doi:10.1093/nar/gks238

A robust PCR primer design platform applied to the detection of *Acidobacteria* Group 1 in soil

Jason D. Gans*, John Dunbar, Stephanie A. Eichorst, La Verne Gallegos-Graves, Murray Wolinsky and Cheryl R. Kuske

Bioscience Division, Los Alamos National Laboratory, Los Alamos, NM, USA

Received November 18, 2011; Revised January 18, 2012; Accepted February 29, 2012

ABSTRACT

Environmental biosurveillance and microbial ecology studies use PCR-based assays to detect and quantify microbial taxa and gene sequences within a complex background of microorganisms. However, the fragmentary nature and growing quantity of DNA-sequence data make group-specific assay design challenging. We solved this problem by developing a software platform that enables PCR-assay design at an unprecedented scale. As a demonstration, we developed quantitative PCR assays for a globally widespread, ecologically important bacterial group in soil, *Acidobacteria* Group 1. A total of 33 684 *Acidobacteria* 16S rRNA gene sequences were used for assay design. Following 1 week of computation on a 376-core cluster, 83 assays were obtained. We validated the specificity of the top three assays, collectively predicted to detect 42% of the *Acidobacteria* Group 1 sequences, by PCR amplification and sequencing of DNA from soil. Based on previous analyses of 16S rRNA gene sequencing, *Acidobacteria* Group 1 species were expected to decrease in response to elevated atmospheric CO_2. Quantitative PCR results, using the *Acidobacteria* Group 1-specific PCR assays, confirmed the expected decrease and provided higher statistical confidence than the 16S rRNA gene-sequencing data. These results demonstrate a powerful capacity to address previously intractable assay design challenges.

INTRODUCTION

Polymerase chain reaction (PCR)-based assays to detect and quantify microbes in environmental samples containing a complex background of microbial DNA are important tools in national defense, public health and microbial ecology. Consequently, many algorithms have been developed for PCR-assay design. Existing algorithms can be divided into four categories based on the number of target and related non-target sequences that are evaluated during design: (i) a single target sequence (1–4), (ii) multiple target sequences (5–15), (iii) a single target sequence and multiple non-targets (16–18) and (iv) multiple target sequences and multiple non-targets (19–24). The last category is the most general and most challenging problem—the group-specific assay design problem.

As more sequence data is deposited in public databases, the number of groups that can be monitored increases. Target groups include pathogens and their closest relatives (25–27), functional groups (28–30) or broad taxonomic groups (30,31). However, the growing amount of sequence data presents substantial challenges for assay design. Three central challenges are (i) the variation in the length and overlap of available sequences, (ii) the absence of target-specific signatures (i.e. oligonucleotide sequences that are present in all target sequences and absent in all near neighbors) and (iii) computational scale, determined by the number and length of target and non-target sequences. The lack of algorithms able to confront these growing challenges makes the design of many group-specific assays intractable.

Variation in the length and overlap of available sequences impedes assay design. Short sequences reduce the regions available for assays, and the lack of overlap (or only partial overlap) (32) forces the omission of data that may contain informative biological variation. For example, if 1200 bp sequences were required as input for assay design, half of the 16S rRNA gene sequences in the Ribosomal Database Project (RDP) (33) would be omitted. About 52% of sequences in the RDP (release 10, 5 April 2011) are less than 1200 bases (the full-length gene is 1400–1500 bp), and the sequences represent different regions of the 16S rRNA gene. Variation in sequence length also impedes efforts to determine the extent to

*To whom correspondence should be addressed. Tel: +1 505 667 3770; Fax: +1 505 665 3024; Email: jgans@lanl.gov
Present address:
Stephanie A. Eichorst, Department of Microbial Ecology, University of Vienna, Vienna, Austria.

which an assay or set of assays covers existing sequences. Current programs do not address the problem of variable length sequences.

The lack of target-specific signatures also restricts assay design. A signature sequence is unique to, and conserved within, the target group and thus confers assay specificity. With the exception of PRISE (22), computer programs for group-specific assay design require signature primers (19–24). This approach is attractive because it is computationally inexpensive, but it prevents the design of assays for groups that lack a unique signature. For such groups, specific assays can still be obtained by exploiting the specificity arising from the *combination* of forward and reverse primers. That is, the primers can be individually non-specific, but group-specific as a pair (Figure 1). This strategy was used to manually design a PCR assay that differentiates *Brucella abortus* from the closely related *B. suis*, *B. melitensis*, and *B. ovis* (34). The assay targets genes that are present in all *Brucella* species but are uniquely arranged in *B. abortus*. Design programs that require a signature primer would not discover this primer pair. The PRISE program (22) is capable of designing this type of assay but cannot accommodate design problems on a large computational scale (determined by the number and length of input sequences).

To address these limitations, we developed two algorithms: SeqStrap and ProSig. SeqStrap enables use of partially overlapping sequences that might otherwise be discarded, while ProSig performs assay design for different assay formats (e.g. PCR, TaqMan PCR and other probe-based PCR assays). Here, we demonstrate the sequential application of these algorithms for design of quantitative PCR (qPCR) assays specific for *Acidobacteria* Group 1. This group was chosen because it is ecologically important and technically challenging to target. The *Acidobacteria* are of ecological interest due to their high abundance in soils, ~20% of the total bacterial community (35), along with their response to pH (36–38), carbon (31,39), soil management (39–41) and elevated atmospheric carbon dioxide (42,43). The phylum contains 26 rRNA gene sequence similarity groups, labeled Groups 1 to 26 (44). *Acidobacteria* Group 1 is of particular interest because it contributes to cellulose degradation (39,45) and responds to ecosystem exposure to elevated atmospheric CO_2 (42,43), a factor in global warming.

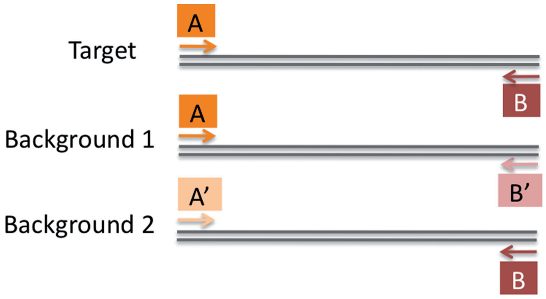


Figure 1. Non-signature primers can produce a target-specific assay. Primer sequences A and B occur in both the target and the background sequences, but only occur as a pair in target sequences. Instead of exploiting unique oligonucleotides (i.e. signatures), assays can exploit unique oligonucleotide *combinations*.

The diversity of *Acidobacteria* Group 1 illustrates the scope of the assay design problem. There is little formal taxonomic structure within Group 1. Formal structure (i.e. the elucidation of taxonomically well-defined Orders, Families and Genera) requires analyses of representative bacterial cultures. These organisms are difficult to culture and there are only 5-sequenced genomes (S. Lucas, A. Copeland, A. Lapidus, J.-F. Cheng, L. Goodwin, S. Pitluck, H. Teshima, J. C. Detter, C. Han, R. Tapia *et al.*, unpublished results; S. Lucas, A. Copeland, A. Lapidus, J.-F. Cheng, L. Goodwin, S. Pitluck, A. Zeytun, J.C. Detter, C. Han, R. Tapia, *et al.*, unpublished results; S. Lucas, A. Copeland, A. Lapidus, J.-F. Cheng, L. Goodwin, S. Pitluck, A. Zeytun, J.C. Detter, C. Han, R. Tapia *et al.*, unpublished results) (46) and 10 formally described cultured species (36,47–51) in Group 1. Consequently, diversity within the Group is known mainly from rRNA gene sequences amplified from environmental samples. Among the 1339 nearly full-length (≥1200 bp) 16S rRNA gene sequences representing *Acidobacteria* Group 1 in the RDP (33), the most divergent sequences are ~85% similar (39). Using a complete linkage clustering algorithm (33) and thresholds corresponding to 95, 90 and 85% sequence similarity, we found the number of clusters formed by the 1339 sequences at each threshold was 132, 10 and 1, respectively. In taxonomic terms, these clusters provide a lower bound estimate of the number of genera, families and orders, illustrating the extensive breadth of diversity within the group. Many failed attempts have been made over the past decade to design PCR assays for *Acidobacteria* Group 1. Here, we demonstrate the success of the SeqStrap and ProSig programs in solving this complex design problem.

MATERIALS AND METHODS

Sequences for assay design

Acidobacteria 16S rRNA gene sequences were obtained from the RDP release 10 (33) on 21 August 2010. The sequence lengths ranged from 262 to 1512 bp for the Group 1 *Acidobacteria*, and from 235 to 1697 bp for the non-Group 1 *Acidobacteria*.

Hardware

SeqStrap and ProSig were run on a 376-core, 2.5 GHz Intel Xeon-based computer cluster with a gigabit Ethernet network. SeqStrap requires a CPU that supports Streaming SIMD Extensions (SSE) version 4.1 (or higher) to compute the Smith–Waterman alignment of a single query sequence against four subject sequences in parallel using 32-bit alignment scores.

SeqStrap and ProSig software

The two algorithms are implemented in C++ and compiled with gcc to run on the Linux operating system. The MPI software library is used to parallelize the calculation on a cluster of Linux computers. Each program is invoked independently via the command line or cluster batch scheduling software. Both ProSig and SeqStrap are publicly available as open source software (GNU public license version 2) and can be downloaded from http://public.lanl.gov/jgans.

Sequence extrapolation algorithm

SeqStrap is a computationally intensive algorithm that performs iterative, pair-wise sequence extrapolation as a preprocessing step before assay design. The algorithm, outlined in Figure 2, iteratively extends each partial target sequence by adding the overhanging sequence from the most similar target sequence (illustrated in Figure 3). The most similar sequences must have an alignment score greater than a preset threshold and must be the highest scoring pair-wise alignment found by a search against all other target sequences. SeqStrap uses a conservative threshold score of 50 (see the scoring scheme described in the legend of Figure 2).

To reduce the number of unproductive sequence comparisons (i.e. alignments that do not contribute to sequence extrapolation), sequences are sorted by length in ascending order at the beginning of each iteration, and the shortest sequences are extrapolated first. Overlap alignments between pairs of sequences are computed with Smith–Waterman dynamic programming (52). Because only the alignment score and the coordinates of overhanging sequence are needed (as opposed to a detailed pair-wise alignment), SeqStrap uses a linear-space variant of the Smith–Waterman dynamic programming that only stores two rows in the dynamic programming matrix (52). This avoids the potential storage limitation in aligning longer (>10^5 bp length) nucleic acid sequences (the dynamic programming matrix size is proportional to the product of the aligned sequence lengths).

To reduce the computational burden of extrapolation with large numbers of sequences, SeqStrap exploits two levels of computer parallelism to compute the independent pair-wise alignments. At the lowest level, alignments between the query sequence (to be extrapolated) and the subject sequences (sources of extrapolated sequence) are computed in parallel (53). By using the 128-bit SSE available on modern Intel and AMD CPUs, four independent alignments (using 32-bit scores) can be computed in parallel. At the highest level, all sequences are uniformly distributed between available CPU cores in a cluster computer using the MPI parallel toolkit. For each query sequence, every core computes the alignments between the query and the locally stored subject sequences.

Even with this parallel implementation, 1 week on a 376-core cluster represents a significant investment of computational resources. The expended computer time reflects our choice of problem (16S rRNA gene sequences, one of the most abundant sequence types in public databases) and the desire for highly accurate, overlap sequence alignments (using Smith–Waterman dynamic programming). The required running time of the extrapolation algorithm is difficult to predict *a priori* and depends on several factors, including number of sequences, distribution of sequence lengths, pair-wise similarities between sequences and cluster hardware. However, a significant speed up in algorithm performance could be obtained by replacing the Smith–Waterman alignment algorithm with a faster (likely heuristic), sequence overlap alignment

```
do {

     sort sequences in ascending order by length

     for each query sequence in sequences {

          if(query sequence length < maximum sequence length){

               compare query against all other sequences

               find best scoring Smith Waterman alignment with
               5′ and/or 3′ sequence that overhangs the query

               if(best alignment score > score threshold){

                    extrapolate query by adding overhanging
                    sequence

                    break out of for-loop
               }
          }
     }

} while still extrapolating
```

Figure 2. Pseudocode description of the SeqStrap algorithm. Sequence alignments are computed using Smith–Waterman dynamic programming with the following scores: 2 for a match, −3 for a mismatch, −5 for gap existence and −2 for gap extension. No penalty is assigned for overhanging ends. When multiple sequence alignments produce the same highest score, the alignment with the largest amount of overhanging sequence is selected. When the amount of overhanging sequence is also the same, the order of the sequences in the input file is used to break the tie (to insure reproducible results). Examples of overhanging sequence are shown in Figure 3. Two heuristic parameters are required: a score threshold and a maximum sequence length. The score threshold prevents spurious matches from creating sequence chimeras. Requiring a maximum sequence length was empirically found to prevent infinite extrapolation loops, e.g. terminal mismatches leading to an endless cycle of sequence A extrapolating sequence B, B extrapolating sequence C, and C extrapolating sequence A. All extrapolations used a score threshold of 50 and a maximum sequence length of 1500 bp. Because the maximum length criterion is applied before extrapolation, it is possible to produce extrapolated sequences that exceed the maximum sequence length.

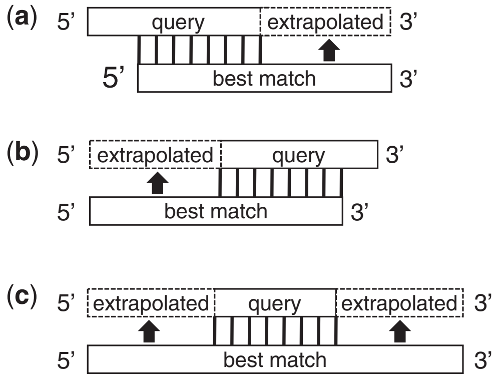


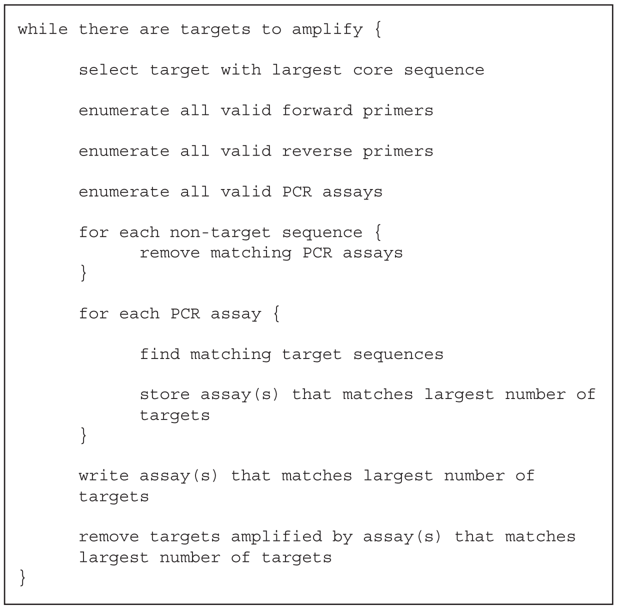
Figure 3. There are three possible ways to extrapolate a query sequence: (**a**) add the 3′-overhanging sequence from the best match, (**b**) add the 5′-overhanging sequence from the best match, and (**c**) add both the 5′- and 3′-overhanging sequence from the best match.

algorithm. The open source implementation of the SeqStrap algorithm can serve as a benchmark for future algorithmic improvements that trade alignment accuracy for alignment speed.

PCR primer design algorithm

ProSig designs PCR, TaqMan PCR and probe-based group-specific assays by large scale assay enumeration followed by subtraction of enumerated assays that potentially cross-react with non-target sequences (Figure 4). A greedy set-coverage solver then identifies the minimal sets of assays required to detect all targets. Multiple sequence alignments and signature oligonucleotides are not used in the design process. Large numbers of target sequences (ranging in size from genes to bacterial genomes) can be processed. In addition, because each forward and reverse primer will participate in multiple assays, we only perform computationally expensive, pair-wise sequence alignments once for each primer and template sequence. Finally, we exploit parallel computing to distribute both the required storage and sequence comparison calculations across a cluster of computers.

Assay enumeration is computationally demanding, but feasible with a mid-sized cluster (i.e. hundreds of cores). When enumerating assays from a selected target sequence, all valid forward and reverse primers from the core (i.e. non-extrapolated) region of the selected targets are enumerated. Valid primers are defined by user-specified constraints in length, primer-template melting temperature, primer-hairpin melting temperature and other common guidelines for primer design (Table 1). Melting temperatures are calculated with a standard nearest-neighbor thermodynamic model (54) that provides both melting temperature and free energy change given a pair-wise sequence alignment that can include mismatches and gaps.



```
while there are targets to amplify {

      select target with largest core sequence

      enumerate all valid forward primers

      enumerate all valid reverse primers

      enumerate all valid PCR assays

      for each non-target sequence {
            remove matching PCR assays
      }

      for each PCR assay {

            find matching target sequences

            store assay(s) that matches largest number of
            targets
      }

      write assay(s) that matches largest number of
      targets

      remove targets amplified by assay(s) that matches
      largest number of targets
}
```

Figure 4. Pseudocode description of the greedy PCR-assay enumeration algorithm. Core sequence refers to the original, non-extrapolated sequence. Primer and assay enumeration used the target parameters listed in Table 1, while the removal of assays used the non-target parameters listed in Table 1.

The potential number of forward and reverse primer candidates is of order 2(nL), where L is the size of the core target sequence (e.g. ~1000 bp for 16S rRNA gene targets), and n is the number of allowed primer lengths (typically less than 10 different lengths ranging from 20 to 30 bp). While each primer can potentially participate in multiple PCR assays, the computationally expensive primer-template and primer-hairpin melting temperatures are only calculated *once* for each primer. When computing the primer-template melting temperature, only the perfect match is considered. Combining the enumerated primers, the number of possible PCR assays is order ($\Delta A n^2 L$), where ΔA is the range of allowed amplicon sizes (i.e. largest amplicon size—smallest amplicon size; typically <200 bp). The calculation of the primer-dimer melting temperature for the forward and reverse primers must be computed for every assay. This calculation dominates the enumeration of valid assays from a single target sequence because the number of primer combinations is far greater than the number of primers. In practice, the number of melting temperature calculations is less than order ($\Delta A n^2 L$), because only a fraction of possible primers satisfy the constraints shown in Table 1, but a general computational framework is still needed that can perform ~10^7–10^{10} independent melting temperature calculations per target sequence. The ProSig program provides this framework by distributing the melting temperature calculations in parallel on multiple CPUs in a computer cluster.

The computational cost of performing PCR-assay subtraction (i.e. matching and removing assays that potentially cross-react with non-target sequences) is of order ($2nL$), which is far less demanding than the assay enumeration step. The melting temperature is calculated for each primer bound to each non-target sequence. Since primer binding can potentially occur anywhere in a non-target sequence, melting temperature calculations are initiated for all non-target sequence subregions that share at least

Table 1. PCR assay design constraints

Allowed target amplicon lengths	70–300 bp
Allowed primer lengths	18–28 bp
Allowed target primer T_m	59–63°C
Maximum primer hairpin T_m	40°C
Maximum primer dimer T_m	40°C
Minimum target primer 3′ clamp	5 bp
Maximum non-target primer T_m	45°C
Maximum non-target amplicon length	1000 bp

The constraints used to (a) enumerate PCR primers and assays from target sequences and (b) subtract PCR assays that matched non-target sequences. An assay matched a non-target sequence if both primers bound in the correct orientation, with melting temperatures (T_m) ≥ 45°C, and produced an amplicon ≤1000 bp. All melting temperatures were computed by a thermodynamic alignment (55) between primer and template, using the nearest neighbor parameters of SantaLucia (54), and assuming a primer concentration of 9×10^{-7} M and a salt concentration of 0.05 M. Target match criteria are more stringent that non-target match criteria.

six consecutive bases of perfect complementarity with a given primer (identified by a hash table lookup). All non-target sequence binding sites with melting temperatures above the user-defined threshold are stored for each primer. Assays that could conceivably produce amplicons with non-target sequences are rejected. Assay rejection occurs if two primer binding sites occur in the correct orientation, within the allowed non-target distance, and with melting temperatures above the non-target T_m threshold (Table 1). Also, at least one of the primers must have one or more 3′ terminal bases that are perfectly matching to the non-target sequence to be suitable for polymerase extension. If both primers have 3′ terminal *mismatches* to the non-target sequence, then the assay is *not* rejected. All possible primer pair combinations are tested (i.e. forward/reverse, forward/forward and reverse/reverse). As with the enumeration step, the assay subtraction step is performed in parallel for each non-target sequence. Only those assays that survive the subtraction process for the *i*-th non-target sequence are searched against the *i*+1 non-target sequence.

When searching candidate assay signatures against multiple non-target sequences, we sorted the non-target genomes by degree of similarity to the target genome (most similar first) in order to remove the largest number of candidate signatures as early in the calculation as possible. Because candidate signatures that match a non-target sequence are removed from further consideration, this strategy reduces the number of signature searches that must be performed for subsequent non-target sequences. For the purpose of prioritizing non-target genomes for signature subtraction, genome similarity is defined by the magnitude of the difference between the normalized dinucleotide composition vectors for each genome [similar to (56), but using vectors of raw dinucleotide counts normalized to length one].

After assay subtraction, all remaining assays are predicted to be specific to one or more target sequences within the group. Identification of the smallest number of assays required to amplify all target sequences in the group is a standard set coverage problem (11). We used a standard greedy heuristic solution (57); after every iteration of enumeration and subtraction for a selected target sequence, the algorithm stores the assay(s) that amplify the largest number of previously unamplified target sequences. Target sequences amplified by previously stored assays are ignored. Not every target sequence is guaranteed to yield a specific assay (e.g. if the same sequence appears in both the target and non-target categories, then none of the enumerated assays will survive non-target subtraction).

The algorithm uses a process loop that consists of (i) selecting a previously unamplified target sequence, (ii) enumerating valid assays, (iii) subtracting assays that match non-target sequences, and (iv) finding the assay (or assays) that amplifies the largest number of previously unamplified targets. This process loop continues until all targets have been selected or amplified (or the process is manually terminated).

This approach fails in easily recognized modes when challenged with mislabeled sequences. If a target sequence is accidentally included in the non-target category, it will be ‘first in line’ for assay subtraction by virtue of its high similarity to other target sequences, and will quickly eliminate the majority of assay candidates. If a non-target sequence is accidentally included in the target category, its assays will be subtracted by other non-target sequences. Any assays that do survive (due to unique features in the mislabeled sequence) will be unlikely to detect other correctly labeled target sequences, and the set coverage algorithm will require a separate assay just to detect the mislabeled sequence. If a sequence that is dissimilar to both target and non-target sequences is included in the target category, it will generate a large number of assays that survive the non-target subtraction step but only detect the dissimilar sequence. Sequences that are dissimilar to both target and non-target sequences and are included in the non-target category will not affect the output (since they will not eliminate any assays during the subtraction step).

qPCR assays

Assays were performed with Biorad iQ SyBr Green Supermix and three primer sets designed for specific detection of *Acidobacteria* Group 1 (Table 2). Primers were obtained from Invitrogen. Each 25 µl qPCR assay contained primers at 0.2 µM. Cycling conditions were as follows: 94°C for 5.0 min; 40 cycles of 94°C for 15 s, 65°C for 30 s; a melt curve of 91 cycles, 30 s each, ramping 0.5°C per cycle from 50.0 to 95.0°C; 4.0°C storage. Standard curves for primer sets AcidoG1_8.1 and AcidoG1_8.17 were generated with purified, genomic DNA from *Acidobacterium capsulatum* ATCC 51196 (hereafter *A. capsulatum*). A standard curve was not obtained for primer set AcidoG1_8.2 because it does not amplify *A. capsulatum*. Assays were applied to soil DNA samples from a field experiment in Rhinelander, Wisconsin, described in (42).

16S rRNA gene clone libraries

Amplicons from triplicate qPCR reactions for each of three replicate soil DNA samples were pooled (i.e. n = 9

Table 2. PCR primer pairs for the specific detection of *Acidobacteria* Group 1 16S rRNA gene sequences

Name	5′-Forward-3′	5′-Reverse-3′	Amplicon length (bp)
acidoG1_8.1	GAACCTTACCTGGGCTCGAAA	GTGCTCAACTAAATGGTAGCAACTG	214
acidoG1_8.2	GGTGCGTGGAATTCCCGG	GCGGATTGCTTATCGCGTTAG	229
acidoG1_8.17	CCCTTGGGACGTAAACTCCTT	TTCCACGCACCTCTCCCA	306

reactions), cloned and sequenced to evaluate assay specificity. Amplicons were purified (Qiagen QIAquick PCR Purification Kit) prior to cloning (Invitrogen TOPO TA Cloning Kit). For each qPCR primer set, 192 clones were picked and the cloned 16S rRNA gene fragments were bi-directionally sequenced with M13 primers.

Sequence processing

Assembled sequences were visually inspected in Sequencher v4.7 (Ann Arbor, MI) to confirm the sequences were full length, as indicated by the presence of forward and reverse primer sites. A total of 149, 158 and 163 useable sequences were obtained for primer sets AcidoG1_8.1, AcidoG1_8.2, AcidoG1_8.17, respectively. Sequences were aligned in SILVA (58). Aligned sequences were compiled in a single database in ARB (59).

Phylogenetic specificity

Phylogenetic placement of the amplicon sequences from qPCR assays was determined in two ways. First, the sequences were classified by an automated classifier in the RDP (60). Secondly, the sequences were added by quick Parsimony to a phylogenetic guide tree in ARB. The guide tree was generated with 145 nearly full-length (ranging from ~1000 to 1500 bp) reference sequences representing *Acidobacteria* Groups 1, 3, 4, 5, 6, 7, and 8. *Archangium gephyra* (β-Proteobacteria, AB218222), *Solimonas soli* (γ-Proteobacteria, EF067861), *Chitinibacter tainanensis* (β-Proteobacteria, AY264287) and *Leeia oryzae* (β-Proteobacteria, DQ280369) were used as additional out groups. The maximum likelihood algorithm (RAxML) in ARB (59) was used to generate the guide tree with a 70% base frequency filter generated with *Acidobacteria* reference sequences representing Groups 1, 3, 4, 5, 6, 7 and 8. Amplicon sequences from the *Acidobacteria* Group 1 assays were added to the guide tree using the ARB parsimony algorithm, since the amplicons were short and represented different regions of the 16S rRNA gene.

qPCR fold-change analysis

The fold-change in relative abundance of groups targeted by qPCR assays was calculated as follows: (amplification efficiency)$^{[(\text{mean CT ambient})–(\text{mean CT elevated CO2})]}$. In this calculation, the amplification efficiency term was computed from the slope of the PCR standard curve generated with *A. capsulatum* using $e^{(-1/\text{slope})}$. The amplification efficiency was 1.85 and 1.94 for assays AcidoG1_8.1 and AcidoG1_8.17. The amplification efficiency of the third assay, AcidoG1_8.2, was not determined because the assay does not amplify *A. capsulatum* (the only available genomic DNA for the test). Therefore, the amplification efficiency of the first assay, AcidoG1_8.1, was used to approximate the fold-change calculations for AcidoG1_8.2. Similar results were obtained when using the amplification efficiency of AcidoG1_8.17 as a substitute value.

Sequence deposition

The 470 amplicon sequences from the three assays are included in both the supplementary online material and the ProSig software package.

RESULTS AND DISCUSSION

Sequence extrapolation

To maximize the number of sequence inputs for PCR-assay design, 8430 *Acidobacteria* Group 1 sequences and 25254 other *Acidobacteria* 16S rRNA gene sequences ranging from 235 to 1697 bp in length were separately processed with SeqStrap. Sequence extrapolation took about 1 week on a 376-core cluster. Extrapolation increased the length of 96% of the *Acidobacteria* Group1 sequences and 94% of the other *Acidobacteria* sequences. Extrapolation increased the average length of target sequences from 759 to 1469 bp and the average length of non-target sequences from 826 to 1512 bp. Sequence extrapolation occurs only between a sequence and the single most similar matching sequence, identified by the largest pair-wise sequence alignment score (above a set threshold) among the pool of sequences to be extrapolated. The default alignment-scoring scheme and the minimum allowed score for sequence extrapolation (Figure 2) allows extrapolation between pairs of sequences with (gap free) sequence identities >60% (in the long sequence alignment limit). This constraint reduces, but does not eliminate, the generation of chimeric artifacts by extrapolation. However, since our group-specific assay design process works by enumerating candidate assays only from the *unextrapolated* portion of target sequences, chimera-specific assays will not arise. It is still possible for chimeric sequences to cause a reduction in predicted assay coverage (i.e. the number of target sequences detected by a given assay). To our knowledge, the SeqStrap algorithm is a unique approach that overcomes the challenge of fragmentary data in sequence databases and, thus, maximizes the data that can be exploited for assay design. In our pipeline, the extrapolated sequences are used during the design process to assess the extent to which assays can amplify existing sequences.

Design of PCR assays specific for *Acidobacteria* Group 1

ProSig was used to enumerate target-specific assays. The *Acidobacteria* Group 1 sequences did not have a specific signature that could be used for a PCR assay. The lack of a signature and the large number of sequences made this target group a suitable test of ProSig's design capability. The cumulative fraction of *Acidobacteria* Group 1 sequences covered by the assays was monitored in real-time by computationally searching the assays against the sequences extrapolated by SeqStrap, and for comparison, to the original, unmodified (i.e. unextrapolated) sequences. The first four assays were predicted to cover ~50% of the extrapolated *Acidobacteria* Group 1 sequences, whereas about 15 assays were predicted to cover 50% of the unextrapolated sequences (Figure 4). This illustrates one of the challenges posed by use of unextrapolated sequences that vary substantially in length and overlap. After the first 40 assays (providing ~95% predicted cumulative coverage), the predicted coverage of *Acidobacteria* Group 1 extrapolated sequences plateaued (Figure 5). We terminated the ProSig run after generating 83 assays because further improvements in predicted coverage were marginal (Figure 5). The 83 assays were predicted to collectively cover ~98% of the *Acidobacteria* Group 1 sequences and required ~48 h to compute on our 376-core cluster. The forward primers for the assays collectively targeted about 24 distinct locations in the 16S rRNA gene. The reverse primers targeted about 19 locations. The inability to detect the entire target group with a single assay emphasizes the difficulty of the assay design problem. Similar results were obtained using the *Actinomycetales* as a target group (data not shown), demonstrating that *Acidobacteria* Group 1 is not unique in posing a design challenge, and emphasizing the need for a robust, flexible design platform like ProSig. The ability of ProSig to predict a minimal set of assays to cover *Acidobacteria* Group 1 illustrates its flexibility and value.

Three *Acidobacteria* Group 1 assays were chosen for experimental testing. To find the optimum subset of three assays predicted to collectively amplify the largest fraction of extrapolated target sequences, a brute force search was performed with all 91 881 possible three-assay combinations from the pool of 83 target-specific assays [i.e. the number of three assay combinations = 83!/(3! × 80!)]. The Perl script used to perform this search is included with the ProSig software. It is not possible to determine an optimal subset of assays simply by examining Figure 5, because each assay may detect overlapping subsets of the target taxa. In addition, because the number of possible assay combinations grows rapidly, the brute force solution to the set-coverage problem is only practical for small combinations (i.e. less than 5). The optimum combination of three assays was AcidoG1_8.1, AcidoG1_8.17 and AcidoG1_8.2 (Table 2). Using *E. coli* numbering, the assays, respectively, target the following three regions of the 16S rRNA gene: basepair 906 to 1071 (between variable regions V5 and V6), bp 373 to 607 (encompassing V4) and basepair 596 to 813 (encompassing V3), respectively. The three assays were predicted to collectively cover 42% of the extrapolated set of 8430 *Acidobacteria* Group 1 sequences. The predicted clade coverage of the three assays is shown in Supplementary Figure S1. None of the primers in the three assays were 'signature' primers, illustrating the unique ability of ProSig to design assays for target groups that lack group-specific signatures.

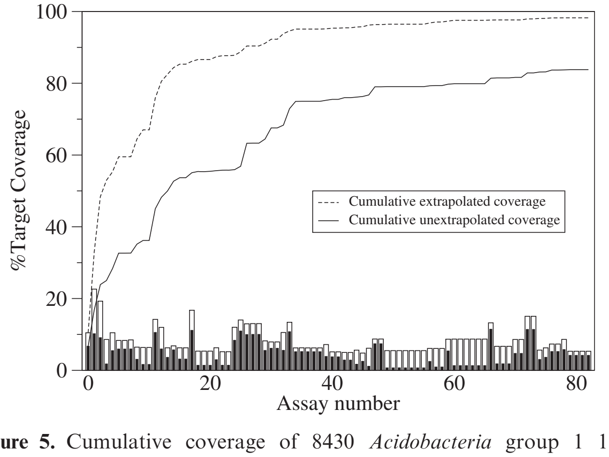


Figure 5. Cumulative coverage of 8430 *Acidobacteria* group 1 16S rRNA gene sequences by 83 target-specific PCR assays. Coverage is the percentage of the extrapolated or unextrapolated (unmodified) target sequences amplified (*in silico*). The bar height represents the extrapolated (white bars) or unextrapolated (black bars) target coverage of individual assays. The dashed and solid lines represent the cumulative extrapolated and unextrapolated target coverage, respectively. PCR assays are plotted along the *x*-axis in the order of discovery. The maximum cumulative coverage is 98% for the extrapolated targets and only 84% for the unextrapolated targets. The coverage of individual assays ranges from a high of 23% to a low of 5% for the extrapolated targets, and from a high of 11% to a low of 1% for the unextrapolated targets.

Validation of assay specificity

The specificity of the assays was confirmed by sequencing PCR amplicons derived from soil samples. Soil is an excellent test-bed for assay specificity because the hyperdiversity of species in soil microbial communities (61,62) increases the opportunity for PCR primer cross-reactivity, and thus, loss of specificity. The RDP classifier (60) identified nearly all of the 149 (for AcidoG1_8.1), 158 (for AcidoG1_8.2) and 163 (for AcidoG1_8.17) sequenced amplicons from each assay as *Acidobacteria* Group 1. The exceptions were 16 sequences from AcidoG1_8.1, which could not be classified with a confidence score >70%. This result was not surprising because the region of the 16S rRNA gene targeted by AcidoG1_8.1 does not provide good resolving power for the RDP classifier (60). More reliable classification was obtained by aligning sequences using a 16S rRNA gene-specific alignment strategy followed by reconstruction of a phylogenetic tree (60). When placed in a phylogenetic tree, all of the amplicon sequences fell within *Acidobacteria* Group 1 (Figure 6), confirming the predicted specificity of the assays.

Application to microbial ecology

The three group-specific assays were used to evaluate the response of *Acidobacteria* Group 1 in a field experiment focused on terrestrial ecosystem responses to a decade of elevated atmospheric CO_2 (42). The field experiment included three replicate field plots under ambient or elevated atmospheric CO_2 conditions. A composite soil sample was obtained from each plot, yielding six samples total. Exploratory Sanger-based, 16S rRNA gene surveys (about 270 sequences each) of the six samples showed a 2-fold decrease in the relative abundance of sequences classified as *Acidobacteria* Group 1 in plots under elevated CO_2 (Figure 7), but the difference was not statistically significant by a pair-wise *t*-test (42). Subsequently, Pyrotag-based, single subunit amplicon libraries (100-fold larger that the Sanger surveys) of the same samples showed a significant ($P = 0.017$), 3-fold decrease in *Acidobacteria* Group 1 under elevated CO_2. To investigate further, the three qPCR assays, AcidoG1_8.1, AcidoG1_8.17 and AcidoG1_8.2 were

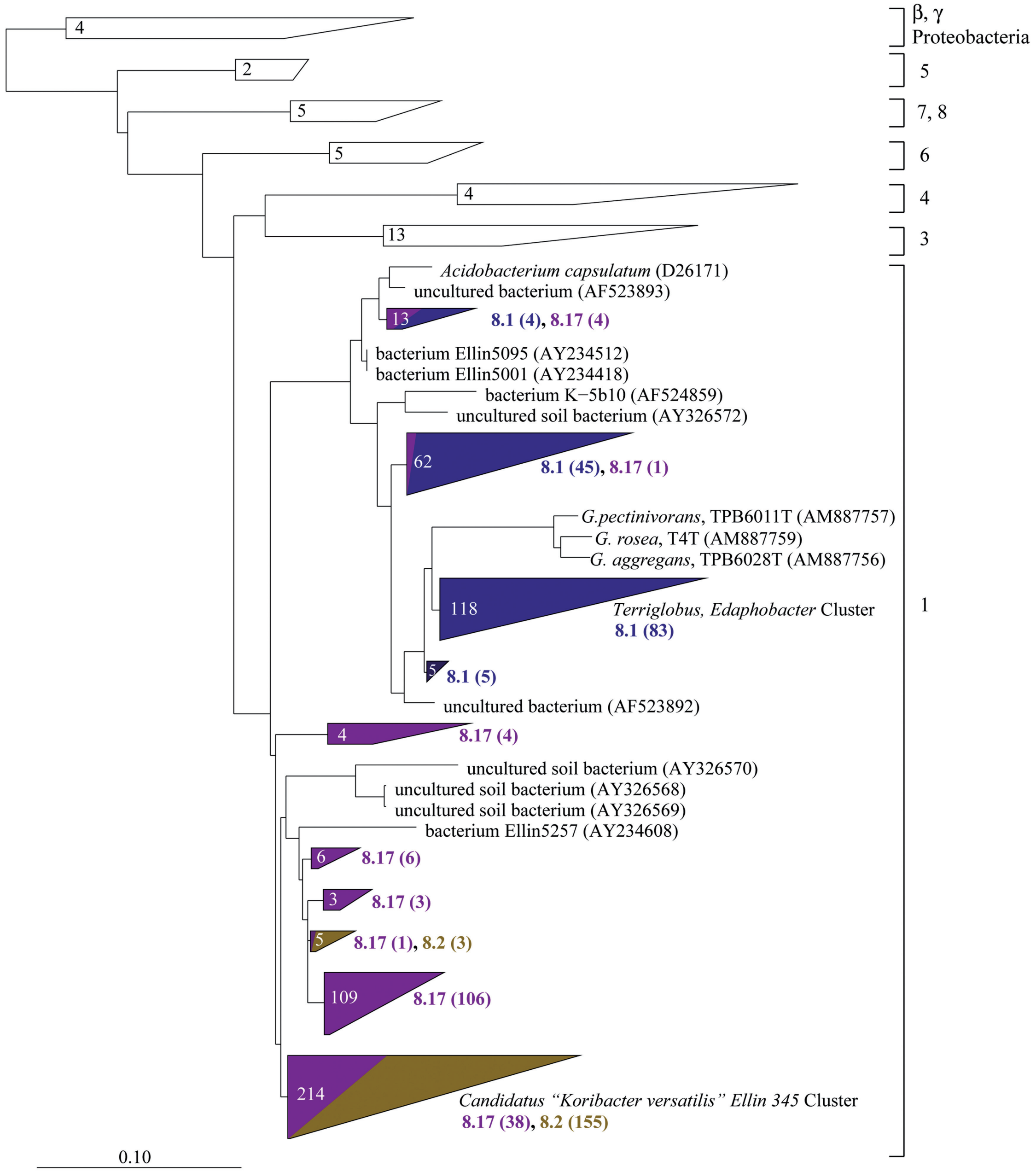

4
2
5
5
4
13
β, γ Proteobacteria
5
7, 8
6
4
3
Acidobacterium capsulatum (D26171)
uncultured bacterium (AF523893)
13
8.1 (4), 8.17 (4)
bacterium Ellin5095 (AY234512)
bacterium Ellin5001 (AY234418)
bacterium K−5b10 (AF524859)
uncultured soil bacterium (AY326572)
62
8.1 (45), 8.17 (1)
G.pectinivorans, TPB6011T (AM887757)
G. rosea, T4T (AM887759)
G. aggregans, TPB6028T (AM887756)
118
Terriglobus, Edaphobacter Cluster
8.1 (83)
5
8.1 (5)
uncultured bacterium (AF523892)
1
4
8.17 (4)
uncultured soil bacterium (AY326570)
uncultured soil bacterium (AY326568)
uncultured soil bacterium (AY326569)
bacterium Ellin5257 (AY234608)
6
8.17 (6)
3
8.17 (3)
5
8.17 (1), 8.2 (3)
109
8.17 (106)
214
Candidatus "Koribacter versatilis" Ellin 345 Cluster
8.17 (38), 8.2 (155)
0.10

Figure 6. Maximum-likelihood tree of *Acidobacteria* 16S rRNA gene sequences. Groups identified as 1, 3, 4, 5, 6, 7 and 8 are indicated to the right of the group. Sequences obtained from soil by PCR amplification with *Acidobacteria* Group 1 primer sets are indicated by color. The total number of sequences in each cluster (including both sequenced amplicons and reference sequences) is labeled at the cluster node. The number of sequenced amplicons for each respective PCR assay is shown in parentheses adjacent to the colored wedge. Wedges are colored to show the approximate fraction of amplicon from each PCR assay in the cluster (blue for 8.1, gold for 8.2 and purple for 8.17). Clusters containing reference sequences from cultured isolates are indicated. The scale bar indicates 0.10 changes per nucleotide.

applied to the same samples. Although the three assays do not detect all members of *Acidobacteria* Group 1, we expected the assays would cover a sufficient number of species to capture the Group 1 responses observed in the 16S rRNA gene surveys. With AcidoG1_8.1, there was no difference in abundance between ambient and elevated CO_2 samples. In contrast, ~5-fold decreases in abundance under elevated CO_2 were measured with AcidoG1_8.17 and AcidoG1_8.2, and the differences were significant ($P = 0.0083$, $P = 0.0023$, respectively). The results demonstrate the value of the assays in validating and further characterizing results obtained from broader survey techniques. The ability to identify responsive target groups from survey data and rapidly follow-up by designing and applying group-specific assays is an important capability in microbial ecology.

Other software

We are aware of only one other program that uses an approach comparable to ProSig. Like ProSig, the PRISE

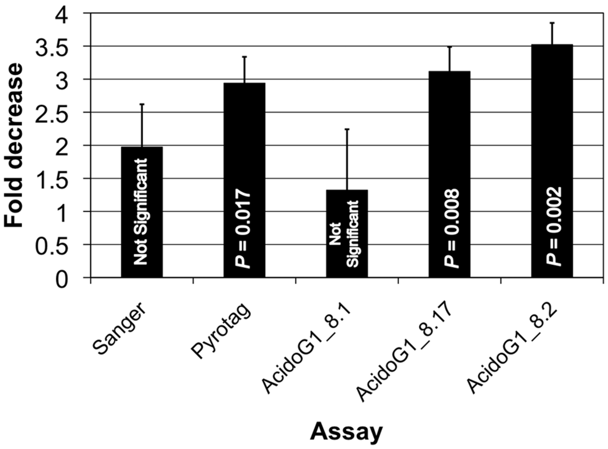


Figure 7. Decrease in the relative abundance of *Acidobacteria* Group 1 in a field experiment measured by different techniques applied to six soil DNA samples from control (ambient atmospheric CO_2) and treated (elevated atmospheric CO_2) plots. The Sanger estimate was obtained from surveys of about 300 16S rRNA gene sequences from each soil DNA sample. The Pyrotag estimate was obtained from surveys of about 50 000 16S rRNA gene sequences from each sample. The variation in the estimates is presumed to arise from the increasing quality of the assays (from left to right) and the specific members of *Acidobacteria* Group 1 sampled by each technique.

program permits discovery of specific PCR assays composed of non-signature primers, but the program has several limitations that severely constrain its capability. The program performs a limited enumeration of valid, target-sensitive PCR assays. Screening the enumerated assays against all non-target sequences then identifies target-specific assays. However, this program requires a greedy enumeration, ignoring primers that do not match a significant fraction of the target sequences. When this constraint is omitted, the program exhausts the available memory, even for small problem sizes (i.e. a small number of target and query sequences). Thus, this program is suitable only for problems in which most of the targets can be detected by an assay.

CONCLUSION

We developed a computational methodology that exploits available nucleic acid sequence information at an unprecedented scale and level of complexity to develop robust PCR-based detection assays. This methodology enables the design of group-specific assays for a wide range of applications in microbial ecology, public health and agricultural safety. The last 20 years have witnessed an exponential growth in the number of available nucleic acid sequences. Improved computational tools that can scale with the growth of sequence information are still needed. The SeqStrap and ProSig help address this need. The software has been released as open source to facilitate future improvements in parallel scalability and novel assay design algorithms.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Figure 1.

FUNDING

US Department of Homeland Security [HSHQDC-08-X-00763, HSHQDC-07-X-00908, HSHQDC-08-X-00793] (to M.W. and J.G.); US Department of Energy (Science Focus Area grant to C.R.K. and J.D.); Office of Biological and Environmental Research (2009LANLF260); US Department of Energy Laboratory Directed Research and Development program (LDRD-ER 20110665ER). Funding for open access charge: US Department of Energy.

Conflict of interest statement. None declared.

REFERENCES

1. Gervais,A.L., Marques,M. and Gaudreau,L. (2010) PCRTiler: automated design of tiled and specific PCR primer pairs. *Nucleic Acids Res.*, **38**, W308–W312.
2. Kaderali,L. and Schliep,A. (2002) Selecting signature oligonucleotides to identify organisms using DNA arrays. *Bioinformatics*, **18**, 1340–1349.
3. Li,K., Brownley,A., Stockwell,T.B., Beeson,K., McIntosh,T.C., Busam,D., Ferriera,S., Murphy,S. and Levy,S. (2008) Novel computational methods for increasing PCR primer design effectiveness in directed sequencing. *BMC Bioinformatics*, **9**, 191.
4. Rozen,S. and Skaletsky,H. (2000) *Primer3 on the WWW for General Users and for Biologist Programmers*. Humana Press, Totowa, NJ.
5. Balla,S. and Rajasekaran,S. (2007) An efficient algorithm for minimum degeneracy primer selection. *IEEE Trans. Nanobiosci.*, **6**, 12–17.
6. Boyce,R., Chilana,P. and Rose,T.M. (2009) iCODEHOP: a new interactive program for designing COnsensus-DEgenerate Hybrid Oligonucleotide Primers from multiply aligned protein sequences. *Nucleic Acids Res.*, **37**, W222–W228.
7. Contreras-Moreira,B., Sachman-Ruiz,B., Figueroa-Palacios,I. and Vinuesa,P. (2009) primers4clades: a web server that uses phylogenetic trees to design lineage-specific PCR primers for metagenomic and diversity studies. *Nucleic Acids Res.*, **37**, W95–W100.
8. Gadberry,M.D., Malcomber,S.T., Doust,A.N. and Kellogg,E.A. (2005) Primaclade–a flexible tool to find conserved PCR primers across multiple species. *Bioinformatics*, **21**, 1263–1264.
9. Giegerich,R., Meyer,F. and Schleiermacher,C. (1996) *Proceedings of the Fourth International Conference on Intelligent Systems for Molecular Biology*, Vol. 4. AAAI Press, St. Louis, MI, pp. 68–77.
10. Graham,K.J. and Holland,M. (2005) PrimerSelect: a transcriptome-wide oligonucleotide primer pair design program for kinetic RT-PCR-based transcript profiling. *Methods Enzymol.*, **395**, 544–553.
11. Jabado,O.J., Palacios,G., Kapoor,V., Hui,J., Renwick,N., Zhai,J., Briese,T. and Lipkin,W.I. (2006) Greene SCPrimer: a rapid comprehensive tool for designing degenerate primers from multiple sequence alignments. *Nucleic Acids Res.*, **34**, 6605–6611.
12. Linhart,C. and Shamir,R. (2002) The degenerate primer design problem. *Bioinformatics*, **18**, S172–S181.
13. Najafabadi,H.S., Torabi,N. and Chamankhah,M. (2008) Designing multiple degenerate primers via consecutive pairwise alignments. *BMC Bioinformatics*, **9**, 55.
14. Souvenir,R., Buhler,J., Stromo,G. and Zhang,W. (2003) String Algorithms. In: Benson,G. and Page,R. (eds), *Workshop on Algorithms in Bioinformatics*. Springer, Budapest, Hungary, pp. 512–526.
15. Wei,X., Kuhn,D. and Narasimhan,G. (2003) *IEEE Computer Society Bioinformatics Conference*. Stanford, CA, pp. 75–83.

16. Frech,C., Breuer,K., Ronacher,B., Kern,T., Sohn,C. and Gebauer,G. (2009) hybseek: pathogen primer design tool for diagnostic multi-analyte assays. *Comput. Methods Programs Biomed.*, **94**, 152–160.
17. Nordberg,E.K. (2005) YODA: selecting signature oligonucleotides. *Bioinformatics*, **21**, 1365–1370.
18. Tembe,W., Zavaljevski,N., Bode,E., Chase,C., Geyer,J., Wasieloski,L., Benson,G. and Reifman,J. (2007) Oligonucleotide fingerprint identification for microarray-based pathogen diagnostic assays. *Bioinformatics*, **23**, 5–13.
19. Ashelford,K., Weightman,A. and Fry,J. (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonuleotide probes and primers in conjunction with the RDP-II database. *Nucleic Acids Res.*, **30**, 3481–3489.
20. Bader,K., Grothoff,C. and Meier,H. (2011) Comprehensive and relaxed search for oligonucleotide signatures in hierarchically clustered sequence datasets. *Bioinformatics*, **27**, 1546–1554.
21. Fitch,J., Gardner,S., Kuczmarski,T., Kurtz,S., Myers,R., Ott,L., Slezak,T., Vitalis,E., Zelma,A. and McCready,P. (2002) Rapid development of nucleic acid diagnostics. *Proc IEEE*, **90**, 1708–1720.
22. Fu,Q., Ruegger,P., Bent,E., Chrobak,M. and Borneman,J. (2008) PRISE (PRImer SElector): software for designing sequence-selective PCR primers. *J. Microbiol. Methods*, **72**, 263–267.
23. Jarman,S.N. (2004) Amplicon: software for designing PCR primers on aligned DNA sequences. *Bioinformatics*, **20**, 1644–1645.
24. Phillippy,A.M., Ayanbule,K., Edwards,N.J. and Salzberg,S.L. (2009) Insignia: a DNA signature search web server for diagnostic assay development. *Nucleic Acids Res.*, **37**, W229–W234.
25. Drosten,C., Gottig,S., Schilling,S., Asper,M., Panning,M., Schmitz,H. and Gunther,S. (2002) Rapid detection and quantification of RNA of Ebola and Marburg Viruses, Lassa Virus, Crimean-Congo Hemorrhagic Fever Virus, Rift Valley Fever Virus, Dengue Virus, and Yellow Fever Virus by real-time reverse transcription-PCR. *J. Clin. Microbiol.*, **40**, 2323–2330.
26. Radnedge,L., Agron,P.G., Hill,K.K., Jackson,P.J., Ticknor,L.O., Keim,P. and Andersen,G.L. (2003) Genome differences that distinguish *Bacillus anthracis* from *Bacillus cereus* and *Bacillus thuringiensis*. *Appl. Environ. Microbiol.*, **69**, 2755–2764.
27. Barns,S.M., Grow,C.C., Okinaka,R.T., Keim,P. and Kuske,C.R. (2005) Detection of diverse new *Francisella*-like bacteria in environmental samples. *Appl. Environ. Microbiol.*, **71**, 5494–5500.
28. Larkin,M.J., Osborn,A.M. and Fairley,D. (2005) A molecular toolbox for bacterial ecologists: PCR primers for functional gene analysis. In: Osborn,A.M. and Smith,C.J. (eds), *Molecular Microbial Ecology*. Taylor and Francis, NY, New York, pp. 249–270.
29. Pereyra,L.P., Hiibel,S.R., Riquelme,M.V., Reardon,K.F. and Pruden,A. (2010) Detection and quantification of functional genes of cellulose- degrading, fermentative, and sulfate-reducing bacteria and methanogenic archaea. *Appl. Environ. Microbiol.*, **76**, 2192–2202.
30. Smith,C.J. and Osborn,A.M. (2008) Advantages and limitations of quantitative PCR (Q-PCR)-based approaches in microbial ecology. *FEMS Microbiol. Ecol.*, **67**, 6–20.
31. Fierer,N., Bradford,M.A. and Jackson,R.B. (2007) Toward an ecological classification of soil bacteria. *Ecology*, **88**, 1354–1364.
32. Lim,J., Shin,S.G., Lee,S. and Hwang,S. (2011) Design and use of group-specific primers and probes for real-time quantitative PCR. *Front. Environ. Sci. Eng. China*, **5**, 28–39.
33. Cole,J.R., Wang,Q., Cardenas,E., Fish,J., Chai,B., Farris,R.J., Kulam-Syed-Mohideen,A.S., McGarrell,D.M., Marsh,T., Garrity,G.M. *et al.* (2009) The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.*, **37**, D141–D145.
34. Probert,W.S., Schrader,K.N., Khuong,N.Y., Bystrom,S.L. and Graves,M.H. (2004) Real-time multiplex PCR assay for detection of Brucella spp., B. abortus, and B. melitensis. *J. Clin. Microbiol.*, **42**, 1290–1293.
35. Janssen,P.H. (2006) Identifying the dominant soil bacterial taxa in libraries of 16S rRNA and 16S rRNA genes. *Appl. Environ. Microbiol.*, **72**, 1719–1728.
36. Eichorst,S.A., Breznak,J.A. and Schmidt,T.M. (2007) Isolation and characterization of soil bacteria that define *Terriglobus* gen. nov., in the phylum. *Acidobacteria Appl. Environ. Microbiol. gen. nov.*, **73**, 2708–2717.
37. Jones,R.T., Robeson,M.S., Lauber,C.L., Hamady,M., Knight,R. and Fierer,N. (2009) A comprehensive survey of soil acidobacterial diversity using pyrosequencing and clone library analyses. *ISME J.*, **3**, 442–453.
38. Sait,M., Davis,K.E. and Janssen,P.H. (2006) Effect of pH on isolation and distribution of members of subdivision 1 of the phylum *Acidobacteria* occurring in soil. *Appl. Environ. Microbiol.*, **72**, 1852–1857.
39. Eichorst,S.A., Kuske,C.R. and Schmidt,T.M. (2011) Influence of plant polymers on the distribution and cultivation of bacteria in the phylum *Acidobacteria*. *Appl. Environ. Microbiol.*, **77**, 586–596.
40. Thomson,B.C., Ostle,N., McNamara,N., Bailey,M.J., Whiteley,A.S. and Griffiths,R.I. (2010) Vegetation affects the relative abundances of dominant soil bacterial taxa and soil respiration Rates in an upland grassland soil. *Microb. Ecol.*, **59**, 335–343.
41. Yin,C., Jones,K.L., Peterson,D.E., Garrett,K.A., Hulbert,S.H. and Paulitz,T.C. (2010) Members of soil bacterial communities sensitive to tillage and crop rotation. *Soil Biol. Biochem.*, **42**, 2111–2118.
42. Dunbar,J., Eichorst,S.A., Gallegos-Graves,L.V., Silva,S., Xie,G., Hengartner,N., Evans,R.D., Hungate,B.A., Jackson,R.B., Megonigal,J.P. *et al.* (2012) Common bacterial repsonses in six ecosystems exposed to 10 years of elevated atmospheric carbon dioxide. *Environ. Microbiol.*, **14**, 1145–1158.
43. Lesaulnier,C., Papamichail,D., McCorkle,S., Ollivier,B., Skiena,S., Taghavi,S., Zak,D. and van der Lelie,D. (2008) Elevated atmospheric CO_2 affects soil microbial diversity associated with trembling aspen. *Environ. Microbiol.*, **10**, 926–941.
44. Barns,S.M., Cain,E.C., Sommerville,L. and Kuske,C.R. (2007) *Acidobacteria* phylum sequences in uranium-contaminated subsurface sediments greatly expand the known diversity within the phylum. *Appl. Environ. Microbiol.*, **73**, 3113–3116.
45. Pankratov,T.A., Ivanova,A.O., Dedysh,S.N. and Liesack,W. (2011) Bacterial populations and environmental factors controlling cellulose degradation in an acidic Sphagnum peat. *Environ. Microbiol.*, **13**, 1800–1814.
46. Ward,N.L., Challacombe,J.F., Janssen,P.H., Henrissat,B., Coutinho,P.M., Wu,M., Xie,G., Haft,D.H., Sait,M., Badger,J. *et al.* (2009) Three genomes from the phylum *Acidobacteria* provide insight into the lifestyles of these microorganisms in soils. *Appl. Environ. Microbiol.*, **75**, 2046–2056.
47. Kishimoto,N., Kosako,Y. and Tano,T. (1991) *Acidobacterium capsulatum* gen. nov., sp. nov., an acidophilic chemoorganotrophic bacterium belonging to the phylum Acidobacteria. *Curr. Microbiol.*, **317**, 138–142.
48. Koch,I.H., Gich,F., Dunfield,P.F. and Overmann,J. (2008) Edaphobacter modestus gen. nov., sp. nov., and Edaphobacter aggregans sp. nov., acidobacteria isolated from alpine and forest soils. *Int. J. Syst. Evol. Microbiol.*, **58**, 1114–1122.
49. Männistö,M.K., Rawat,S., Starovoytov,V. and Häggblom,M.M. (2011) *Terriglobus saanensis* sp. nov., an acidobacterium isolated from tundra soil. *Int. J. Syst. Evol. Microbiol.*, **61**, 1823–1828.
50. Pankratov,T.A. and Dedysh,S.N. (2010) *Granulicella paludicola* gen. nov., sp. nov., *Granulicella pectinivorans* sp. nov., *Granulicella aggregans* sp. nov. and *Granulicella rosea* sp. nov., acidophilic, polymer-degrading acidobacteria from Sphagnum peat bogs. *Int. J. Syst. Evol. Microbiol.*, **60**, 2951–2959.
51. Pankratov,T.A., Kirsanova,L.A., Kaparullina,E.N., Kevbrin,V.V. and Dedysh,S.N. (2011) *Telmatobacter bradus* gen. nov., sp. nov., a cellulolytic facultative anaerobe from subdivision 1 of the Acidobacteria, and emended description of *Acidobacterium capsulatum* Kishimoto et al. 1991. *Int. J. Syst. Evol. Microbiol.*, **62**, 430–437.
52. Durbin,R., Eddy,S., Krogh,A. and Mitchison,G. (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

53. Alpern,B., Carter,L. and Gatlin,K. (1995) *ACM/IEEE Conference on Supercomputing*. ACM (Association for Computing Machinery), NY, New York.
54. SantaLucia,J. Jr and Hicks,D. (2004) The thermodynamics of DNA structural motifs. *Annu. Rev. Biophys. Biomol. Struct*., **33**, 415–440.
55. Gans,J.D. and Wolinsky,M. (2008) Improved assay-dependent searching of nucleic acid sequence databases. *Nucleic Acids Res*., **36**, e74.
56. Karlin,S. and Mrazek,J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Natl Acad. Sci. USA*, **94**, 10227–10232.
57. Gardner,S.N., Kuczmarski,T.A., Vitalis,E.A. and Slezak,T.R. (2003) Limitations of TaqMan PCR for detecting divergent viral pathogens illustrated by hepatitis A, B, C, and E viruses and human immunodeficiency virus. *J. Clin. Microbiol*., **41**, 2417–2427.
58. Pruesse,E., Quast,C., Knittel,K., Fuchs,B., Ludwig,W., Peplies,J. and Glockner,F. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*., **35**, 7188–7196.
59. Ludwig,W., Strunk,O., Westram,R., Richter,L., Meier,H., Yadhukumar., Buchner,A., Lai,T., Steppi,S., Jobb,G. *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res*., **32**, 1363–1371.
60. Wang,Q., Garrity,G.M., Tiedje,J.M. and Cole,J.R. (2007) Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl. Environ. Microbiol*., **73**, 5261–5267.
61. Gans,J., Wolinsky,M. and Dunbar,J. (2005) Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science*, **309**, 1387.
62. Gans,J., Wolinsky,M. and Dunbar,J. (2006) Response to comment by Bunge et al. on ‘Computational improvements reveal great bacterial diversity and high metal toxicity in soil’. *Science*, **313**, 917.