Published online 18 September 2009

Nucleic Acids Research, 2009, Vol. 37, No. 21 ***e144***
doi:10.1093/nar/gkp746

Phylogenetic screening of a bacterial, metagenomic library using homing endonuclease restriction and marker insertion

Pui Yi Yung[1], Catherine Burke[1], Matt Lewis[2], Suhelen Egan[1], Staffan Kjelleberg[1] and Torsten Thomas[1,*]

[1]Centre for Marine Bio-Innovation and School of Biotechnology and Biomolecular Sciences, University of New South Wales, 2035 NSW, Australia and [2]J. Craig Venter Institute, 9704 Medical Center Drive, Rockville, MD 20850, USA

Received June 1, 2009; Revised July 8, 2009; Accepted August 24, 2009

ABSTRACT

Metagenomics provides access to the uncultured majority of the microbial world. The approaches employed in this field have, however, had limited success in linking functional genes to the taxonomic or phylogenetic origin of the organism they belong to. Here we present an efficient strategy to recover environmental DNA fragments that contain phylogenetic marker genes from metagenomic libraries. Our method involves the cleavage of 23S ribsosmal RNA (rRNA) genes within pooled library clones by the homing endonuclease I-*Ceu*I followed by the insertion and selection of an antibiotic resistance cassette. This approach was applied to screen a library of 6500 fosmid clones derived from the microbial community associated with the sponge *Cymbastela concentrica*. Several fosmid clones were recovered after the screen and detailed phylogenetic and taxonomic assignment based on the rRNA gene showed that they belong to previously unknown organisms. In addition, compositional features of these fosmid clones were used to classify and taxonomically assign a dataset of environmental shotgun sequences. Our approach represents a valuable tool for the analysis of rapidly increasing, environmental DNA sequencing information.

INTRODUCTION

Metagenomics is defined as the study of the total genomic composition of the biota from a particular environment. In recent years metagenomics has revealed many novel phylogenetic and functional properties of microbial systems (1–4) and has provided insight into the microbial communities of seawater (5–7), soil (8,9), biogas plants (10), indoor air handling unit (11) and in animal- or human-associated niches (12–15). One major aim of metagenomics is to define the functional properties of microorganisms, which are yet to be cultured or difficult to isolate, by directly analyzing their genomes from the environmental sample.

Random shotgun sequencing of environmental DNA has established itself as a core methodology in metagenomics and recent advances in high-throughput sequencing technology have made this approach economically feasible and hence widely used (16). Typically, shotgun sequencing data are functionally annotated on the level of individual sequence reads or assembled contigs. The individual read annotation and poor success in assembling complex (i.e. species-rich) samples (17) often results in functional genes being physically detached from genes containing taxonomic information (such as the 16S rRNA gene). This has limited the progress in revealing the phylotype–function relationships for many uncultivated organisms and several computational and laboratory-based approaches have been developed to overcome this situation (18). Computational methods mainly use binning of DNA fragments based on conserved compositional features (such as oligonucletotide or k-mer frequencies) followed by taxonomic classification of bins. Algorithms like PhyloPythia (19), TETRA (20) or TACOA (21) have been successfully employed; however, their performance is heavily dependent on the availability of reference (or training) sequences that are long enough to provide sufficient compositional information to define a classifier for shorter, unknown fragments (typically a few thousand base pairs in length). These reference fragments should ideally also contain well-established and universal markers for detailed phylogenetic analysis and taxonomic assignment. Fosmids and BAC clones containing phylogenetic markers such as the 16S rRNA gene have often been

*To whom correspondence should be addressed. Tel: +61 2 93853467; Fax: +61 2 93851779; Email: t.thomas@unsw.edu.au

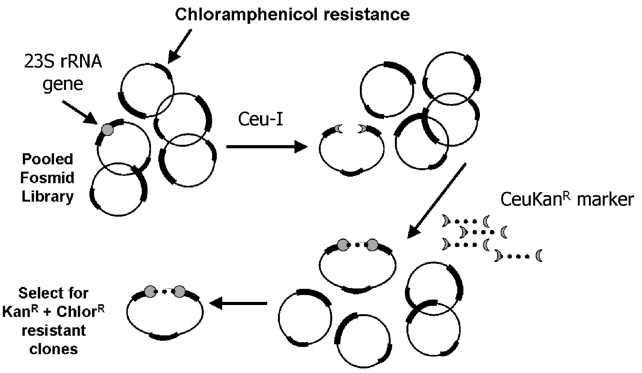

Figure 1. Schematic diagram showing the steps involved in the HERMI process.

quoted as fitting these requirements for suitable reference sequences (19,20).

Laboratory-based molecular screens for phylogenetic markers from fosmids or BAC clones have been developed and involve PCR- or fluorescent *in situ* hybridization-based screening of individually arrayed clones or small pools of clones (around 96 individual clones) (22–25). PCR-based screening of metagenomic libraries clones is frequently employed to identify phylogenetic marker genes (26–30). Potential cross-hybridization of the PCR primers or probes with the host background of the library (e.g. *Escherichia coli*) limits these approaches to either very specific organisms (e.g. targeting defined species) or to groups distantly related to the library host (e.g. screen for archaeal markers in an *E. coli* host library). Furthermore, these methods are rather labor intensive, potentially quite costly and often reach their practical limits with increasing clone numbers in the library.

Here we present an efficient phylogenetic screening strategy for metagenomic libraries using *h*oming *e*ndonuclease *r*estriction and *m*arker *i*nsertion (HERMI) (Figure 1). This approach allows rapid selection of clones containing rRNA genes from large pools by employing a restriction digest with the homing endonuclease I-*Ceu*I. The recognition site of I-*Ceu*I is 26-bp long and cuts very specifically in a conserved part of the intron-free 23S rRNA gene to produce a 4-nucleotide (CTAA) 3′ overhang (31). The slow evolutionary rate of the recognition site as well as the tolerance of the enzyme for minor sequence changes means that rRNA genes from a wide range of organisms can usually be cut (32). Hence, applying a I-*Ceu*I digest to pooled metagenomic clones should only linearize clones containing 23S rRNA genes. The linearized clones are subsequently isolated by inserting a selectable marker gene (i.e. antibiotic cassette), followed by retransformation and selection from the library pool (Figure 1).

MATERIALS AND METHODS

Bacterial strains and media

Escherichia coli EPI300 (Epicentre, Madison, WI, USA) and DH5α strains were used throughout this study. Cultures were maintained in Luria Bertani medium plus 10% NaCl (LB10) or on LB10 with 1.5% agar solid media. For clones containing only the pCC1FOS vector backbone the media was supplemented with 12.5 μg/ml chloramphenicol. For pGEM:CeuKan clones 50 μg/ml kanamycin and 100 μg/ml ampicillin and for pCC1FOS:HERMI clones 12.5 μg/ml chloramphenicol and 50 μg/ml kanamycin were used.

Construction of pooled fosmid library

Total DNA was extracted from the microbial community associated with the marine sponge *Cymbastela concentrica* (Thomas *et al*., unpublished data). A metagenomic clone library was constructed using the CopyControl™ Fosmid Library Production Kit (Epicentre, Madison, WI, USA) according to the manufacturer's protocol with minor modifications. Briefly, environmental DNA was size fractionated and DNA with molecular weight of 36-40 kb was gel excised followed by an end-repair reaction. The DNA (~1 μg) was ligated with the fosmid vector pCC1FOS followed by packaging the DNA into lambda phage particles, which were used to infect *E. coli* EPI300 cells. Transformants were plated onto LB10 agar with chloramphenicol and grown at 37°C for 16 h. The resulting library contains approximately 6500 clones with average insert size of 36 kb (data not shown). The *E. coli* library clones were pooled followed by the addition of 5 volumes of fresh LB10 broth supplemented with chloramphenicol. Induction of the fosmids to high copy number was done by addition of 0.01% (w/v) arabinose and incubating the cultures at 37°C for 4 h. After induction the cells were collected and the fosmid DNA was extracted using the Illustra plasmidPrep Mini Spin Kit (GE Bio-Science Corp, NJ, USA) according to manufacturer's instructions.

Construction of a kanamycin cassette with flanking I-*Ceu*I recognition sequences

The kanamycin cassette with flanking I-*Ceu*I recognition sequences was generated via a three-step PCR as direct PCR with primers containing the 26-bp I-*Ceu*I recognition sequence caused severe primer–dimer formation. The first round of PCR includes the amplification of the kanamycin resistance cassette from plasmid pACYC177 (NEB, Beverly, MA, USA) with flanking partial I-*Ceu*I recognition sequences. The PCR conditions as per 50 μl reaction were 10 ng of pACYC177 plasmid, 1 × iProof HF Buffer, 200 μM of each dNTP, 0.5 μM of each primer (KanFCeu/KanRCeu) (Table 1) and 0.05 U/μl iProof DNA polymerase (Bio-Rad, Hercules, CA, USA), 1 min/98°C, followed by 10s/98°C, 30 s/53°C, 1 min/72°C, 30 cycles. The second round of PCRs was performed in two separate reactions to add the entire I-*Ceu*I restriction sequence on each of the ends of the cassette; one reaction used the primer set CeuF/KanRCeu and the other used primer set CeuR/KanFCeu (Table 1). PCR conditions were: 3 ng of DNA template from the first PCR, remaining reaction mix composition as for the first-round PCR, 1 min/98°C, followed by 10 s/98°C, 30 s/45°C, 1 min 30 s/72°C, 30 cycles. The two PCR products were then pooled (final volume of 100 μl) and

Table 1. Primers used in this study

Name	Sequence (5′ to 3′)
CeuF	GGGTAACTATAACGGTCCTAAGGTAGCGAG
CeuR	GGGTCGCTACCTTAGGACCGTTATAGTTAAA
KanFCeu	GTCCTAAGGTAGCGAGGTTGATGAGAGCTTT GTTG
KanRCeu	AGGACCGTTATAGTTAAAGTCAGCGTAATG CTCTGC
KanFSeq	GACCGTTCCGTGGCAAAG
KanRSeq	GCTCGATGAGTTTTTCTAATC
23S–129F	CYGAATGGGRVAACC
23S–2241R	ACCGCCCCAGTHAAACT
16S–27F	AGAGTTTGATCMTGGCTCAG
16S–1492R	ACGGTTACCTTGTTACGACTT
pEpiFosFor	GGATGTGCTGCAAGGCGATTAAGTTGG
pEpiFosRev	CTCGTATGTTGTGTGGAATTGTGAGC CCTACGGGAGGCAGCAG
907R	CCGTCAATTCCTTTGAGTTT
GM5FplusGC	CCTACGGGAGGCAGCAGCGCCCGCCGCGCC CCGCGCCCGTCCCGCCGCCCCCGCCCG

the 1.1 kb product was gel extracted to purify the kanamycin cassette. The third round of PCR involved hybridizations and fill-in reaction steps to generate a PCR product with the full I-*Ceu*I recognition site on both ends of the kanamycin cassette. The conditions were: ~50 µl of gel-extracted DNA from the second round of PCR, 1 × RedTaq PCR buffer, 10 µM of each dNTP and 1 U of RedTaq DNA polymerase (Promega, Madison, WI, USA), 2 min/94°C, 20 s/94°C, 30 s/45°C, 20 s/72°C, 20 cycles. The final PCR product was cloned into pGEM vector (pGEM®-T Easy Vector System, Promega, Madison, WI, USA) according to manufacturer's instructions followed by the transformation into *E. coli* DH5α. Transformants (named pGEM:CeuKan) with the correct insert (after confirmation via restriction digest with I-*Ceu*I) were selected for miniprep purification. The kanamycin cassette with flanking I-*Ceu*I recognition sequences (CeuKan) was gel purified after digestion of pGEM:CeuKan.

Homing endonuclease restriction and marker insertion

One microgram of pooled fosmid DNA was digested with the I-*Ceu*I homing endonuclease in a reaction containing two units of I-*Ceu*I enzyme (NEB), 1 × NEB buffer 4, 0.5 µl of BSA (NEB) and sterile, deionized water up to a volume of 50 µl. The digest was incubated at 37°C for 3 h followed by a heat inactivation at 65°C for 20 min. The digested fosmids were dephosphorylated using two units of Antarctic alkaline phosphatase (NEB) according to manufacturer's instructions. The fosmids were ligated with the CeuKan cassette in 1:30 (vector:insert) molar ratio overnight at room temperature. The ligation mix was transformed into *E. coli* EPI300 cells via electroporation and cells were recovered at 37°C for 1 h and then plated onto LB10 agar supplemented with appropriate antibiotics to select for pCC1FOS:HERMI clones. Transformants that grew on the selective agar were purified, fosmid DNA was extracted and subjected to denaturing gradient gel electrophoresis (DGGE) analysis, end sequencing and 16S/23S rRNA gene PCR as described below.

Denaturing gradient gel electrophoresis

The 16S rRNA gene was PCR amplified as described by Muyzer *et al.* (33) with the universal primer GM5FplusGC and 907RC (Table 1). Various DNA samples were used: (i) Pooled fosmid library DNA, (ii) *E. coli* genomic DNA and (iii) fosmid DNA of individual HERMI clones. PCR conditions were 10 ng of DNA template, 1 × RedTaq buffer, 0.5 µM of each forward and reverse primers, 200 µM of each dNTP, 300 µg of BSA, 1 U of RedTaq DNA polymerase (Promega, Madison, WI, USA), 3 min/96°C, hot start at 80°C, 30 s/94°C, 30 s/57°C, 1 min 10 s/72°C, 25 cycles. The PCR products were cleaned using the QIA quick PCR purification kit (QIAGEN, Hilden, Germany) and the DNA was examined with a DCode DGGE unit (BIO-RAD, Hercules, CA, USA) using the following parameters: 10% acrylamide gel, a denaturant gradient containing 45–60% urea-formamide, 1× TAE buffer, 75 V at 60°C for 16 h. Bands from the DGGE gel were extracted, dialysed overnight at 4°C with 50 µl of molecular grade water and re-amplified using primers GM5F and 907RC for sequencing.

Sequencing and phylogenetic analysis of fosmid clones

End sequencing of the HERMI clones were performed using the primer pair pEpiFosFor and pEpiFosRev (Table 1). PCR amplification and sequencing using universal primers for 23S and 16S rRNA gene (Table 1) were also performed on selected HERMI clones as described previously (34,35). Briefly, PCR conditions were 3 min/94°C, 1 min/94°C, 1 min/57°C, 3 min/72°C, 30 cycles (for 23S PCR) and 3 min/94°C, 80°C hot start, followed by 30 s/94°C, 1 min/50°C, 3 min/72°C, 25 cycles (for 16S PCR). The PCR products were subjected to sequencing using the same primers. Other sequencing reactions were also performed using the KanFSeq and KanRSeq primers (Table 1) to obtain 23S rRNA gene sequence flanking the kanamycin cassette.

The complete sequence of PCR products were obtained and searched with the BLAST algorithm (36) against the NCBI and Silva database (37) and closest representatives were selected, aligned using the Aligner tool provided in Silva, and imported into the Silva 16S rRNA and 23S rRNA database using the ARB program for phylogenetic tree construction (38). Maximum likelihood trees were constructed with default parameters.

Whole fosmid sequencing and analysis

HERMI fosmid clones were shotgun sequenced as outlined in Rusch *et al.* (5). Shotgun reads were processed and assembled using Phred/Phrap and the assembly was manually checked in Consed (39). Binning of the fosmid sequences and an assembled shotgun-reads dataset from the bacterial community of *C. concentrica* (Thomas *et al.*, unpublished data) was done according to the strategy outlined by Woyke *et al.* (40). Briefly, tetranucleotide patterns were determined using TETRA (20) and

exported as normalized *Z*-scores for fosmids and scaffolds of the shotgun datasets longer than 20 kb. Clustering was performed with Euclidian distance and complete linkage using the software Cluster 3.0 and visualized with JavaTreeView (41).

RESULTS

Specificity of the HERMI screen

Applying the HERMI method in a single-tube reaction to a pooled library of approximately 6500 fosmid clones derived from the microbial community associated with the sponge *C. concentrica* resulted in 52 kanamycin-resistant transformants. End sequencing of these clones rejected 40 as being sister clones leaving 12 for further analysis. PCR amplification and sequencing of the region flanking the CeuKan cassette showed that none of the clones contains more than one CeuKan cassette. In addition, all insertion sites have similarity to the 23S rRNA gene indicating that no unspecific insertion had taken place. The I-*Ceu*I recognition sequences for four of the HERMI clones that contained unique phylotype (HERMI06, 11, 16 and 30) were compared with the I-*Ceu*I recognition sequence described in (32). This demonstrated that HERMI06, 11 and 30 have perfect matches to the recognition sequence (Figure 2), while HERMI16 had base pair insertions between positions 4–5, 5–6, 6–7 and 7–8 (respective to the original recognition sequence). This indicates that base pair insertions in these positions of the recognition sequence do not abolish I-*Ceu*I cleavage activity.

To evaluate the phylogenetic spectrum of sequences that can be cleaved by I-*Ceu*I, we compared the 26-bp recognition sequence to the entire bacterial 23S rRNA Silva database (37) with various thresholds (no mismatch, 1 mismatch and 2 or more mismatches). Partial sequences (those did not cover the entire region of the I-*Ceu*I recognition sequence), mitochondrial and chloroplast DNA, and sequences belonging to the actinobacteria phyla [previously shown not to be cleaved by I-*Ceu*I (42)] were excluded. We also considered sequences containing mismatches that were previously shown to abolish the activity of the enzyme in our comparison (31). Our analysis showed that only 1.1% of the 23S rRNA gene sequences (59 sequences out of 5525 sequences) contain mismatches that can potentially abolish the enzymatic activity. Phylotypes that might be excluded from our screening approach include a number of species within the genus *Chlamydophila* and *Chlamydia*, the unclassified gamma-proteobacteria *Candidatus Carsonella ruddii*, the Candidate division TM7 group GTL1 and a number of species in the genus *Chloroflexi* and *Roseiflexus* (all one mismatch) as well as some species within the *Planctomycete, Coxiella burnetii* and *Francisella tularensis* subsp. (two or more mismatches) (Figure 2). This updated, theoretical consideration is consistent with previous observations (43) and given the degree of mismatch tolerance described for homing endonucleases (44,45), we estimate that our HERMI strategy is capable of capturing at least 98% of the known diversity from a wide phylogenetic spectrum of 23S rRNA gene sequences.

Phylogenetic analysis of HERMI clones from *C. concentrica*

The 12 unique HERMI clones (defined as clones having unique fosmid end sequences) were subjected to 16S and 23S rRNA gene sequencing and their phylogenetic identities were examined. Based on 16S rRNA sequences four unique phylogenetic groups (with a percentage identity cut-off of 99%) were identified from these 12 clones. These four unique 16S rRNA gene sequences

I-CeuI	TAAC-T-A-T-AACGGTCCTAA ▼	GGTAGCGA
HERMI06	TAAC-T-A-T-AACGGTCCTAA	GGTAGCGA
HERMI11	TAAC-T-A-T-AACGGTCCTAA	GGTAGCGA
HERMI30	TAAC-T-A-T-AACGGTCCTAA	GGTAGCGA
HERMI16	TAAC**A**T**C**AA**T**GAACGGTCCTAA	GGTAGCGA
1 mismatch		
Chloroflexi & *Roseiflexus*	TAAC-T-A-T-AAC**A**GTCCTAA	GGTAGCGA
Candidate division TM7	TAAC-T-A-T-AAC**C**GTCCTAA	GGTAGCGA
Chlamydophila & *Chlamydia* sp.	TAAC-T-A-T-AACGGT**G**CTAA	GGTAGCGA
Candidatus Carsonella ruddii	TAAC-T-A-T-AACGGTCC**A**AA	GGTAGCGA
Plantomycetes sp.	TAAC-T-A-T-AA**G**GGTCCTAA	GGTAGCGA
2 or more mismatches		
Coxiella burnetii	TAAC-T-A-T-AAA**T**TAA**C**CG**T**	**T**GTAGCGA
	TAAC-T-A-T-AG**A**T**C**TCCTAA	GGTTGCGA
Francisella tularensis subsp.	TAAC-T-A-T-A**C**CGGTC**G**TAA	GGTTGCGA
	TAAC-T-A-T-A**A**CGGTCCTAA	GG**AT**GCGA

Figure 2. Alignment (5′ to 3′) of the I-*Ceu*I recognition sites with insertion sites of HERMI clones and different sequences from the Silva LSU 23S rRNA database. Black triangle represents the cleavage site for I-*Ceu*I and where the CeuKan marker was inserted for the HERMI clones. Letters underlined in the recognition sequence of HERMI16 indicate basepair insertions into the recognition sequence while letters underlined for other sequences highlight mismatches in positions that have previously been shown to negatively impact I-*Ceu*I enzymatic activity.

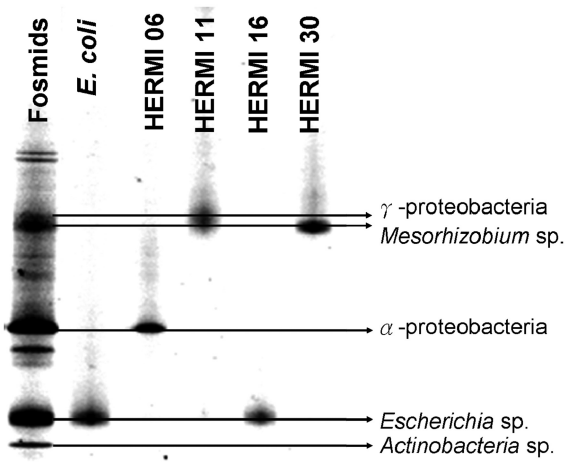

Figure 3. DGGE gel of the 16S rRNA gene of a pooled fosmid library (lane 1), *E. coli* genomic DNA (lane 2) and HERMI clones (lanes 3–6). Individual bands were excised and sequenced; the taxonomic assignment of the band is shown next to the arrow.

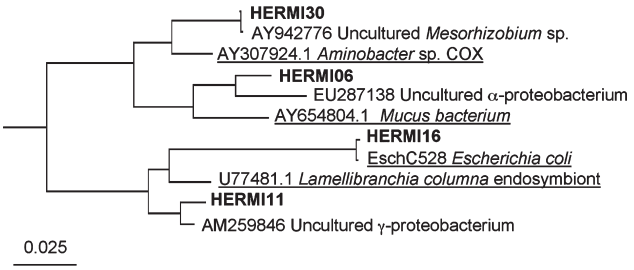

Figure 4. Maximum likelihood tree of the 16S rRNA gene for the HERMI clones (bold). and the closest representatives of cultured (from the RDP database, underlined) and uncultured organisms (from Silva SSU database) with accession numbers preceding the name of the organism. Scale bar indicates 0.025 sequence divergence.

(using clone HERMI06, 11, 16 and 30 as representatives) were also found by band sequencing and migration distance to correspond to the four dominant bands of a 16S rRNA gene DGGE analysis of the pooled fosmids. This indicates that we have captured the major phylotypes present in the library (Figure 3). One minor band of the pooled fosmids was found to belong to the group *Actinobacteria*, which is known not to be cleaved by the I-*Ceu*I enzyme (see above).

A maximum likelihood tree was constructed for the 16S rRNA sequences of the HERMI clones (Figure 4) and shows that all sequences (except HERMI16, see below) are related to uncultured organisms and at least 7.5% divergent from their nearest cultured bacterium. Taxonomic classification assigned them to novel clades within the α- and γ-proteobacteria group (HERMI06 and HERMI11) and a *Mesorhizobium*-related group (HERMI30). The 16S rRNA sequence of HERMI16 is identical to *E. coli*, but further analysis of this fosmid showed that only the 23S rRNA was cloned and that the 16S rRNA sequence resulted from amplification of *E. coli* host background. A maximum likelihood tree

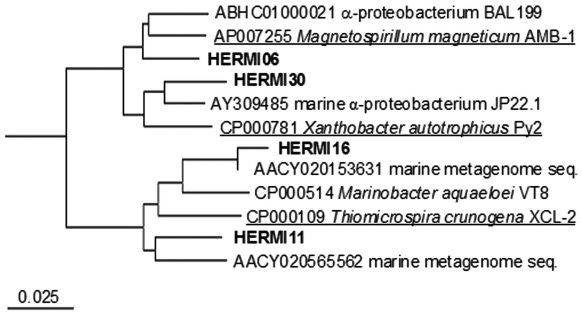

Figure 5. Maximum likelihood tree of the 23S rRNA gene for the HERMI clones (bold) and the closest representatives from the Silva 23S LSU database. Underlined sequence highlights closest cultured representatives. Scale bar indicates 0.025 sequence divergence.

was also constructed for the 23S rRNA gene of the HERMI clones (Figure 5) and their closest representatives imported from the 23S rRNA Silva database (37). Here the HERMI clones are at least 11% divergent to their closest cultured representatives. The longer branch length in this tree to the next related sequences is a reflection of sparseness of available 23S rRNA gene data, but nevertheless allows us to link the 16S rRNA with 23S rRNA gene phylogeny (Figure 5).

Binning and taxonomic assignment of shotgun metagenomic sequences

To illustrate the value of whole fosmid sequences containing 16S or 23S rRNA genes for the assignment of unknown DNA fragments, we clustered the whole sequence of HERMI11 with sequences longer than 20 kb from an assembled shotgun sequence dataset of the bacterial community of *C. concentrica*. As illustrated in Figure 6, the HERMI11 clone sequence fell into a cluster of unknown sequences (red branches and yellow box) with a high correlation index (0.884). We therefore can now assign the shotgun fragments in this cluster to the novel phylotype within the γ-proteobacteria and link this taxon with 372 predicted protein contained within these DNA sequences.

DISCUSSION

In this study we have shown that HERMI is a suitable tool for screening a metagenomic library for clones with phylogenetic markers (i.e. the 23S rRNA gene). Handling time for the HERMI procedure is less than 2 h and allows for the simultaneous screening of thousands of clones in a single-tube reaction, hence making our approach faster and more convenient than other phylogenetic marker screens (22–25). We also note that the selectable marker applied in this study (i.e. kanamycin resistance marker) could be replaced with any marker cassette. For example, a green fluorescent protein cassette could be employed allowing for subsequent screening of clones with fluorescence-activated cell sorting (FACS).

Nesbø *et al.* (42) also recently used I-*Ceu*I to clone fragments containing 23S rRNA genes from

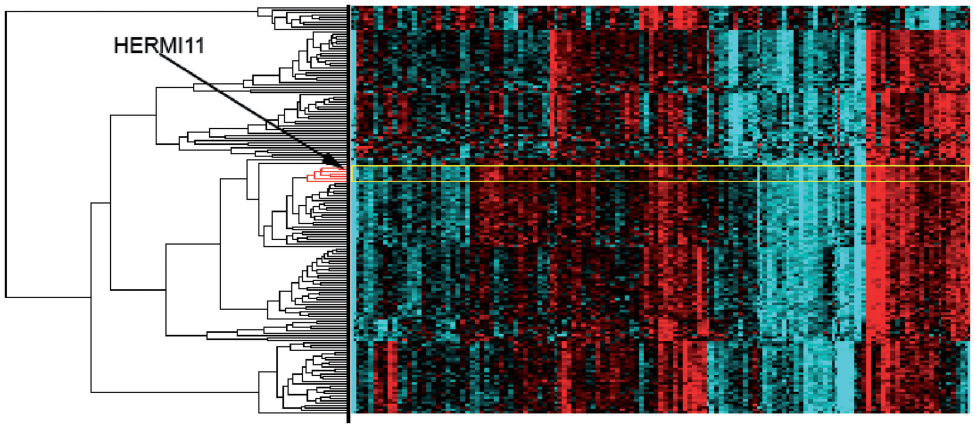

Figure 6. Clustering of HERMI11 fosmid sequence with assembled shotgun sequences of unknown taxonomic assignment. *Arrow* indicates the position of HERMI11 in a cluster of sequences (correlation of 0.884) that is highlighted by red branches and a yellow box. Cyan and red colors in the data matrix represent over- and under-representation of tetranucleotide *Z*-scores.

environmental DNA samples. Their strategy involved the construction of a specialized vector that allowed for the insertion of DNA digested with a blunt-end restriction enzyme and I-*Ceu*I. However, their approach did not yield full-length 23S and 16S rRNA genes, limiting the phylogenetic and taxonomic value of their clones. The HERMI approach does not have this limitation and, as 16S and 23S rRNA genes are often tightly clustered together, will frequently yield clones with 16S rRNA information, for which more comprehensive databases are currently available. We also note that the linkage of 16S and 23S rRNA information will also help to improve annotation of unknown clades in current 23S rRNA databases, as illustrated by the phylogenetic analysis in our study (Figure 4 and 5). Another benefit of the HERMI strategy is that it can be applied retrospectively to any existing environmental DNA library, many of which have so far only been screened for expressed, functional genes (46–48). Screening those libraries for phlyogenetic markers will generate more information about compositional-features of the DNA from uncultured organisms, which in turn will improve our ability to classify bins of environmental DNA sequences (Figure 6) and hence reveal more phylotype–function relationship.

Sequence analysis of the 26-bp recognition site in the Silva 23S rRNA database indicated that only a small proportion (<2%) of all currently known 23S rRNA genes might not be captured with our approach. Importantly, this limitation can be quantified for each sample prior to the screening, for example, by constructing a 23S or 16S rRNA gene PCR library and determine its phylogenetic composition. This is typically done as a first step for most environmental diversity studies (49) and hence would not impose an additional work load.

Finally, homing endonucleases are not only restricted to cleaving the 23S rRNA genes. Arnould and coworkers have recently demonstrated that a semi-rational design can yield I-*Cre*I mutant homing endonucleases with highly specific recognition patterns, which are distinct from the wild-type activity (50). As such other homing endonuclease mutants are available or could be designed that target other common genes used for phylogenetic analysis (e.g. the 16S rRNA gene, *recA*), functional genes or non-coding regions, and which could then be applied in a analogous screening strategy as described here. This will further enhance the utility and versatility of the HERMI approach.

ACCESSION NUMBERS

GQ160460–GQ160467.

ACKNOWLEDGEMENTS

The authors thank Robert Friedman and acknowledge the J. Craig Venter Institute (JCVI) Joint Technology Center, under the leadership of Yu-Hui Rogers, for producing whole-fosmid sequence data. The shotgun sequencing of *Cymbastela concentrica* is available at GenBank under GenomeProject ID 34751 and at the Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis (CAMERA) website (http://camera.calit2.net/). The ribosomal RNA gene and fosmid sequences reported here have been deposited in Genbank under accessions GQ160459–GQ160467.

FUNDING

The Gordon and Betty Moore Foundation and the Australian Research Council (LP0668235). Funding for open access charge: internal funds from the Centre for Marine Bio-Innovation.

Conflict of interest statement. None declared.

REFERENCES

1. Ferrer,M., Beloqui,A., Timmis,K.N. and Golyshin,P.N. (2009) Metagenomics for mining new genetic resources of microbial communities. *J. Mol. Microbiol. Biotechnol.*, **16**, 109–123.

2. Handelsman,J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev.*, **68**, 669–685.
3. Li,X. and Qin,L. (2005) Metagenomics-based drug discovery and marine microbial diversity. *Trends Biotechnol.*, **23**, 539–543.
4. Streit,W.R. and Schmitz,R.A. (2004) Metagenomics – the key to the uncultured microbes. *Curr. Opin. Microbiol.*, **7**, 492–498.
5. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M., Remington,K. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77:0398–0431.
6. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Nature*, **304**, 66–74.
7. Martin-Cuadrado,A.B., Rodriguez-Valera,F., Moreira,D., Alba,J.C., Ivars-Martínez,E., Henn,M.R., Talla,E. and López-García,P. (2008) Hindsight in the relative abundance, metabolic potential and genome dynamics of uncultivated marine archaea from comparative metagenomic analyses of bathypelagic plankton of different oceanic regions. *ISME J.*, **2**, 865–886.
8. Allen,H.K., Moe,L.A., Rodbumrer,J., Gaarder,A. and Handelsman,J. (2009) Functional metagenomics reveals diverse beta-lactamases in a remote Alaskan soil. *ISME J.*, **3**, 243–251.
9. Daniel,R. (2005) The metagenomics of soil. *Nat. Rev. Microbiol.*, **3**, 470–478.
10. Schlüter,A., Bekel,T., Diaz,N.N., Dondrup,M., Eichenlaub,R., Gartemann,K.H., Krahn,I., Krause,L., Krömeke,H., Kruse,O. *et al.* (2008) The metagenome of a biogas-producing microbial community of a production-scale biogas plant fermenter analysed by the 454-pyrosequencing technology. *J. Biotechnol.*, **136**, 77–90.
11. Tringe,S.G., Zhang,T., Liu,X., Yu,Y., Lee,W.H., Yap,J., Yao,F., Suan,S.T., Ing,S.K., Haynes,M. *et al.* (2008) The airborne metagenome in an indoor urban environment. *PLoS ONE*, **3**, e1862.
12. Fieseler,L., Quaiser,A., Schleper,C. and Hentschel,U. (2006) Analysis of the first genome fragment from the marine sponge-associated, novel cadidate phylum *Poribacteria* by environmental genomics. *Environ. Microbiol.*, **8**, 612–624.
13. Jones,B.V., Begley,M., Hill,C., Gahan,C.G. and Marchesi,J.R. (2008) Functional and comparative metagenomic analysis of bile salt hydrolase activity in the human gut microbiome. *Proc. Natl Sci. USA*, **105**, 13580–13585.
14. Warnecke,F., Luginbühl,P., Ivanova,N., Ghassemian,M., Richardson,T., Stege,J., Cayouette,M., McHardy,A., Djordjevic,G., Aboushadi,N. *et al.* (2007) Metagenomic and functional analysis of hindgut microbiota of a wood-feeding higher termite. *Nature*, **450**, 560–565.
15. Dinsdale,E., Edwards,R., Hall,D., Angly,F., Breitbart,M., Brulc,J., Furlan,M., Desnues,C., Haynes,M., Li,L. *et al.* (2008) Functional metagenomic profiling of nine biomes. *Nature*, **452**, 629–632.
16. Mardis,E. (2008) Next-generation DNA sequencing methods. *Annu. Rev. Genomics Hum Genet.*, **9**, 387–402.
17. Tringe,S., von Mering,C., Kobayashi,A., Salamov,A., Chen,K., Chang,H., Podar,M., Short,J., Mathur,E., Detter,J. *et al.* (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554–557.
18. Kunin,V., Copeland,A., Lapidus,A., Mavromatis,K. and Hugenholtz,P. (2008) A bioinformatician's guide to metagenomics. *Microbiol. Mol. Biol. Rev.*, **72**, 557–578.
19. McHardy,A.C., Martin,H.G., Tsirigos,A., Hugenholtz,P. and Rigoutsos,I. (2007) Accurate phylogenetic classification of variable-length DNA fragments *Nat. Methods*, **4**, 63–72.
20. Teeling,H., Waldman,J., Lombardot,T., Bauer,M. and Glöckner,F. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**, 163.
21. Diaz,N., Krause,L., Goesmann,A., Niehaus,K. and Nattkemper,T. (2009) TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformatics*, **10**, 56.
22. Beja,O., Suzuki,M.T., Koonin,E.V., Aravind,L., Hadd,A., Nguyen,L.P., Villacorta,R., Amjadi,M., Garrigues,C., Jovanovich,S.B. *et al.* (2000) Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environ. Microbiol.*, **2**, 516–529.
23. Moreira,D., Rodríguez-Valera,F. and López-García,P. (2006) Metagenomic analysis of mesopelagic Antarctic plankton reveals a novel deltaproteobacterial group. *Microbiology*, **152**, 505–517.
24. Stein,J.L., Marsh,T.L., Yu,K.Y., Shizuya,H. and DeLong,E.F. (1996) Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *J. Bacteriol.*, **178**, 591–599.
25. Liles,M.R., Manske,B.F., Bintrim,S.B., Handelsman,J. and Goodman,R.M. (2003) A census of rRNA genes and linked genomic sequences within a soil metagenomic library. *Appl. Environ. Microbiol.*, **69**, 2684–2691.
26. Gilbert,J.A., Mühling,M. and Joint,I. (2008) A rare SAR11 fosmid clone confirming genetic variability in the 'Candidatus Pelagibacter ubique' genome. *ISME J.*, **2**, 106–116.
27. Meng,J., Wang,F., Zheng,Y., Peng,X., Zhou,H. and Xiao,X. (2009) An uncultivated crenarchaeota contains functional bacteriochlorophyll a synthase. *ISME J.*, **3**, 106–116.
28. Martín-Cuadrado,A., López-García,P., Alba,J., Moreira,D., Monticelli,L., Strittmatter,A., Gottschalk,G. and Rodríguez-Valera,F. (2007) Metagenomics of the deep Mediterranean, a warm bathypelagic habitat. *PLoS ONE*, **2**, e914.
29. Quaiser,A., López-García,P., Zivanovic,Y., Henn,M., Rodriguez-Valera,F. and Moreira,D. (2008) Comparative analysis of genome fragments of Acidobacteria from deep Mediterranean plankton. *Environ. Micro.*, **10**, 2704–2717.
30. Massana,R., Karniol,B., Pommier,T., Bodaker,I. and Béjà,O. (2008) Metagenomic retrieval of a ribosomal DNA repeat array from an uncultured marine alveolate. *Environ. Micro.*, **10**, 1335–1343.
31. Marshall,P. and Lemieux,C. (1992) The I-CeuI endonuclease recognizes a sequnce of 19 base pairs and preferentially cleaves the coding strand of the *Chlamydomonas moewusii* chloroplast large subunit rRNA gene. *Nucleic Acids Res.*, **20**, 6401–6407.
32. Marshall,P. and Lemieux,C. (1991) Cleavage pattern of the homing endomuclease encoded by the fifth intron in the chloroplast large subunit rRNA-encoding gene of *Chlamydomonas eugametos*. *Gene*, **104**, 241–245.
33. Muyzer,G., Brinkhoff,T., Nübel,U., Santegoeds,C., Schäfer,H. and Wawer,C. (1997) In Akkermans,A.D.L., van Elsas,J.D. and de Bruijn,F.J. (eds), *Molecular Microbial Ecology Manual*, Vol. 3.4.4. Kluwer, Dordrecht, The Netherlands, pp. 1–27.
34. Hunt,D.E., Klepac-Ceraj,V., Acinas,S.G., Gautier,C., Bertilsson,S. and Polz,M.F. (2006) Evaluation of 23S rRNA PCR primers for use in phylogenetic studies of bacterial diversity. *Appl. Environ. Microbiol.*, **72**, 2221–2225.
35. Lane,D.J. (1991) 16S/23S rRNA sequencing. In Stachebrandt,E. and Goodfellow,M. (eds), *Nucleic Acid Techniques in Bacterial Systematics*. Wiley, Chichester, New York, pp. 115–175.
36. Altschul,S., Gish,W., Miller,W., Myers,E. and Lipman,D. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
37. Pruesse,E., Quast,C., Knittel,K., Fuchs,B., Ludwig,W., Peplies,J. and Glöckner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.*, **35**, 7188–7196.
38. 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.
39. Gordon,D., Abajian,C. and Green,P. (1998) Consed: a graphical tool for sequence finishing. *Genome Res.*, **8**, 195–202.
40. Woyke,T., Teeling,H., Ivanova,N., Huntemann,M., Richter,M., Gloeckner,F., Boffelli,D., Anderson,I., Barry,K., Shapiro,H. *et al.* (2006) Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature*, **443**, 950–955.
41. Eisen,M., Spellman,P., Brown,P. and Botstein,D. (1998) Cluster analysis and display of genome-wide expression patterns. *Proc. Natl Acad. Sci. USA*, **95**, 14863–14868.
42. Nesbø,C.L., Boucher,Y., Dlutek,M. and Doolittle,W.F. (2005) Lateral gene transfer and phylogenetic assignment of environmental fosmid clones. *Environ. Microbiol.*, **7**, 2011–2026.

43. Liu,S.-H., Hessel,A. and Sanderson,K.E. (1993) Genomic mapping with I-Ceu I, an intron-encoded endonuclease specific for genes for ribosomal RNA, in Salmonella spp., Escherichia coli, and other bacteria *Proc. Natl Acad. Sci. USA*, **90**, 6874–6878.
44. Aagaard,C., Awayez,M.J. and Garrett,R.A. (1997) Profile of the DNA recognition site of the archaeal homing endonuclease I-DmoI. *Nucleic Acids Res.*, **25**, 1523–1530.
45. Jurica,M.S., Monnat,R.J. Jr and Stoddard,B.L. (1998) DNA recognition and cleavage by the LAGLIDADG homing endonuclease i-crei. *Mol. Cell*, **2**, 469–476.
46. Lim,H.K., Chung,E.J., Kim,J.C., Choi,G.J., Jang,K.S., Chung,Y.R., Cho,K.Y. and Lee,S.W. (2005) Characterization of a forest soil metagenome clone that confers indirubin and indigo production on *Escherichia coli*. *Appl. Environ. Microbiol.*, **71**, 7768–7777.
47. Rondon,M.R., August,P.R., Bettermann,A.D., Brady,S.F., Grossman,T.H., Liles,M.R., Loiacono,K.A., Lynch,B.A., MacNeil,I.A., Minor,C. *et al.* (2000) Cloning the soil metagenome: a strategy for accessing the genetic and functional diversity of uncultured microorganisms. *Appl. Environ. Microbiol.*, **66**, 2541–2517.
48. MacNeil,I.A., Tiong,C.L., Minor,C., August,P.R., Grossman,T.H., Loiacono,K.A., Lynch,B.A., Phillips,T., Narula,S., Sundaramoorthi,R. *et al.* (2001) Expression and isolation of antimicrobial small molecules from soil DNA libraries. *J. Mol. Microbiol. Biotechnol.*, **3**, 301–308.
49. Tringe,S. and Hugenholtz,P. (2008) A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.*, **11**, 442–446.
50. Arnould,S., Chames,P., Perez,C., Lacroix,E., Duclert,A., Epinat,J., Stricher,F., Petit,A., Patin,A., Guillier,S. *et al.* (2006) Engineering of large numbers of highly specific homing endonucleases that induce recombination on novel DNA targets. *J. Mol. Biol.*, **355**, 443–458.