

ThermoPhyl: a software tool for selecting phylogenetically optimized conventional and quantitative-PCR taxon-targeted assays for use with complex samples

Brian B. Oakley¹, Scot E. Dowd² & Kevin J. Purdy¹

¹University of Warwick, School of Life Sciences, Coventry, UK; and ²Research and Testing Laboratories, Lubbock, TX, USA

Correspondence: Brian B. Oakley, USDA ARS, Richard Russell Research Center, 950 College Station Road, Athens, GA 30605, USA. Tel.: +1 706 546 3166; fax: +1 706 546 3771; e-mail: brian.oakley@ars.usda.gov

Received 6 July 2010; revised 23 February 2011; accepted 24 February 2011.
Final version published online 21 March 2011.

DOI:10.1111/j.1574-6941.2011.01079.x

Editor: Christoph Tebbe

Keywords

PCR primer design; specificity; sensitivity; phylogenetic assays; qPCR; PCR.

Introduction

A basic task for many environmental and clinical researchers is to target a specific genotype in a complex sample to understand, for example, the abundance of a particular taxon or the expression of a particular gene. For conventional PCR, users have traditionally designed PCR assays manually, starting with a visual comparison of the alignments of multiple sequences. This approach is at best laborious and, even when a phylogenetically optimal assay (i.e. maximally sensitive and specific) can be identified, empirical tests are time-consuming and can still produce poor PCR results. Software is available that can facilitate primer and probe design and analysis, but none of these completely meet the needs of designing PCR assays to detect target clades specifically. In particular, there is no software presently available to the community that can assess the probe and primer sets required for quantitative-PCR (qPCR) assays using an internal probe, the most accurate

Abstract

The ability to specifically and sensitively target genotypes of interest is critical for the success of many PCR-based analyses of environmental or clinical samples that contain multiple templates. Next-generation sequence data clearly show that such samples can harbour hundreds to thousands of operational taxonomic units, a richness that precludes the manual evaluation of candidate assay specificity and sensitivity using multiple sequence alignments. To solve this problem, we have developed and validated a free software tool that automates the identification of PCR assays targeting specific genotypes in complex samples. ThermoPhyl uses user-defined target and nontarget sequence databases to assess the phylogenetic sensitivity and specificity of thermodynamically optimized candidate assays derived from primer design software packages. ThermoPhyl derives its name from its central premise of testing *Thermodynamically* optimal assays for *Phylogenetic* specificity and sensitivity and can be used for two primer (traditional PCR) or two primers with an internal probe (e.g. TaqMan[®] qPCR) application and potentially for oligonucleotide probes. Here, we describe the use of ThermoPhyl for traditional PCR and qPCR assays. PCR assays selected using ThermoPhyl were validated using 454 pyrosequencing of a traditional specific PCR assay and with a set of four genotype-specific qPCR assays applied to estuarine sediment samples.

and specific qPCR technique. Given the wealth of sequence data rapidly accumulating from high-throughput sequencing and the diversity such methods are detecting in environmental and clinical samples (Sogin *et al.*, 2006; Acosta-Martinez *et al.*, 2008; Andersson *et al.*, 2008; Biddle *et al.*, 2008; Dowd *et al.*, 2008; Hamady *et al.*, 2008), it is clear that high-throughput PCR primer and probe design approaches are required to specifically and sensitively target genotypes in a complex sample.

The software we describe here, ThermoPhyl, exploits the output of proprietary software that can produce large numbers of thermodynamically optimized candidate assays for PCR and qPCR [e.g. Primer Express (Applied Biosystems (ABI), Warrington, UK), BatchPrimer3 (You *et al.*, 2008)] and subsequently assessing each individual assay for sensitivity (proportion of target group perfectly matched) and specificity (number of nontarget organisms perfectly matched) to user-defined target and nontarget sequences in a local database. ThermoPhyl is built around a simple

pattern-matching script and is designed for applications in which a user wishes to target a taxonomic group of interest in a complex sample. The name ThermoPhyl derives from its central goal, which is to test thermodynamically optimal PCR assays for phylogenetic sensitivity and specificity to arrive at an assay that is both *thermodynamically* and *phylogenetically* optimal. ThermoPhyl is designed to analyse a very large number of candidate assays simultaneously and to facilitate primer and probe set choice by providing an output that clearly summarizes the specificity and sensitivity of each candidate assay.

ThermoPhyl complements several software tools currently available such as NCBI Primer-Blast, probeCheck (Loy *et al.*, 2008) and ARB (Ludwig *et al.*, 2004), which meet a range of user needs related to determining the phylogenetic specificity of targeted primers and probes. ThermoPhyl performs the high-throughput assessments of assays with one (e.g. FISH), two (conventional PCR) or three (e.g. TaqMan[®] qPCR) oligonucleotides, which no other available software can perform. Additionally, ThermoPhyl is installed locally, which allows users to utilize personal databases and fully control processing speed and throughput. ThermoPhyl can efficiently harness the power of large datasets by rapidly performing a large number of comparisons that are summarized in output sorted by specificity and sensitivity.

In this paper, we describe the rationale behind and the use of ThermoPhyl, compare it with other commonly used primer PCR primer and probe design and assessment programs and validate its use in the design of both conventional PCR and qPCR assays. For conventional PCR, we used ThermoPhyl to select a PCR primer set to amplify the *mcrA* gene of the methanogenic archaeal genus *Methanosaeta* and experimentally tested its specificity and sensitivity using pyrosequencing. For qPCR, we used ThermoPhyl to select four *Desulfobulbus* genotype-specific qPCR assays and compared the data produced by this analysis with existing data. The program, a user's manual, and a training dataset are made freely available to the research community at <http://go.warwick.ac.uk/thermophyl>.

Materials and methods

Program validation: traditional PCR primer set

Sequence database and target group definition

Using *mcrA* sequences from the functional gene pipeline/repository (<http://fungene.cme.msu.edu/>), a database was constructed in ARB (Ludwig *et al.*, 2004). Based on maximum-likelihood phylogenetic reconstructions, a monophyletic clade containing 19 *Methanosaeta* sequences, including sequences from the characterized three isolates *Methanosaeta concilii*, *Methanosaeta harundinacea*, and *Methanosaeta*

thermophila, was identified as a target group. BatchPrimer3 (You *et al.*, 2008) was used with default parameters to design 50 primer pairs for each of these target sequences with amplicon lengths set between 400 and 500 bp.

After comparison with the target sequences, two primer pairs ranked highly by ThermoPhyl, F1-1044 (5'-CTACATGTCCG GYGGTGTC-3') and R1-1507 (5'-TAGTTRGCGCCYCTCAKC TC-3'), and F2-1060 (5'-GTCGGWTTACMCAGTACGC-3') and R2-1470 (5'-TGCCCTCGTCKGACTGGTA-3'), were chosen for empirical testing after the inclusion of several degeneracies based on a manual comparison with the sequence database.

Primer assessments

Primer specificity, sensitivity, and amplification efficiency were evaluated empirically using genomic DNA from *M. concilii* (DSM6752), *M. harundinacea* (DSM 17206), *M. thermophila* (DSM4774), *Methanosarcina mazei* (DSM2053), *Methanosarcina acetivorans* (DSM2834), and environmental DNA from sediment in the Colne Estuary, UK. DNA was extracted from actively growing cultures using the DNeasy Blood and Tissue kit (Qiagen, Crawley, UK) and environmental DNA was extracted from sediment samples representing marine and freshwater conditions [sites 1 and 10 in Oakley *et al.* (2010)] as described previously (Purdy *et al.*, 1996; Purdy, 2005). The first round of PCR was performed in 25 µL volumes containing 1 × EpiCentre FailSafe Master Mix G (EpiCentre, Madison WI), 600 nM each primer F1 and R1, and 1.25 U EpiCentre FailSafe Enzyme Mix. Touchdown PCR was performed with an initial denaturation of 96 °C for 2 min, followed by 10 cycles of 94 °C for 30 s, 52 °C (– 1 °C per cycle) for 30 s, and 72 °C for 40 s, followed by 22 cycles with 42 °C annealing. Because of the low amounts of *Methanosaeta* DNA present in the environmental samples, 1 µL of these PCR products was used in a second round of PCR containing 1 × Promega PCR buffer, 2.5 mM MgCl₂, 10 µg bovine serum albumin, 600 nM each primer F2 and R2, 200 nM dNTPs, and 1 U Promega Taq polymerase. Thermal cycling consisted of 96 °C for 2 min, 30 cycles of 94 °C for 30 s, 50 °C for 30 s, 72 °C for 40 s, and a final extension at 72 °C for 10 min. Triplicate PCRs from each of three biological replicates [sediment samples taken within 50 cm at each site (Hawkins & Purdy, 2007)] were cleaned using the QiaQuick Gel Extraction kit (Qiagen) and pooled after normalization based on the quantification of PCR products using QuantIt PicoGreen (Invitrogen, Paisley, UK) as per the manufacturer's instructions and fluorescence was measured using a Perkin-Elmer Wallac Victor² 1420 plate reader.

Pyrosequencing methods

Pyrosequencing was performed at Research and Testing Laboratory (Lubbock, TX: <http://www.researchandtesting>).

com) using tagged amplicon methods similar to those described previously (Dowd *et al.*, 2008) modified for titanium chemistry (Roche, Indianapolis). In brief, concatamer primers were synthesized using the construct '5' 454TitaniumLinkerA-tag-primer 3'', where the 454 linker A was biotin labeled, and based on Roche amplicon sequencing Titanium Linker A, the tag was a random 10mer (GC content 40–60%) and utilized to bin out sequences resulting from a specific sample and the primers used for this study. The reverse concatenate was in the format '5' 454Titanium-LinkerB-primer', where the primer was the appropriate reverse primer for the reaction. Twenty cycles of PCR were utilized (94 °C for 30 s, 50 °C for 30 s, 72 °C for 40 s), with a final extension at 72 °C for 10 min to incorporate the linkers and tag. Pyrosequencing based on titanium bulk sequencing methods was utilized based on the manufacturers' protocols by introducing the amplicon into the steps in the protocol following library creation. A 200 flow Titanium sequencing run was performed according to the Roche protocols with amplicon signal processing. Following the sequencing and image processing, the sequences were binned out based on tag sequence into individual MULTI-FASTA files and used for data analysis.

Pyrosequence data analysis

Raw sequence data were edited using a series of custom PERL and BIOPERL scripts, which performed the following initial steps: trimming of pyrosequencing tag sequences, removal of sequences with one or more ambiguous base calls, and removal of sequences shorter than 410 bp. Sequences were screened for the presence of both forward and reverse primer sequences and then translated in all three forward frames and screened for the presence of a conserved motif (VGF) within the forward primer region; translated sequences without stop codons and no more than one unknown amino acid passed the screen and leading and trailing nucleotides were trimmed to complete codons in the appropriate frame. A total of 1745 sequences from site 10 (out of 7661) and 4517 sequences from site 1 (out of 10874) passed all screens. The minimum, median, and maximum sequence lengths were 411, 415, and 419 bp respectively.

To determine the identity of sequences, a BLASTP analysis was performed by querying each translated sequence against a custom database of 44 *mcrA* reference sequences from pure cultures including all known *Methanosaeta* strains. The only changes to default BLASTP parameters were the use of soft-masking (-F 'm S') to enable filtering for low-complexity subsequences during the word seeding phase, but not the extension phase of the BLASTP algorithm.

Sequences were aligned with MUSCLE (Edgar, 2004) invoked from a BIOPERL shell, which first appended an anchor-

ing oligo sequence (5'-ACCACACAAAAACCCACA-3') to both the 5' and the 3' ends of the alignment and then randomly split all sequences into subsets of 1000 sequences and aligned these first to each other, and then to a single reference sequence from *M. concilii* (AF313802). Each alignment of 1000 sequences was then appended to the previous, and finally, the entire alignment was aligned as a profile to the reference sequence. Alterations to default MUSCLE parameters were gap-open and gap-extend penalties of -500. This alignment strategy was arrived at empirically by optimization of a training data set of pyrosequencing data derived from a clonal *M. concilii* sample and resulted in significantly increased accuracy and reduced computational time (*c.* 1 vs. 66 h) relative to the default command line invocation of MUSCLE (data not shown).

Distance matrices were constructed with PHYLIP (Felsenstein, 1989) using the Kimura two-parameter model and sequences were grouped into operational taxonomic units (OTUs) using the furthest-neighbour method of DOTUR (Schloss & Handelsman, 2005). PERL was used to generate a custom import filter to incorporate sequences into ARB (Ludwig *et al.*, 2004) including a data table of sequence membership for each OTU. Phylogenetic trees were built with maximum-likelihood algorithms in ARB using representative sequences from each site for the 27 OTUs identified by a 20% sequence dissimilarity cutoff conservatively based on a pairwise nucleotide difference between *M. concilii* and *M. harundinacea* of 25% for the amplicon region.

Program validation: qPCR methods

Probe and primer sets were designed using Primer Express (ABI), which outputs Taqman probe and primer sets that should function effectively using ABI's standard qPCR conditions. Multiple potential assays were derived from all of the target sequences that were available and all candidate assays were tested using ThermoPhyl.

qPCR assays were performed using Applied Biosystems TaqMan[®] gene expression master mix with MGB probes as described previously (Oakley *et al.*, 2010). The primer and probe sequences for assays M, Mh, FW1, and FW2 are listed in Table 1. In brief, PCR conditions were as per the manufacturer's recommendations on an ABI 7000 or ABI 7500. Each assay was optimized with titrations of the primer and probe and the sensitivity and specificity were validated with plasmids from representative target and closely related nontarget clones. Optimized assays contained 300 nM (600 nM for FW2) each of the forward primer and reverse primers, 200 nM MGB probe, and 1 × ABI GeneExpression Master Mix in 25 µL reactions. Assays were considered valid when all no-template controls were negative, calibration-curve R^2 values were > 98%, and the amplification efficiency was between 90% and 115%. All replicates from each

sampling site were run on a single plate for consistency. Assays have been utilized and analysed previously in Oakley *et al.* (2010).

Results and discussion

Program structure and function

Conventional methods of PCR primer design for the analysis of complex communities rely on an initial visual comparison of sequence alignments and then a painstaking comparison of potential primers. While programs such as ARB and Primer-Blast can perform these functions and do produce good PCR primers and probes, performing comparisons between differ-

ent primer sets is laborious and does not guarantee success. Furthermore, designing a phylogenetically coherent, thermodynamically optimized Taqman-like qPCR assay is simply not possible in any of the presently available programs. Primer design programs can produce large numbers of thermodynamically optimized primer (Primer3, <http://www.sourceforge.net>) or primer/probe (Primer Express, ABI) sets, but only do so for individual sequences and so cannot produce assays that specifically target only the members of a clearly defined clade without the need for very laborious comparisons of individual candidate assays. ThermoPhyl fills this gap by comparing the output of high-quality primer design software with sequence data derived from a coherent phylogeny. A flow diagram showing the steps required to use ThermoPhyl is given in Fig. 1.

Table 1. *Desulfobulbus* clade-specific qPCR assay primer and probe sequences selected using ThermoPhyl

Assay	Forward primer	Reverse primer	Probe	Position
M	TGATTGACCACACCCGTATTACC	GCCGTTACCTCAGCCTTAG	ATCTCTGCTTGTCGGCTC	673–783
Mh	CGCTGTTCATGCTTCCGATA	GATCGATCATCGGCGGTTT	CCTCGGTGTGCATCG	620–681
FW1	TCGCCATTCTCGGTATCCAT	CCGGTGATCCGGTCTGT	CAAACCGCCGATGAT	640–704
FW2	CCGGTTAAGGCGGTTATGG	CGCCGGCAAGGTCATG	TGATCTGTTGAGTATTTTGGTT	595–653

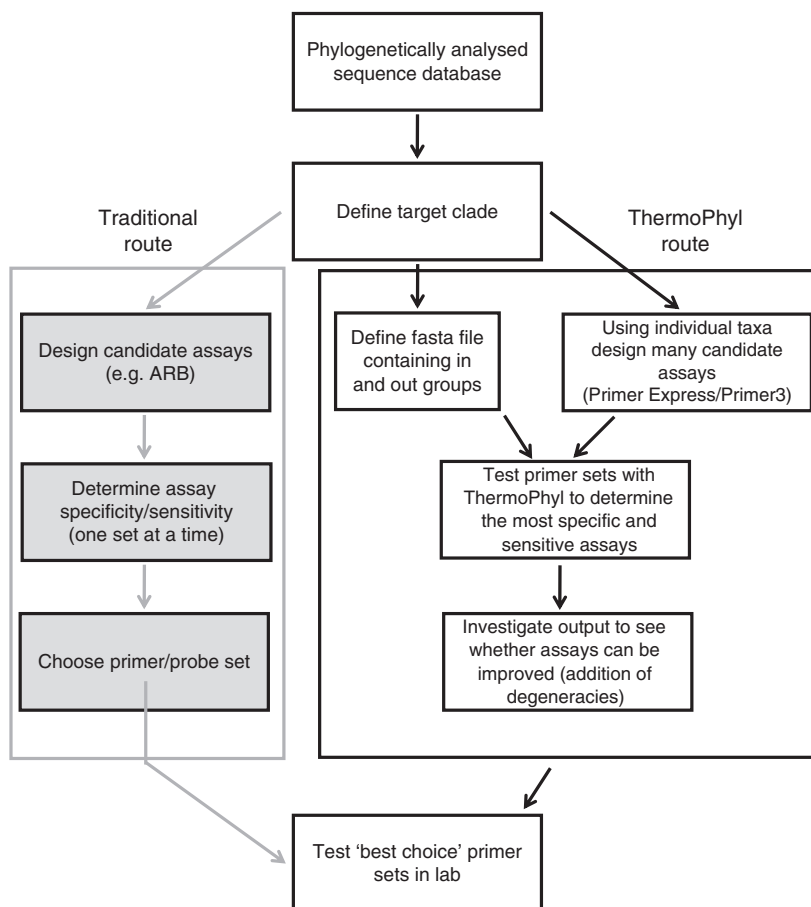


Fig. 1. Flow diagram showing how ThermoPhyl is used in comparison with existing standard practice.

(a) 'sorted_search_results'

	A	B	C	D	E	F	G	H
1	Primer_pair	Target_matches	Non_target_matches	F	F_pos	R_motif	R_pos	amp_length
2	776	1	0	CTACATGTCGGCGGTGTC	6	GAGCTGAGAGGCGCCAATA	448	442
3	786	6	2	CTACATGTCGGCGGTGTC	23	GAGATGAGGGGCGCCAATA	465	442
4	781	1	0	CTACATGTCGGCGGTGTC	6	GAGCTGAGGGGCGTAATA	448	442
5	783	1	0	CTACATGTCGGCGGTGTC	6	GAGCTGAGAGGCGTAATA	448	442
6	211	7	2	AAGCTACATGTCGGTGGTC	20	TCCTACCAGTCAGACGAGGG	435	415
7	222	7	2	AAGCTACATGTCGGTGGTC	20	GTCCAACCAACGTGTTCTCC	418	398
8	230	7	2	AAGCTACATGTCGGTGGTC	20	TCCAACCAACGTGTTCTCT	419	399
9	234	7	2	AGCTACATGTCGGTGGTC	21	TCCTACCAGTCAGACGAGGG	435	414

(b) 'raw_search_results'

	A	B	C	D	E	F	G	H	I
1	Target(Y/N)	Primer_pair	Cumulative_Matches	F_pos	R_pos	Amplicon_length	F	R_motif	Match
2	Y	786	1	6	448	442	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	RS-MCR14 AF313820
3	Y	786	2	6	448	442	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	RS-MCR24 AF313828
4	Y	786	3	6	448	442	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	RS-MCR30 AF313834
5	Y	786	4	6	448	442	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	RS-MCR32 AF313836
6	Y	786	5	6	448	442	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	MRR-mcr32
7	Y	786	6	6	447	441	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	PMmcr28
8	N	786	1	23	465	442	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	ARB_2275F6E9
9	N	786	2	23	465	442	CTACATGTCGGTGGTGC	GAGATGAGGGGCGCCAATA	ARB_F7253DB1
10	Y	776	1	6	448	442	CTACATGTCGGCGGTGTC	GAGCTGAGAGGCGCCAATA	Methanosaela concilii AF313802
11	Y	781	1	6	448	442	CTACATGTCGGTGGTGC	GAGCTGAGGGGCGTAATA	RS-MCR42 AF313845
12	Y	783	1	6	448	442	CTACATGTCGGTGGTGC	GAGCTGAGAGGCGTAATA	RS-MCR17 AF313823
13	Y	1	1	42	451	409	TACTGCTGCCTACCAACG	CTGAGAGGCGCAACTACC	Methanosaela concilii AF313802
14	Y	2	1	49	451	402	GCCTACACCAACGATGTCCT	CTGAGAGGCGCAACTACC	Methanosaela concilii AF313802
15	Y	3	1	42	452	410	TACTGCTGCCTACCAACG	TGAGAGGCGCAACTACC	Methanosaela concilii AF313802
16	Y	4	1	49	452	403	GCCTACACCAACGATGTCCT	TGAGAGGCGCAACTACC	Methanosaela concilii AF313802

Fig. 2. Output files from ThermoPhyl (text files opened in Excel). (a) Results sorted according to the highest number of 'target group' hits for each individual. The F1/R1 primers have been moved to the top of table and highlighted in grey. (b) Raw data, showing how each assay accumulates matches with target and nontarget sequences. The F1/R1 primers have been moved to the top of table and highlighted in grey.

ThermoPhyl is a simple pattern-matching PERL script that compares primers and probes with two user-defined datasets: the 'target group' and the 'nontarget group'. For any number of possible primer/probe sets, ThermoPhyl determines assay sensitivity, that is, how many of the 'target group' are a perfect match for each primer/probe set, and specificity, how many of the 'nontarget group' are a perfect match for the primer/probe sets, for each individual assay. ThermoPhyl then outputs a 'sorted' assay file detailing assays in order of the highest sensitivity and specificity first (see Fig. 2a). It also outputs a 'raw-data file' showing which members of the 'target group' and 'nontarget group' matched with each assay (Fig. 2b). From this output, it is possible to determine whether the addition of degeneracies

in the primers might improve assay sensitivity, although such changes may adversely affect the assay as a whole.

Program requirements

ThermoPhyl, a freeware PERL script program (download PERL from <http://www.activestate.com/activeperl>), is a simple matching program that runs on both windows and Unix machines. In tests on moderately fast WinXP machines (e.g. 2 GHz Pentium CPU with 3 Gb of RAM), testing 5000 candidate qPCR assays against a database with 5000 taxa, and makes the required 25 million comparisons in about 2.5 h. However, most users will have many fewer comparisons than this and for most applications ThermoPhyl generally produces output in seconds to minutes.

To run ThermoPhyl, three input files are required: first, a FASTA file that contains all of the desired target and nontarget sequences, all with unique names, thought to be present in the samples of interest. This file should contain as many representative sequences as possible (typically 100–50 000 depending on the application) to maximize confidence in distinguishing between target and nontarget groups. Because many public databases [e.g. GreenGenes (DeSantis *et al.*, 2006) and Silva (Pruesse *et al.*, 2007) for 16S rRNA genes] now contain many very similar sequences, users may want to reduce these databases to representative sequences.

The second file required is a text file containing only the names of the target sequences. The names *must* correspond exactly to those in the FASTA file above and should be unique, such as a GenBank accession number.

Finally, a list of candidate assays based on the target sequences must be provided. These can be produced by a number of primer design programs. For traditional PCR BatchPrimer3 (You *et al.*, 2008), <http://probes.pw.usda.gov/cgi-bin/batchprimer3/batchprimer3.cgi>, can provide specific candidate assays for a number of different target sequences from a single FASTA file and is a good high-throughput solution to creating candidate assays to test.

For qPCR, software such as ABI's Primer Express can quickly generate a list of candidate assays to test and so allow the use of standardized protocols for qPCR. Using these approaches, we have typically generated 50 candidate assays per target sequence, which are then compiled into a single tab-delimited text file, the candidate-assay file.

Comparison of ThermoPhyl with other primer analysis programs

Several programs have been developed to allow users to design or assess probes and primers to determine whether they are specific and sensitive. However, ThermoPhyl was designed when it became apparent that there was no software available that was designed to assess whether a Taq-man-like qPCR probe/primer set was specific and sensitive. To highlight the differences and similarities between these programs, ThermoPhyl is compared with several other commonly used programs in Table 2. The primary advantage that ThermoPhyl has over all the other programs is that it is the only program that is capable of assessing qPCR probe/primer sets and that is designed to assess many candidate assays at one time. The other programs all have

Table 2. Comparison of ThermoPhyl with four other commonly used oligonucleotide analysis programs

Characteristics	Programs				
	ThermoPhyl	ARB*	Primer-Blast	PRIMROSE†	probeCheck‡
Types of assay					
qPCR probe/primer sets	Yes	No	No	No	NA
Standard PCR	Yes	Yes	Yes	No	NA
Oligonucleotide	Yes	Yes	No	Yes	Yes
Design primers/oligos	No	Yes	Yes	Yes	No
No. of sequences assays designed against	User defined (external program)	User defined (oligos) 1 (primers)	1	User defined	User defined (external program)
Assessment					
Phylogenetic basis	Yes	Yes (oligos) No (primers)	Yes	Yes	Yes
Databases	User defined	16/23S	RNA ref Seq	User defined	16S
No. of target sequences assessed	User defined	User defined (oligos) 1 (primers)	User defined	User defined	NA
Target/nontarget definition	Yes	Yes	Yes	Yes	Yes
Defines nontarget hits	Yes	Yes (oligo) No (primers)	No	Yes	Yes
Degenerate bases allowed?	No	Yes	No	Yes	Yes
No. of assays assessed	No limit§	User defined (oligos) 0 (primers)	≤ 2000	NA	10
Output	Text	Text	Graphic html	Specific	Text
Local/Remote	Local	Local	Remote	Local	Remote
Reference	This study	Ludwig <i>et al.</i> , 1997	NCBI website	Ashelford <i>et al.</i> , 2002	Loy <i>et al.</i> , 2008

*ARB designs and compares oligonucleotide probes with reference to the entire database, but does not do so for PCR primer sets, which are designed with reference to a single sequence only and the subsequent primer sets cannot be directly compared against the database.

†PRIMROSE designs and assesses single oligonucleotides, not PCR primer sets.

‡probeCheck is designed to check single oligonucleotide specificity.

§Limit is based on the processing time, which depends of the speed of the local computer, the number of candidate assays, and the sizes of the target and nontarget databases.

additional limitations, which makes ThermoPhyl a valuable tool in the design of probes and primers. One of ARB's strengths is its oligonucleotide probe design capability, which it does with reference to its sequence database and allows the user to visually assess the newly designed probes. However, ARB's PCR primer design is much more limited. It utilizes just a single target sequence to produce candidate assays and offers no simple possibility of assessing these primer sets against the database. Primer-Blast also designs PCR primer sets (using Primer3) and, like ARB this is presently limited to using a single target sequence. Primer sets are then compared with a choice of databases, including the entire nucleotide database, but with a strong recommendation to use nonredundant databases such as Refseq RNA. The user can set the target clade; however, it is then entirely dependent on the GenBank taxonomy to define the target clade and there is no option to use a user-defined database. PRIMROSE (Ashelford *et al.*, 2002) does allow the use of user-defined databases, although it is designed to work with the RDP database. Its major drawbacks are that it only designs oligonucleotides not PCR primer sets and each oligonucleotide needs to be individually assessed against the database to determine exactly what it matches, which is time-consuming. Therefore, after a search for potential oligonucleotides, potential primer sets will need to be further assessed for thermodynamic suitability. Finally, probeCheck is designed to assess previously designed oligonucleotides, but does so individually and not as a primer set. Therefore, ThermoPhyl along with programs such as Primer Express and BatchPrimer3, allows the very rapid design and assessment of very many thermodynamically optimized PCR and qPCR assays against any dataset that the user chooses.

Validation of ThermoPhyl-selected PCR and qPCR assays

Validation of a ThermoPhyl-selected conventional PCR assay

As the first empirical test of ThermoPhyl, conventional PCR primers specifically targeting the α -subunit of the methyl-coenzyme M reductase gene (*mcrA*) of *Methanosaeta* were selected following the scheme set out in Fig. 1. Over 700 candidate primers sets were designed using BatchPrimer3 from a range of available *Methanosaeta mcrA* sequences from both isolates and environmental clone sequences. As this particular taxon generally represents a small proportion of DNA in the estuarine sediments analysed here, a nested PCR approach was adopted and so two primer sets were selected. After routine manual optimization of thermal cycling conditions, the primers selected by ThermoPhyl produced a strong single band from both genomic

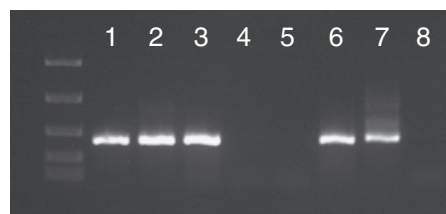


Fig. 3. PCR amplification of the *Methanosaeta mcrA* gene from Colne estuary sediment using the primers F2 and R2 as described in the text. Lanes: 1, *Methanosaeta concilii*; 2, *Methanosaeta harundinacea*; 3, *Methanosaeta thermophila*; 4, *Methanosarcina mazei*; 5, *Methanosarcina acetivorans*; 6, site 1 sediment DNA; 7, site 10 sediment DNA; and 8, no-template control.

DNA prepared from the three available *Methanosaeta* isolates (*M. concilii*, *M. harundinacea* and *M. thermophila*) and environmental DNA preparations, but did not amplify DNA from the closely related *M. mazei* or *M. acetivorans* (Fig. 3).

Using these nested primer sets, *Methanosaeta*-specific PCR products were amplified from DNA extracted directly from the two contrasting environmental sediment samples (marine-dominated site 1 and freshwater-dominated site 10) from the River Colne, Essex, UK (Hawkins & Purdy, 2007; Oakley *et al.*, 2010). These amplicons were analysed using only a small proportion (~1%) of a 454 pyrosequence read. After screening to remove poor-quality sequences as described above, 6262 high-quality sequences remained (4517 sequences from site 1 and 1745 sequences from site 10). These sequences were checked using a local BLASTP analysis to a database of 44 *mcrA* sequences, and across the two sampling sites, 99.9% (6257/6262) of sequences were most closely related to *Methanosaeta* (Table 3). Rarefaction analysis of these sequences showed that, at a sequence dissimilarity of 20% (the difference between the *mcrA* genes of the two mesophilic *Methanosaeta* isolates, *M. concilii* and *M. harundinacea*, is 25%; therefore, 20% is a reasonable species-level definition), the *Methanosaeta* communities at both sites have been completely sampled (Fig. 4). Pyrosequencing of these environmental amplicons also revealed an extensive novel diversity within the *Methanosaeta* clade (Fig. 5). Twenty-seven OTUs were defined at a 20% cut-off, all falling within the *Methanosaeta* clade, and yet many clearly represent novel lineages affiliated with *Methanosaeta*. Therefore, these nested PCR primer sets are both specific and sensitive and show the value of using ThermoPhyl in primer selection.

Validation of ThermoPhyl-selected qPCR assays

The second validation test was to use ThermoPhyl to select Taqman qPCR primer and probe sets, designed using Primer Express (ABI), that targeted four *Desulfobulbus* genotypes

detected in the Colne estuary, UK (data presented previously in Oakley *et al.*, 2010). The four *Desulfobulbus* clades were selected for qPCR analysis because they exhibited a differential distribution along the estuary based on an initial denaturing gradient gel electrophoresis analysis, a distribution that was subsequently supported by clone sequence data (Fig. 6). Candidate assays were designed to target each of the four genotypes; the specificity and sensitivity of these assays were determined using ThermoPhyl and the best

Table 3. Summary of the local BLASTP analysis of the *Methanosaeta mcrA* sequence data

Top BLASTP hit	Site 1	Site 10
Targets		
<i>Methanosaeta concilii</i> AF313802	2563	1674
<i>Methanosaeta harundinacea</i> AY970348	1825	13
<i>Methanosaeta concilii</i> VeAc9 AF313803	115	148
<i>Methanosaeta thermophila</i> PT gb ABK14360.1	9	10
Nontargets		
<i>Methanothermobacter marburgensis</i> X07794	2	
<i>Methanothermus fervidus</i> J03375	1	
<i>Methanococcus jannaschii</i> mrtA U67465	1	
<i>Methanococcoides burtonii</i> U22234	1	
Total	4517	1745

Pyrosequence data were analysed by BLASTP run locally on a custom database containing 44 pure culture *mcrA* sequences, including all *Methanosaeta* strains. The only changes to default parameters were the use of soft-masking (-F 'm S') to enable filtering for low-complexity subsequences during the word seeding phase, but not the extension phase of the BLASTP algorithm.

assays were selected. Data from these four assays supported our previous data showing that all four genotypes have a restricted distribution along the estuary (Fig. 6), indicating that the assays were targeting the correct genotypes. However, these data do not prove that the assays are specific. It can be reasoned that a 'good' qPCR assay should produce PCR products that, if sequenced and analysed phylogenetically, should produce a monophyletic clade, with the caveat that as qPCR assays usually produce very short fragments, the resultant trees are unlikely to be very robust. Therefore, we cloned and sequenced ~12 amplicons for each assay and all four assays produce monophyletic groups after sequence analysis (data not shown). Therefore, ThermoPhyl was successful in selecting highly specific and sensitive qPCR primers and probes from a large number of thermodynamically optimized candidate assays.

These two validation tests show that ThermoPhyl is capable of analysing large numbers of potential PCR and qPCR assays for specificity and sensitivity using a user-defined sequence database and thus will allow the user to make a phylogenetically informed choice about which primer sets to use for a specific target group. This is particularly powerful with qPCR assays as no presently available phylogenetic program is capable of assessing the validity of even a single qPCR primer/probe set, let alone many hundreds of candidate assays. Therefore, while ThermoPhyl is in itself a simple pattern-matching program, it fills a gap in the available software by linking a wholly user-defined dataset to powerful PCR and qPCR primer design software.

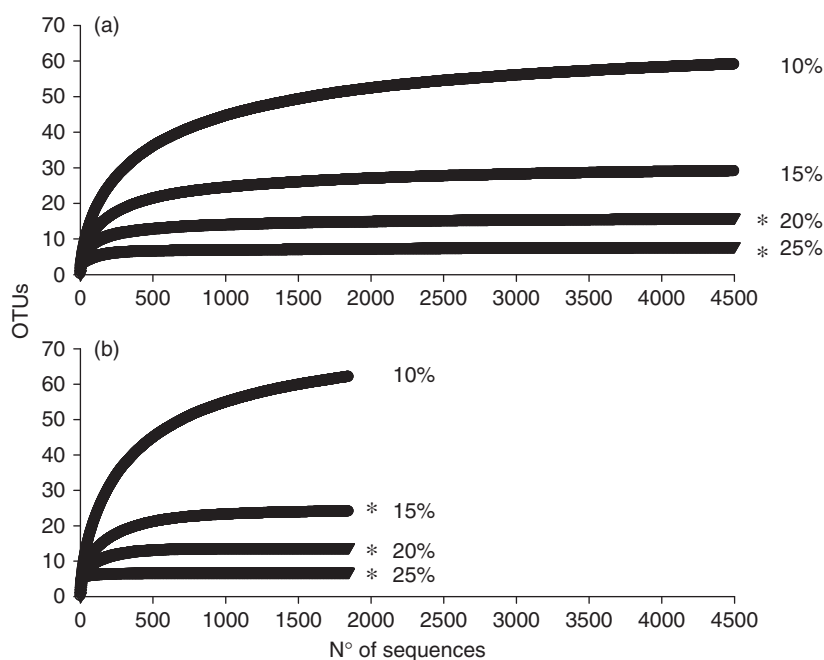


Fig. 4. Rarefaction analysis of *Methanosaeta mcrA* pyrosequence data with OTUs defined at 10%, 15%, 20%, and 25% sequence dissimilarity from (a) site 1 and (b) site 10. Curves marked with * were saturated for OTU definitions. The sequence dissimilarity between *Methanosaeta concilii* and *Methanosaeta harundinacea* is 25% for the amplicon region.

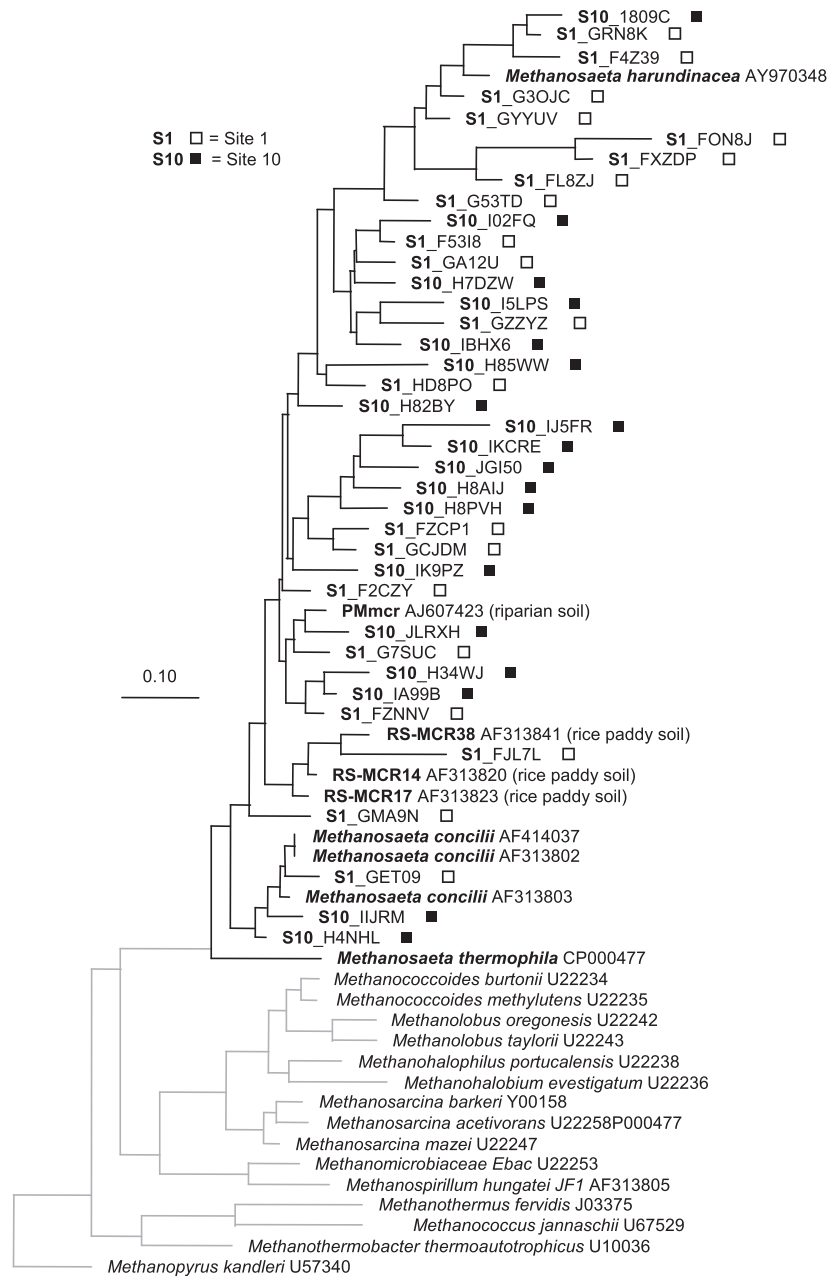


Fig. 5. Phylogenetic representation of *mcrA* sequence diversity recovered by ThermoPhyl-generated primers. The tree is a maximum-likelihood phylogenetic reconstruction based on alignment of nucleotides restricted to the amplicon region. Sequences are labelled with either S1 or S10, indicating whether they are from site 1 or site 10, respectively, and those shown are representatives of the 27 OTUs defined as described in the text.

Potential pitfalls and specific recommendations

To use ThermoPhyl effectively, target groups must form a natural phylogenetic group. Before using ThermoPhyl, sequences should be properly placed in some sort of a phylogenetic tree to evaluate this and to designate target and nontarget sequences in a way that reflects the evolutionary history of the gene in question. If the target sequences do not form a coherent phylogenetic group, it will be difficult to design an accurate assay, although it is

possible that different sequence data (e.g. another gene) for the same taxa could still be used in such a case.

Additionally, the more sequence data available for both target and nontarget groups, the better. The strength of ThermoPhyl in fact its central goal, is to summarize a very large number of comparisons to arrive at a single 'best' assay. However, users should be aware that some genes or clades may prove more challenging than others, especially if the targeted gene is highly variable or does not carry a strong phylogenetic signal. Additional guidance is provided in the

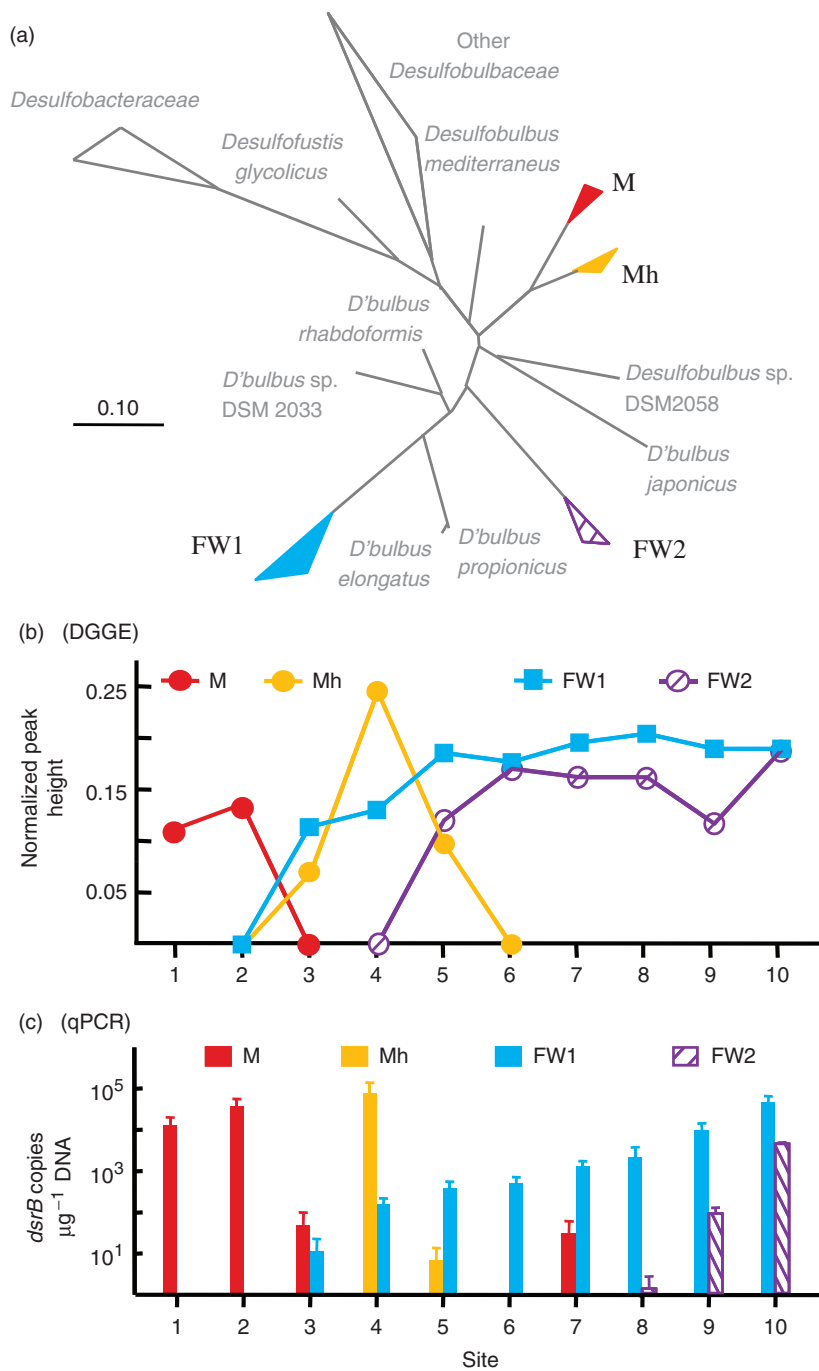


Fig. 6. Identity and distributions of *Desulfobulbus*-affiliated *dsrB* ecotypes (adapted from fig. 3 in Oakley et al., 2010). (a) Phylogenetic positions of the four assayed clades within *Desulfobulbus*. Tree was reconstructed from amino acid informed DNA alignment using the maximum-likelihood algorithm AxML. (b) Distributions of these four genotypes across the estuary as assessed by denaturing gradient gel electrophoresis. Values represent peak heights normalized within each lane to control for loading differences. (c) Distributions of four genotypes as assessed by clade-specific qPCR assays. Values represent means of three biological replicates (error bars = 1 SEM).

user's manual, and common questions are listed in the FAQ, both accessible via the ThermoPhyl website.

While ThermoPhyl can perform the most laborious aspects of selecting primer sets, it is necessary for the user to engage with the ThermoPhyl output to determine how well the 'best' assays suit their purpose. We have found that using ThermoPhyl's output within programs such as ARB can rapidly confirm the potential value of a primer set and highlight where degenerate bases could improve sensitivity

without unduly compromising specificity, although this is not recommended for qPCR probe and primer sets unless absolutely necessary.

Conclusions

ThermoPhyl can utilize large sequence datasets now commonly available to identify phylogenetically specific and sensitive assays for traditional and qPCR. ThermoPhyl is

run locally on a user's computer, avoiding the constraints of internet data transmission, and allowing for customized, personal databases. ThermoPhyl can provide a high-throughput data-driven solution to the problem of targeted assay design in complex samples and is made available free to the research community at <http://go.warwick.ac.uk/thermophyl/>.

Acknowledgements

This work is part of the Marie Curie Excellence Grant for Teams project, MicroComXT (MEXT-CT-2005-024112), funded by the European Commission under FP6.

References

- Acosta-Martinez V, Dowd S, Sun Y & Allen V (2008) Tag-encoded pyrosequencing analysis of bacterial diversity in a single soil type as affected by management and land use. *Soil Biol Biochem* **40**: 2762–2770.
- Andersson AF, Lindberg M, Jakobsson H, Backhed F, Nyren P & Engstrand L (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS ONE* **3**: e2836.
- Ashelford KE, Weightman AJ & Fry JC (2002) PRIMROSE: a computer program for generating and estimating the phylogenetic range of 16S rRNA oligonucleotide probes and primers in conjunction with the RDP-II database. *Nucl Acids Res* **30**: 3481–3489.
- Biddle JE, Fitz-Gibbon S, Schuster SC, Brenchley JE & House CH (2008) Metagenomic signatures of the Peru Margin subseafloor biosphere show a genetically distinct environment. *P Natl Acad Sci USA* **105**: 10583–10588.
- DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P & Andersen GL (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microb* **72**: 5069–5072.
- Dowd SE, Sun Y, Secor PR, Rhoads DD, Wolcott BM, James GA & Wolcott RD (2008) Survey of bacterial diversity in chronic wounds using pyrosequencing, DGGE, and full ribosome shotgun sequencing. *BMC Microbiol* **8**: 43.
- Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792–1797.
- Felsenstein J (1989) PHYLIP – Phylogeny Inference Package (Version 3.2). *Cladistics* **5**: 164–166.
- Hamady M, Walker JJ, Harris JK, Gold NJ & Knight R (2008) Error-correcting barcoded primers for pyrosequencing hundreds of samples in multiplex. *Nat Methods* **5**: 235–237.
- Hawkins RJ & Purdy KJ (2007) Genotypic distribution of an indigenous model microorganism along an estuarine gradient. *FEMS Microbiol Ecol* **62**: 187–194.
- Loy A, Arnold R, Tischler P, Rattei T, Wagner M & Horn M (2008) probeCheck – a central resource for evaluating oligonucleotide probe coverage and specificity. *Environ Microbiol* **10**: 2894–2898.
- Ludwig W, Strunk O, Westram R *et al.* (2004) ARB: a software environment for sequence data. *Nucleic Acids Res* **32**: 1363–1371.
- Oakley BB, Carbonero E, van der Gast CJ, Hawkins RJ & Purdy KJ (2010) Evolutionary divergence and biogeography of sympatric niche-differentiated bacterial populations. *ISME J* **4**: 488–497.
- Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J & Glockner FO (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.
- Purdy KJ (2005) Nucleic acid recovery from complex environmental samples. *Method Enzymol* **397**: 271–292.
- Purdy KJ, Embley TM, Takii S & Nedwell DB (1996) Rapid extraction of DNA and rRNA from sediments by a novel hydroxyapatite spin-column method. *Appl Environ Microb* **62**: 3905–3907.
- Schloss PD & Handelsman J (2005) Introducing DOTUR, a computer program for defining operational taxonomic units and estimating species richness. *Appl Environ Microb* **71**: 1501–1506.
- Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM & Herndl GJ (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *P Natl Acad Sci USA* **103**: 12115–12120.
- You FM, Huo NX, Gu YQ, Luo MC, Ma YQ, Hane D, Lazo GR, Dvorak J & Anderson OD (2008) BatchPrimer3: a high throughput web application for PCR and sequencing primer design. *BMC Bioinformatics* **9**: 253.