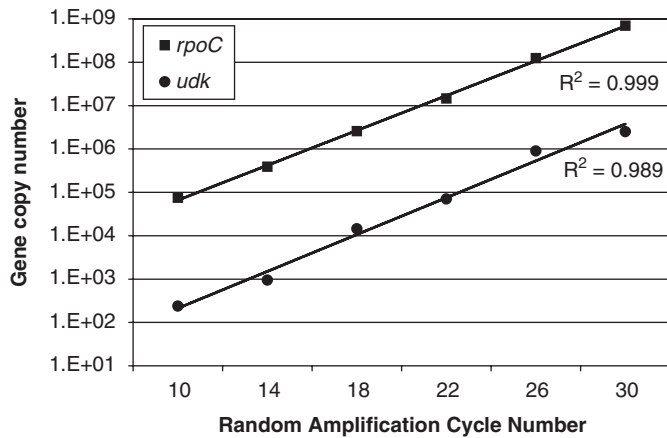


Fig. 3. Gene capture and random amplification of two non-homologous genes, *rpoC* and *udk* from a pure culture of *S. oneidensis*. Aliquots were sacrificed over successive rounds of random amplification and genes measured using quantitative PCR. Gene capture reflects an increase in the signal:noise ratio of *rpoC:udk* by over 300 times, and both genes are amplified non-specifically over four orders of magnitude at equivalent rates.

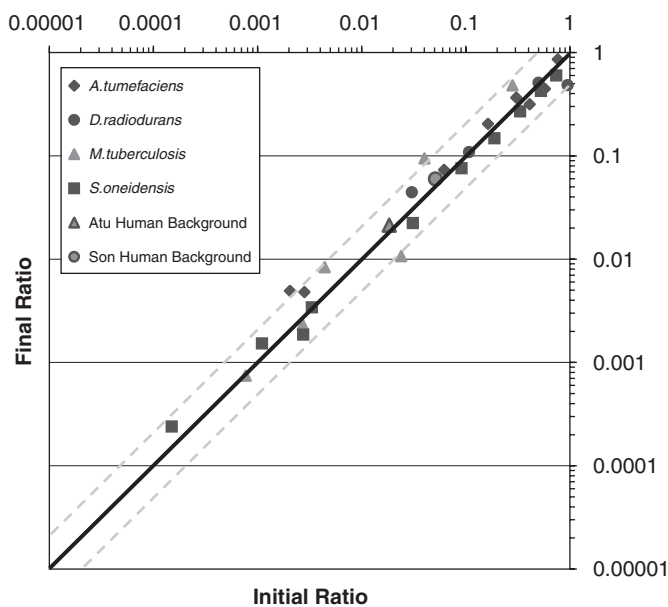


Fig. 4. Gene capture with a bead-based DNA capture probe targeting the *rpoC* gene. Initial ratios of a mixture of five different organisms were measured by quantitative PCR and compared to final ratios after gene capture with *V. cholerae* as the internal standard. The solid line with slope = 1 represents 100% recovery relative to the standard, and the dashed lines represent recovery by a factor of two above and below the expected values. For gene capture, each sample was sacrificed and measured for each of the five members. Subsequent gene capture experiments with human genomic DNA as background were added to the same data plot. These points represent three independent capture experiments for the same sample, although error bars were sufficiently small as to be obscured by the symbol.

concentrations of *rpoC* genes differed from the internal standard by up to four orders of magnitude.

After gene capture, samples were divided into three independent replicates and amplified in two steps using a random hexanucleotide PCR protocol. Aliquot samples

were sacrificed after successive rounds of random amplification and ratios of the different organisms were measured using Q-PCR. Amplification of *rpoC* genes from the *V. cholerae* standard showed exponential growth and strong agreement among replicates, but initial measurements for the other organisms were unexpectedly low and replicate samples diverged up to an order of magnitude (data not shown). This degree of variation ran counter to the observations with random amplification of the two non-homologous genes, so it seemed unlikely that concentration differences between the organisms should be a factor. A review of the Q-PCR primer design suggested a possible explanation: primers for *V. cholerae* were located within 400 base pairs of the capture site, whereas the ideal Q-PCR priming sites for the other organisms were located up to an additional 400 bases downstream. Considering that random amplification produces a truncated set of amplicons relative to the initial fragment sizes, it seemed likely that Q-PCR priming sites distal to the capture site were being lost or disrupted during the random amplification step. After redesigning Q-PCR primer pairs for *A. tumefaciens* and *S. oneidensis* to narrow the proximity between the capture and priming sites to 400 bp, amplification of the *rpoC* genes from these organisms matched the rate for *V. cholerae*. However, redesigned primers for *M. tuberculosis* were not suitable under the standard Q-PCR conditions, and primers for *D. radiodurans* were not identified; thus, it was only possible to accurately measure three of the five organisms in the mixed sample after random amplification. This again illustrates the challenge of using conventional PCR to differentiate and quantify highly conserved homologous genes.

After redesigning the primers, the full CAPRA assay was evaluated for *A. tumefaciens* and *S. oneidensis* with *V. cholerae* as the internal standard, and the ratios of *rpoC* genes and the percent recoveries, Q_f/Q_i , were calculated as before. Again, the product ratios reflected the initial template ratios despite orders of magnitude differences in the concentrations of homologous genes (Fig. 5). Interestingly, the ratios remained preserved within a factor of two whether the organisms were present at nearly 1:1 or 1:10,000 relative to the standard. This suggests that deviation in the quantitative measurements is independent of the template ratios of the different genes. Instead, the deviation from expected ratios may reflect a variety of non-systematic errors, including sample transfer via pipette, variations in gene capture efficiency, measurement error in the Q-PCR analysis technique, and/or PCR drift during random amplification. As in conventional PCR, the effects of these types of error can be dampened by combining replicate reactions. Assuming that quantitative measurements are independent of concentration and that all CAPRA experiments behaved as replicates for a given organism, the average percent recovery for *S. oneidensis* was $98 \pm 22\%$ (95% confidence interval). Likewise, averaging the ratio measurements for the *A. tumefaciens* samples gave a percent recovery of $97 \pm 39\%$ (95% confidence interval). These results demonstrate that

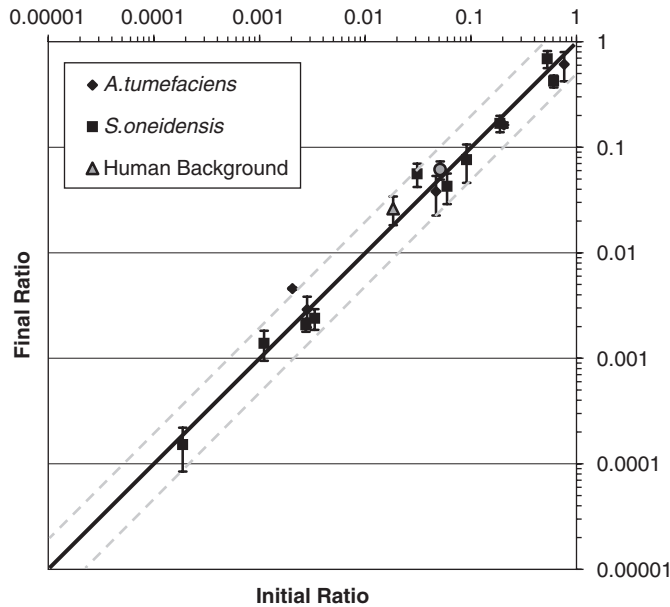


Fig. 5. Ratio measurements before and after the full CAPRA assay targeting the *rpoC* gene. Points represent one captured sample and three replicates of the random amplification reaction with error bars representing the 95% confidence interval. (*D. radiodurans* and *M. tuberculosis* were also present in the mixture, but not quantified.) Samples captured from a background of human DNA are also included, reflecting three independent random amplification reactions from each of two identical gene capture replicates. (*A. tumefaciens* is represented by the triangle, and *S. oneidensis* is represented by the circle in the human DNA background experiments.) Again, samples fell within a factor of two compared to the line representing 100% recovery.

quantitative ratios are accurately preserved in these mixtures despite differences of several orders of magnitude between the test populations and the standard.

Microbial *rpoC* genes were also captured and amplified from among a background of human genomic DNA. Three identical replicates were captured with the universal *rpoC* probe with *V. cholerae* as the internal standard, and average percent recovery for the three replicates was measured as $142 \pm 51\%$ for *A. tumefaciens*, and $106 \pm 11\%$ for *S. oneidensis* (95% confidence). After random amplification, two of the 3 replicates showed an average percent recovery of $138 \pm 27\%$ and $120 \pm 23\%$ for *A. tumefaciens* and *S. oneidensis*, respectively. However, one of the three samples failed to provide an accurate measure of recovery due to primer homodimer formation, as confirmed by gel electrophoresis.

4. Discussion

CAPRA offers a promising strategy for universal and quantitative recovery of homologous genes from complex mixtures of DNA. Support for this concept was demonstrated with the universal DNA-dependent RNA polymerase (*rpoC*) gene as a target, where mixtures of genes were recovered within a factor of two compared to their initial concentrations. Five genomes were accurately represented after gene capture, although only three of these could be

accurately measured with Q-PCR after random amplification. This reflects the conserved nature of the *rpoC* gene and the challenge in selecting specific primers using conventional PCR. In addition, the proximity between the capture site and detection sites appears to play a role in the accuracy of the detection method. This is an interesting point to consider in future experiments, especially since initial shear fragment size is a variable in the assay and longer fragment sizes may be more inclusive of distal detection sites.

Although Q-PCR allows for discrimination of a simple mixture of genomes, DNA oligonucleotide microarray hybridization technology offers a better opportunity for multiplexed discrimination of diverse mixtures of homologous genes [23]. PCR and random amplification protocols have already been used to increase the detection sensitivity in DNA microarray diagnostics, and results suggest that random PCR approaches introduce considerably less amplification bias for whole genomes compared to conventional PCR [24,25]. The addition of gene capture with DNA probes as an enrichment step, as the results of this paper suggest, allows for gene selectivity and may further enhance the sensitive and quantitative power of microarray detection by reducing the interference of background genomic DNA. Though undesired background sequences may be captured and co-amplified with the desired targets, microarray hybridization stringency will help eliminate the influence of these signals in the final analysis. Refinements of the CAPRA methodology are currently underway to optimize the characteristics of the amplification products for use with DNA oligonucleotide microarrays. Continued experimentation will help determine the quantitative potential as well as the full range of detection sensitivity for the assay.

In terms of the quantitative potential of the CAPRA assay, further work needs to be done to overcome the biases inherent in cell lysis and DNA extraction efficiency between different cell types. This bias was recently described by DeSantis et al. [26], where the vigor of bead-based extraction led to variations in peak intensity of different PCR amplicons hybridized to a 16S rRNA-based DNA oligonucleotide microarray. Considering that PCR bias may have exacerbated the error due to DNA extraction, a random amplification approach such as CAPRA may partially mitigate the distortion. Still, a conservative approach to quantitative analysis would be to limit the study to similar cell types, or to examine only the relative community changes for a given extraction method.

In an era of rapidly emerging applications in biotechnology, CAPRA represents a promising new approach for identifying and monitoring organisms. The CAPRA assay offers several advantages for selectively and quantitatively amplifying multiple homologous loci without *a priori* knowledge of PCR priming sites. Although the method was tested with a single highly conserved gene, the principles of gene capture can be applied to any target for which suitable capture and detection probes can be

identified. This allows for the identification and monitoring of genes ranging from the ubiquitous housekeeping genes to those that encode for more highly specialized functions. In addition to microbial diagnostics, CAPRA may be further developed to monitor complex community ecology and functional dynamics, as well as identify and classify genotypes in higher organisms. This ability to retrieve and amplify sequences in an unbiased manner has broad implications for developing accurate, universal, and quantitative gene-based diagnostic tools.

Acknowledgments

This work was funded by the Office of Science Biological and Environmental Research NABIR Program, U.S. Department of Energy (DOE) under Grant DOEAC05-00OR22725, and by the STC Program of the National Science Foundation under Agreement Number CTS-0120978 to the University of Illinois Urbana-Champaign. The authors wish to thank Dr. Jizhong Zhou for hosting L. C. in his laboratory, and for providing samples of *D. radiodurans*. Samples of *A. tumefaciens*, and *M. tuberculosis* and *V. cholerae*, were also graciously provided by Dr. Eugene Nester and Dr. Gary Schoolnik, respectively. The authors declare that they have no competing financial interests. Correspondence and requests for materials should be addressed to L.D. Crosby (E-mail address: laurel@stanford.edu).

References

- [1] Wilson KH, Wilson WJ, Radosevich JL, DeSantis TZ, Viswanathan VS, Kuzmarski TA, et al. High-density microarray of small-subunit ribosomal DNA probes. *Appl Environ Microbiol* 2002;68(5):2535–41.
- [2] Palmer C, Bik EM, Eisen MB, Eckburg PB, Sana TR, Wolber PK, et al. Rapid quantitative profiling of complex microbial populations. *Nucl Acids Res* 2006;34(1):e5.
- [3] von Wintzingerode F, Gobel UB, Stackebrandt E. Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiol Rev* 1997;21(3):213–29.
- [4] Polz MF, Cavanaugh CM. Bias in template-to-product ratios in multitemplate PCR. *Appl Environ Microbiol* 1998;64(10):3724–30.
- [5] Suzuki MT, Giovannoni SJ. Bias caused by template annealing in the amplification of mixtures of 16S rRNA genes by PCR. *Appl Environ Microbiol* 1996;62(2):625–30.
- [6] Jacobsen CS. Microscale detection of specific bacterial DNA in soil with a magnetic capture-hybridization and PCR amplification assay. *Appl Environ Microbiol* 1995;61(9):3347–52.
- [7] Manganapan G, Vokurka M, Schouls L, Cadranell J, Lecossier D, van Embden J, et al. Sequence capture-PCR improves detection of mycobacterial DNA in clinical specimens. *J Clin Microbiol* 1996;34(5):1209–15.
- [8] Stinear T, Davies JK, Jenkin GA, Hayman JA, Oppedisano F, Johnson PD. Identification of *Mycobacterium ulcerans* in the environment from regions in Southeast Australia in which it is endemic with sequence capture-PCR. *Appl Environ Microbiol* 2000;66(8):3206–13.
- [9] Bohlander SK, Espinosa R, Le Beau MM, Rowley JD, Diaz MO. A method for the rapid sequence-independent amplification of microdissected chromosomal material. *Genomics* 1992;13(4):1322–4.
- [10] Woese CR. Bacterial evolution. *Microbiol Rev* 1987;51(2):221–71.
- [11] DeLong EF. Marine microbial diversity: the tip of the iceberg. *Trends Biotechnol* 1997;15(6):203–7.
- [12] Maidak BL, Larsen N, McCaughey MJ, Overbeek R, Olsen GJ, Fogel K, et al. The ribosomal database project. *Nucl Acids Res* 1994;22(17):3485–7.
- [13] Farrelly V, Rainey FA, Stackebrandt E. Effect of genome size and *rnm* gene copy number on PCR amplification of 16S rRNA genes from a mixture of bacterial species. *Appl Environ Microbiol* 1995;61(7):2798–801.
- [14] Acinas SG, Marcelino LA, Klepac-Ceraj V, Polz MF. Divergence and redundancy of 16S rRNA sequences in genomes with multiple *rnm* operons. *J Bacteriol* 2004;186(9):2629–35.
- [15] Crosby LD, Criddle CS. Understanding bias in microbial community analysis techniques due to *rnm* operon copy number heterogeneity. *Biotechniques* 2003;34(4):790–4 796, 798 passim.
- [16] Forney LJ, Zhou X, Brown CJ. Molecular microbial ecology: land of the one-eyed king. *Curr Opin Microbiol* 2004;7(3):210–20.
- [17] Schloss PD, Handelsman J. Status of the microbial census. *Microbiol Mol Biol Rev* 2004;68(4):686–91.
- [18] Harris JK, Kelley ST, Spiegelman GB, Pace NR. The genetic core of the universal ancestor. *Genome Res* 2003;13(3):407–12.
- [19] Dahllöf I, Baillie H, Kjelleberg S. *rpoB*-based microbial community analysis avoids limitations inherent in 16S rRNA gene intra-species heterogeneity. *Appl Environ Microbiol* 2000;66(8):3376–80.
- [20] Toledo G, Palenik B. *Synechococcus* diversity in the California current as seen by RNA polymerase (*rpoC1*) gene sequences of isolated strains. *Appl Environ Microbiol* 1997;63(11):4298–303.
- [21] Zaychikov E, Martin E, Denisova L, Kozlov M, Markovtsov V, Kashlev M, et al. Mapping of catalytic residues in the RNA polymerase active center. *Science* 1996;273(5271):107–9.
- [22] Dieci G, Hermann-Le Denmat S, Lukhtanov E, Thuriaux P, Werner M, Sentenac A. A universally conserved region of the largest subunit participates in the active site of RNA polymerase III. *EMBO J* 1995;14(15):3766–76.
- [23] Hughes TR, Mao M, Jones AR, Buchard J, Branton MJ. Expression profiling using microarrays fabricated by an ink-jet oligonucleotide synthesizer. *Nature Biotechnol* 2001;19:342–7.
- [24] Wang D, Coscoy L, Zylberberg M, Avila PC, Boushey HA, Ganem D, et al. Microarray-based detection and genotyping of viral pathogens. *Proc Natl Acad Sci USA* 2002;99(24):15687–92.
- [25] Vora GJ, Meador CE, Stenger DA, Andreadis JD. Nucleic acid amplification strategies for DNA microarray-based pathogen detection. *Appl Environ Microbiol* 2004;70(5):3047–54.
- [26] DeSantis T, Stone CE, Murray SR, Mcberg JP, Anderson GL. Rapid quantification and taxonomic classification of environmental DNA from both prokaryotic and eukaryotic origins using a microarray. *FEMS Microbiol Lett* 2005;245(2):271–8.