MOLECULAR ECOLOGY
RESOURCES
Molecular Ecology Resources (2012) doi: 10.1111/j.1755-0998.2012.03166.x

A fast, highly sensitive double-nested PCR-based method to screen fish immunobiomes

SÉBASTIEN BOUTIN[1], MAELLE SEVELLEC[1], SCOTT A. PAVEY, LOUIS BERNATCHEZ and NICOLAS DEROME
Institut de Biologie Intégrative et des Systèmes (IBIS), Département de Biologie, Université Laval, 1030 Avenue de la Médecine – Québec, QC, Canada G1V 0A6

Abstract

Efficient methods for constructing 16S tag amplicon libraries for pyrosequencing are needed for the rapid and thorough screening of infectious bacterial diversity from host tissue samples. Here we have developed a double-nested PCR methodology that generates 16S tag amplicon libraries from very small amounts of bacteria/host samples. This methodology was tested for 133 kidney samples from the lake whitefish *Coregonus clupeaformis* (Salmonidae) sampled in five different lake populations. The double-nested PCR efficiency was compared with two other PCR strategies: single primer pair amplification and simple nested PCR. The double-nested PCR was the only amplification strategy to provide highly specific amplification of bacterial DNA. The resulting 16S amplicon libraries were synthesized and pyrosequenced using 454 FLX technology to analyse the variation of pathogenic bacteria abundance. The proportion of the community sequenced was very high (Good's coverage estimator; mean = 95.4%). Furthermore, there were no significant differences of sequence coverage among samples. Finally, the occurrence of chimeric amplicons was very low. Therefore, the double-nested PCR approach provides a rapid, informative and cost-effective method for screening fish immunobiomes and most likely applicable to other low-density microbiomes as well.

Keywords: Coregonus clupeaformis, immunobiome, next-generation sequencing, pathogens, pyrosequencing

Received 17 January 2012; revision received 8 May 2012; accepted 15 May 2012

Introduction

Diagnosis of bacterial infection is often limited to consider the presence and/or abundance of a specific pathogen identified a priori in host tissues. Thanks to the development of novel high-throughput sequencing technologies, we can now explore entire bacterial communities with no prior information. Rapid and accurate characterization of unknown bacterial consortia requires documenting both taxonomic diversity and structure (i.e. relative abundance of every bacterial strain or ribotype). Reaching these two objectives implies first selecting a molecular marker specific to bacteria and archae. This marker should contain both variable and conserved regions. These properties allow both exclusion of host DNA amplification and accurate discrimination among species of bacteria. The 16S ribosomal RNA subunit gene meets these two criteria, thus making it the marker of choice in microbial ecology (Amann & Ludwig 2000). Second, describing the taxonomical dynamics of a bacterial community necessitates using both quantitative and qualitative approaches: characterizing taxonomic diversity and quantifying the relative abundance of the different species.

Pyrosequencing PCR products of the 16S ribosomal gene are currently the best method for simultaneous qualitative and quantitative screening of complex or unknown bacterial populations (Amend *et al.* 2010; Zhang *et al.* 2011). First, this procedure facilitates a more in-depth analysis of the community compared with other molecular techniques, such as cloning or denaturing gradient gel electrophoresis (DGGE) and can even detect rare microflora (Petrosino *et al.* 2009). Second, it allows the simultaneous quantification of the relative amount of each unique ribotype. Furthermore, this method provides the opportunity to screen many samples in parallel by tagging individual samples with specific barcodes (Humblot & Guyot 2009; Engelbrektson *et al.* 2010). Finally, next-generation sequencing is both time and cost-effective for community profiling.

The most informative region of 16S ribosomal RNA gene is the hypervariable 2 and 3 (V2 and V3) regions,

Correspondence: Nicolas Derome, Fax: (+1) 418 656 2043;
E-mail: nicolas.derome@bio.ulaval.ca
[1]S.B. and M. S. share equal first co-authorship.

which can efficiently discriminate taxa down to the generic level. For instance, Chakravorty *et al.* (2007) showed that the V2 region or V3 region were suitable for identification to the genus level in 110 bacteria samples. Thus, PCR-based methods targeting the V3 regions are widely used for profiling microbial communities (16S + V3 region + bacterial community = 181 hits in the Web of Knowledge, March 2012).

A major concern with these methods is the contamination by environmental bacteria and/or eukaryotic DNA (Ampe *et al.* 1999; Humblot & Guyot 2009). Bacterial contamination is mostly owing to contamination by handlers, PCR products or water (Bottger 1990; Rand & Houck 1990; Schmidt *et al.* 1991; Von Wintzingerode *et al.* 1997; Kulakov *et al.* 2002). Eukaryotic contamination occurs frequently with environmental samples and samples coming from host tissues (Humblot & Guyot 2009; Petrosino *et al.* 2009). Although eukaryotic DNA should theoretically not affect 16S rRNA gene-based methods, several studies reported that primers commonly used to target the 16S gene also amplify the eukaryotic 18S rRNA gene (Ampe & Miambi 2000).

In this study, we developed a new method for building a 454 library with small amounts of bacteria/host DNA to cope with problem of eukaryotic contamination and small parasite/host DNA ratio. In particular, we were interested in developing an approach to diagnose the presence and relative abundance of putative pathogenic bacterial communities in fish kidney tissue from different wild populations. It is generally assumed that healthy fish should have a kidney free from bacteria such that the presence of bacteria in this tissue is interpreted as evidence of pathogen infection (Cahill 1990; Uhland *et al.* 2000; Dionne *et al.* 2009).

Materials and methods

Sample collection

Lake whitefish (*Coregonus clupeaformis*) samples were collected using gill nets in five different lakes (Cliff, East, Témiscouata, Webster and Indian) from the St John River drainage: two of these lakes are in Québec, Canada (East and Témiscouata), and three from Maine, United States (Cliff, Webster and Indian Pond). The sampling was undertaken between 15 June and 15 July 2010, as bacterial infections has been shown in other studies to be highest at this time of the year (Larsen *et al.* 2004; Dionne *et al.* 2009). Fish were dissected in sterile conditions, and kidneys were individually stored in a sterile Eppendorf © tube and flash-frozen in liquid nitrogen. Kidney samples were then transported to the laboratory and kept at −80 °C.

Bacterial analysis

Extraction. Kidney tissue DNA was extracted in a laminar flow cabinet, using the QIAamp© DNA mini kit (Qiagen) with sterile tools and supplies. The protocol used for this extraction was a modified from QIAamp© DNA mini kit protocol for tissues. A lysozyme step (lysozyme = 4 μL (100 mg/mL) of 30 min at 37 °C was added before the proteinase K step to break down cell walls of gram+ bacteria. Each extraction product was eluted with 60 μL of sterile water. DNA extractions were quantified with a Nanodrop® (Thermo Scientific) which revealed DNA concentrations higher than 500 ng/μL and displaying satisfying ratios ($mean_{280/260}$ = 1.7; $mean_{280/230}$ = 2.7)

PCR approaches. Three different techniques were tested to assess the reproducibility of amplifying bacteria in kidney tissue: Double-nested PCR, classical one-step PCR and single-nested two-step PCR. First, a double-nested PCR was performed. This technique was based on the nested PCR method developed by Yourno (1992). A nested PCR is an association of two PCR amplifications, the first amplification product being used as template for the second one. The primers used in the second step hybridize on the amplicon generated by the previous PCR step. Basically, our double-nested PCR contained three successive amplification steps, which were conducted with three different primer pairs: 1389R-9F, 907R-23F and 519R-63F (Table 1) (Lane *et al.* 1985; Burggraf *et al.* 1991; Marchesi *et al.* 1998; Yoon *et al.* 1998; Turner *et al.* 1999). The first PCR step targeted the full 16S rDNA sequence (1380 bp), which was amplified with 1389R-9F primers. Then, for the second PCR step, 907R and 23F primers allowed the specific reamplification of the hypervariable region V1-V2-V3-V4-V5 (884 bp). Finally, 519R-63F primers were used to specifically reamplify the rDNA hypervariable region V2-V3 (456 bp). The double-nested PCR was performed in 12.5 μL of reaction mixture volume with DNA thermal cycler (Biometra®).

Table 1 PCR primers used for the classical PCR, the nested PCR and the double-nested PCR

Primer	Sequence	Reference
1389R	5′-ACGGGCGGTGTGTACAAG-3′	Marchesi *et al.* 1998
907R	5′-CCGTCAATTCCTTTRAGTTT-3′	Lane *et al.* 1985
519R	5′-GWATTACCGCGGCKGCTG-3′	Turner *et al.* 1999
9F	5′-GAGTTTGATCCTGGCTCAG-3′	Yoon *et al.* 1998
23F	5′-TGCAGAYCTGGTYGATYCT GCC-3′	Burggraf *et al.* 1991
63F	5′-CAGGCCTAACACATGCAA GTC-3′	Marchesi *et al.* 1998

Two other experiments were performed to compare this method with classical (one-step) and single-nested (two-step) PCR. The second experiment differed from the first in that the third step of the double-nested PCR was performed alone, starting directly from extracted DNA and using the same PCR conditions. For the third experiment, a two-step amplification nested PCR was performed. This two-step amplification was equivalent to the first and the third step of the double-nested PCR, and was carried out in identical PCR conditions.

The details of these three approaches are as follows. For the first experiment, the first step involved a reaction mixture consisting of 6.25 μL of Taq Polymerase (TaKaRa Ex Taq), 0.25 μL (1.25 μM) of each specific primers (forward 9F and reverse 1389R), 3.75 μL of sterile nuclease-free water (DEPC-treated Water Ambion®) and 2 μL of specify amount of DNA. After denaturation (95 °C; 2 min), 45 cycles were performed: denaturation (54 °C; 1 min), annealing (55 °C; 1 min) and extension (72 °C; 1 min) followed with a final extension step consisting of 5 min at 72 °C. For the second step, the PCR mixture consisted of 6.25 μL of Taq polymerase, 0.25 μL for each primer (907R-23F), 5.25 μL of sterile nuclease-free water and 0.5 μL of the step one reaction. The same thermal regime was performed as the first amplification step. For the third and final PCR steps, the reaction mixture contained 6.25 μL of Taq polymerase, 1.25 μL (6.25 μM) for each primer (519R-63F), 3.25 μL of RNA-DNA-bacteria-free water and 0.5 μL of product of PCR 2. The annealing step was changed to 59 °C, 1 min for the third amplification step. To eliminate contamination, the DEPC-treated water was stored at −20 °C in sterile, individually packed Eppendorf tubes. We performed the three PCR experiments for each sample in triplicate, each sample exhibiting a band of 500 bp for at least two of three times was deemed an infected fish. Bacterial cultures in liquid medium were used as a positive control.

Library. The third step of the double-nested PCR was performed a second time for individuals that yielded a visible band on agarose gel (infected individuals). For this library production, third-step primers 519R and 63F included, respectively, the 454 Life Sciences Adapter B and A. Then, 45 different bar-coded MID-tags (Multiplex identifiers) were added to allow parallel sample sequencing. These MIDs were linked with the 5′ end of the amplicon sequence specific 63F primer. A total of 133 samples were successfully amplified with the double-nested PCR, and labelled with MIDs and 454 primers.

Purification and pyrosequencing. PCR products were purified using the AMPure© bead calibration method for the preparation of DNA libraries in the GS FLX Titanium sequencing system. Then, all purified PCR products were quantified with a Nanodrop® spectrophotometer. Each sequence tagged library was produced by pooling 45 PCR products in equimolar amounts. Then, each group was sequenced using the GS-20 (Genome Sequencer 20) (Roche, Basel, Switzerland) at the Plateforme d'Analyses Génomiques (Université Laval, Québec, Canada). The emulsion PCR was performed according to the emPCR Method Manual (Roche).

Data analysis

The data were analysed in two steps. First, CLC Genomics Workbench 3.1 (CLC Bio, Aarhus, Denmark CLC work bench BIO®) was used to trim sequences for quality and recover the primers sequences and tags. Second, preprocessing and analysis were performed using the MOTHUR software (Schloss *et al.* 2009). Among the three analysis options available in MOTHUR, Operational Taxonomic Unit (OTU) was used. This approach was most appropriate for our data set because it is not dependent on a predefined taxonomy. Next, the Costello stool analysis protocol was used for cleaning, classified the sequences in OTU and analyse the diversity between samples (Costello *et al.* 2009). The first step of this protocol is to merge redundant sequences to get a list of unique sequences. Next, all the unique sequences were aligned to a bacterial reference with SILVA. In the third step, all sequences smaller than 300 bp were deleted to ensure an accurate and precise taxonomic identification by covering at least V1 and V2 hypervariable regions of the 16S gene (Chakravorty *et al.* 2007). Then, the precluster command clustered sequences with only one nucleotide different into the same OTU. Finally, the chimeras were detected using the SILVA-based alignment of the Gold reference set and then deleted.

Two indices were retained to assess the quality of pyrosequencing: the sequence coverage index and the Chao index. The sequence coverage index is a metric used to estimate the quality of a data set. This coverage is the average proportion of the bacterial community that is sequenced. The Chao index is a nonparametric richness estimator of the theoretical number of sequences, for which accuracy has been demonstrated in previous studies (Hughes *et al.* 2001). Furthermore, to analyse phylogenetic distances between each community, we calculate the weighted Unifrac distance (Lozupone & Knight 2005). This index calculates a distance between each pair of samples based on the abundance of taxa contained by the samples. We choose the weighted Unifrac distance because it takes account to the abundance and not only the presence or absence of taxa in the samples. UniFrac distance values range from 0 to 1. Zero indicates that the communities under comparison exhibit 100% similarity

in their phylogenetic structure while '1' indicates that the communities are highly differentiated.

Results

Comparison of PCR, nested PCR and double-nested PCR

The three amplification strategies, double-nested PCR, one-step PCR and nested PCR, were performed in duplicates for the same samples. The double-nested PCR (Fig. 1c.) obtained one single intense light band at 500 bp for the samples which were considered infected (1 and 2). Conversely, samples considered as uninfected did not exhibit any bands at all (3 and 4). This first PCR experiment showed relatively high repeatability across triplicates (86.3%). In the second experiment, which involved a one-step PCR, products were also obtained only from infected fish. However, both infected samples exhibited only a very faint positive band at 500 base pairs (bp) (Fig. 1a), and this could not be improved by varying the annealing temperature, the cycle number, the PCR-mix reaction composition or the Taq polymerase. In the third experiment, which involved two amplification steps (single-nested PCR), infected samples displayed two bands around 300 and 200 bp (Fig. 1b). Thus, the bacterial DNA represented by a 500-bp band was absent for these two putative infected samples. Moreover, the single-nested PCR allowed the amplification of bacterial DNA with a weak repeatability of 40% for 10 tested samples (data not shown), thus suggesting potential eukaryotic contamination and nonspecific bacterial genome amplification.

Description of data set

More than 700 000 reads for the entire data set of 133 infected individuals were obtained (Table 2). The number of sequencing reads did not differ significantly among lakes (*P*-value = 1, Chi-squared test). The 16S amplicon sequence length distributions before the preprocessing of the data set is presented in Fig. 2. This distribution was similar among four lakes: Cliff, East, Témiscouata and Webster. There was a first peak at 25–50 bases followed by a rising curve until a second peak at 350–375 bases. This second peak was followed by the main peak at 425–450 bases. The distribution for Indian

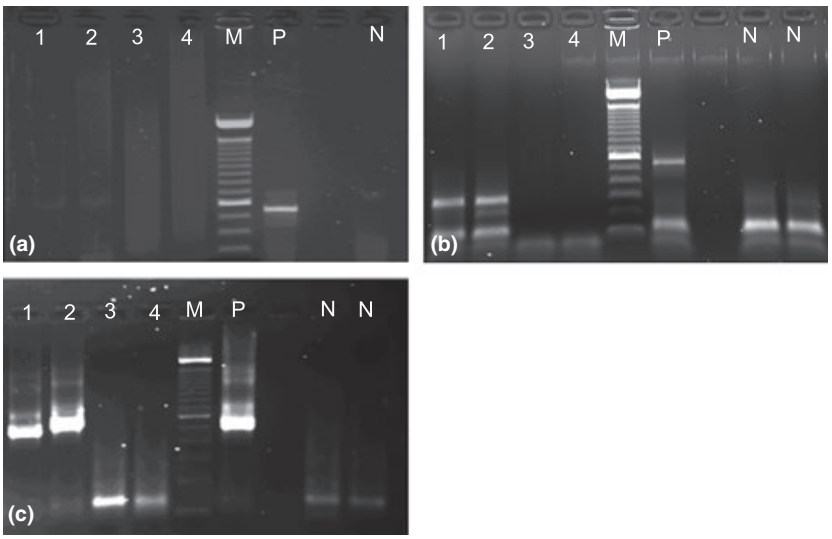

Fig. 1 (a) Results of classical PCR (519R-63F) with infected (1,2) and noninfected (3,4) samples (b) Results of nested PCR (907R-23F & 519R-63F) with infected (1,2) and noninfected (3,4) samples (c) Results of double-nested PCR (1389R-9F, 907R-23-F and 519R-63F) with samples 1,2,3 and 4. All PCR product migrations were carried out on agarose gels, using a 100-bp sharp DNA marker (Life Technologies) as ladder (M). The well P is the positive control, and the well N is the negative control.

Table 2 Read number analysis. The total of initial reads is the number of reads before the data set cleaning. This total initial read number was divided by the number of samples for each lake resulting in the mean initial reads per sample. The final number of sequences is the amount of sequences after the data set cleaning. This final number of sequences was divided by the number of samples to obtain the mean of final sequences per sample

Lakes	Cliff	East	Indian	Témiscouata	Webster	Total
Number of samples	27	27	37	18	24	133
Total of initial reads	135 978	168 924	163 124	122 041	104 848	694 915
Mean of initial reads per sample	5036	7345	4798	6780	4993	5650
Final number of sequences	11 687	18 461	18 766	19 127	11 105	79 146
Mean of final sequences per sample	433	803	552	1063	529	643

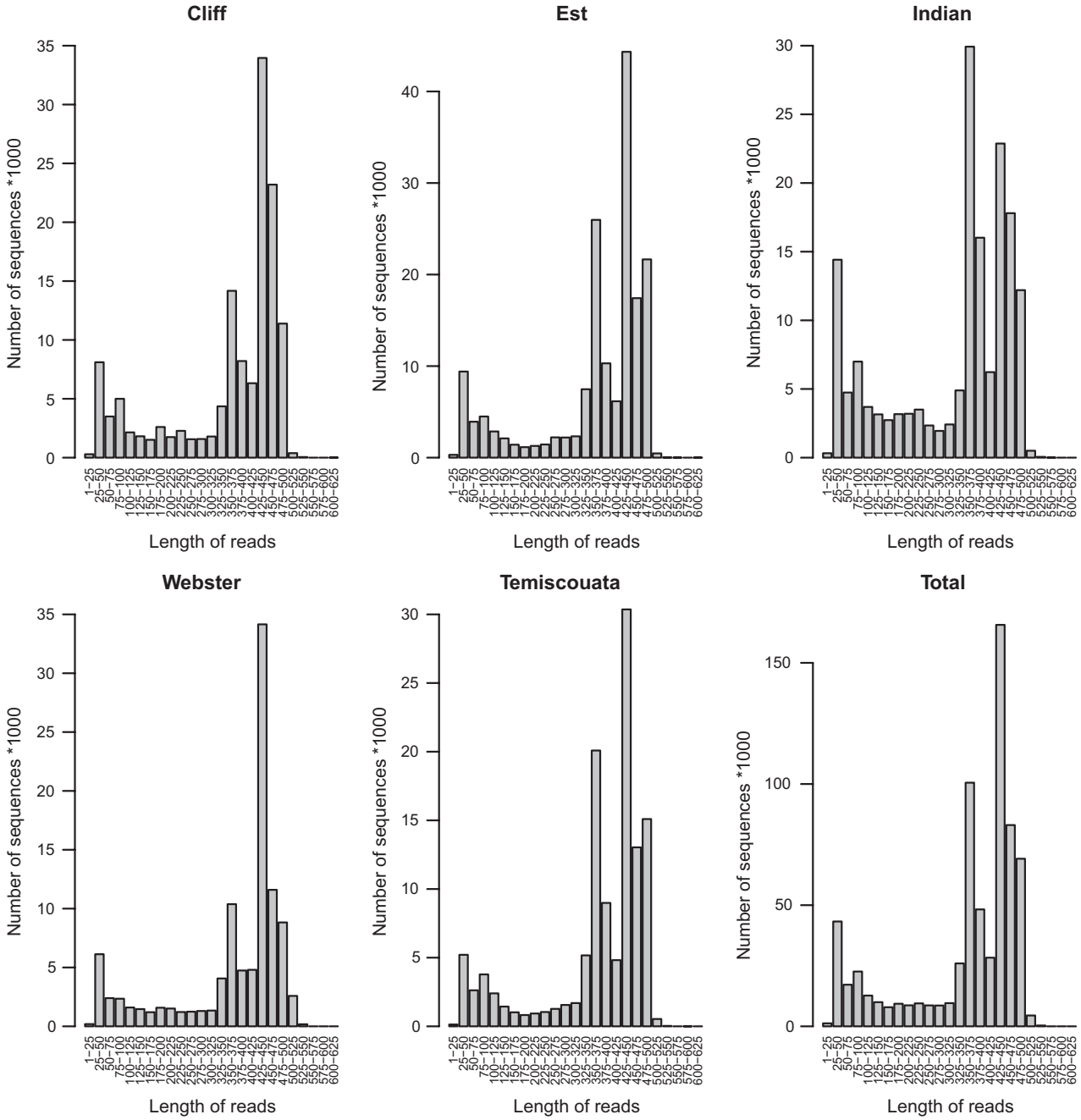

Fig. 2 Sequence length distribution. This distribution was computed using all 133 infected fish.

Pond differed from the other lakes. There was one main peak at 350–375, followed by three secondary peaks: 425–450, 450–475 and 475–500 bases. For all five lakes, sequences longer than 300 bases composed 77% of the data set. The five different bar-coded MID-tags were mostly reliable as only two did not work in all three libraries (MID 20 and 21). Thus, the six samples labelled with these MID-tags were not sequenced. Aside of this, four other samples tagged with different MID, were either chimeras or sequences shorter than 50 bases pairs, which were removed from the data set.

Data cleaning

Data cleaning steps were undertaken first with CLC Genomics, and then with the preprocessing algorithms of MOTHUR. First, 170 169 redundant sequences were deleted (Table 3). After this sorting step, 115 203 reads

Table 3 Data set cleaning steps. The first step is a command named 'unique' filtering out all redundant sequences. This set of unique sequences was then aligned to a bacterial reference. The 'screen' step deleted all sequences smaller than 300 bp. The third step is another 'unique' command used to remove all redundant sequences after the alignment. The 'pre-cluster' 'chimera', and 'remove' steps allowed eliminating sequences with a pyrosequencing error and the chimeras, respectively

Step	Cliff	East	Indian	Témiscouata	Webster	Total	Percentage
Number of reads	135 978	168 924	163 124	122 041	104 848	694 915	100
Unique and alignment	101 952	128 555	120 899	95 378	77 962	524 746	75.51
Screen	79 774	104 469	89 638	77 858	57 804	409 543	58.93
Unique	31 749	48 811	48 706	42 271	28 235	199 772	28.75
Pre-cluster	19 890	32 866	32 715	29 299	18 888	133 658	19.23
Chimera and remove	11 687	18 461	18 766	19 127	11 105	79 146	11.39

smaller than 300 bp (16.5%) and 209 771 redundant aligned sequences were removed. The precluster command allowed the elimination of 66 114 reads with a pyrosequencing error. Finally, 54 512 chimeric (7.8%) reads were deleted. Finally, 525 200 sequences (75.6% of the all data set) of excellent quality were used for further analyses. From those sequences, we used 79 146 unique sequences (11% of the data set) to maximize the calculation time. The abundance of each sequence (number of count) was set aside to keep all the information necessary for further quantitative analyses.

Analysis

Four of the lakes presented the same sequence coverage at 96%, whereas Indian Pond had coverage of 93%, for a mean of 95.4%. A subset of 10 individuals from the complete data set of 133 was randomly chosen to provide a detailed coverage analysis (Table 4). (Given the methodological nature of this paper, the detailed analysis and interpretations of the entire data set will be the topic of a full regular paper). The Chao curve compares the Chao index vs. the numbers of sequences observed (Fig. 3). Three of five lakes reached a plateau (East, Indian and Témiscouata) whereas the two other lakes were still in the exponential phase. Moreover, there was a large variation of bacterial family diversity among the ten samples (Table 4) and each bacterial community are phylogenetically differentiated (Mean Unifrac distances = 0.42, $P < 0.001$).

Discussion

To ensure a thorough and accurate diagnosis of bacterial infection in fish kidney, we developed a three-step PCR approach to diagnose individuals that were infected; then reperformed the third PCR with MID labelled primers to construct 454 library of 16S amplicons. To reach the goal, three different PCR protocols were tested in parallel. The classical PCR with one set of primers failed to discriminate efficiently between infected and healthy fish because its amplification rate was weak even after 45 cycles (Fig. 1a). This method was not conclusive for either the infection diagnosis or for amplicon library construction. Total DNA concentrations of each sample were very high, so it is unlikely this parameter was truly responsible for the weakness of this single-step amplification strategy. Rather, the pervasive influence of the host eukaryotic DNA compared with the small amount of bacterial DNA caused the more likely weak, nonspecific amplifications. The nested PCR approach was thus chosen to circumvent this problem. Results of the two-step PCR protocol (Fig. 1b) were consistent with eukaryotic contamination, potentially owing to cross amplification of host DNA. Although of good quality, amplicons were not specific to bacterial 16S ribosomal genes as evidenced by a 300-bp band obtained in 73% cases. In other experiments, we observed that those primers sets anneal partially with the 18S rDNA gene of other salmonids (*Salvelinus fontinalis)* (S. Boutin, personal communication) confirming that the 300-bp band was probably a product of the amplification of 18S rDNA gene of whitefish. Furthermore, repeatability of the 500-bp band amplification was poor (40%), and did not allow accurate diagnostic of bacterial presence.

In contrast, the new protocol involving a third-step amplification with a new set of primers increased both the specificity and amount of the targeted bacterial amplicon, thus improving the experimental repeatability. Thus, the resulting double-nested PCR provided consistent results (Fig. 1c): a unique, well defined 500-bp band was obtained with 86.3% repeatability, indicating a highly specific amplification of bacterial DNA. Overall, the double-nested PCR was the only amplification strategy that provided reliable qualitative diagnosis between infected and uninfected fish. Moreover, this three-step nested PCR strategy was the only one to generate a sufficient yield of 16S DNA to construct an amplicon library suitable for pyrosequencing.

To assess further the specificity and the repeatability of the three steps nested PCR, the resulting 16S amplicon

libraries were synthesized and pyrosequenced using 454 FLX. In a first step, read quality, number and length were measured. The 16S amplicon library sequencing was very efficient as it generated 700 000 reads for a 3/4 plate, which exceeds the expected mean number of reads for the FLX chemistry by 17% (200 000 for a 1/4 plate, Brian Boyle, personal communication) (Droege & Hill 2008). Furthermore, no significant differences were observed for the read number between individual samples. This indicates that all samples were sequenced equally. This is remarkable because in other studies read are not fairly divided between tags (Stoeck *et al.* 2009). The only problem encountered was the noneffectiveness of MIDs 20 and 21, but those two tags were noneffective in all three quarters of the plate, so the problem came from the MID sequences in these primers. It may be a problem in detecting the key sequence in the primers, or in the quality filters used, or a secondary structure formation with those two MIDs.

The majority of reads were usable for phylogenic assignation as the proportion of short reads (<300 bp) did not exceed 23% before preprocessing. All reads were cleaned using Mothur. Chimera formation is a common source of artefacts during PCR amplification. Haas *et al.* (2011) showed that the average percentage of chimera formed with amplification of the V1–V3 region is 16%, ranging from 15% to over 20% (Haas *et al.* 2011). Here, we detected chimeras using Chimera slayer implemented in the software MOTHUR, and only 7.84% of chimeras were rejected using the same algorithm. The low number of chimera recorded in our study suggests that the nested PCR is a safer methodology to increases the DNA yield and reduces the formation of chimera than increasing the number of cycles of a single-step PCR (Takahiro 2003).

The second step evaluated the representativeness of sequencing of each sample and to check for potential bias induced by PCR. As there was either not enough material or the wrong fragment size, we did not perform pyrosequencing on the single and nested PCRs and we cannot compare the performance of these methods with respect to pyrosequencing. But the coverage was calculated using the Good's estimator index and Chao's curves indicate whether all samples are deeply sequenced and accurately representative. The coverage of each of the ten samples was between 90% and 99%. Finally, all sequences were classified in different OTU by the assignation in the Ribosomal Database Project database (Cole *et al.* 2005) (Table 4). All samples were composed by a large diversity of families, and unifrac metrics between each sample was important and statistically significant indicating that this method did not induced any bias in favour of one family. Indeed, if any bias was induce by the double-nested PCR, owing to the choice of primers, we would expect to observe more similar patterns of abundance in the bacterial communities by favouring certain genera. To the contrary, each sample is mostly represented by different genera. Therefore, the resulting 16S amplicon library was representative of the sample taxonomic diversity and do not exhibit any bias in favour of one or more genera.

To conclude, the double-nested PCR developed in this study is a dedicated protocol to diagnose thoroughly and accurately the taxonomic diversity of fish immunobiomes, even in samples with relatively low ratio of bacteria/host eukaryotic DNA. The pyrosequencing of the 16S amplicon library provided a reliable data set, featuring a deep coverage of sample diversity and, importantly, with low chimera occurrences.

Table 4 Taxonomy of the ten randomly chosen samples

Lake of sample	Phylum	Class	Order	Familly	Number of sequences
East	Actinobacteria	Actinomycetales	Dermatophilaceae	Kineosphaera	1
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	1195
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	5
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	61
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	937
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Macromonas	1
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Delftia	3
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	1189
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Alkanindiges	9
Webster	Clostridia	Clostridiales	Incertae_Sedis_XI	Finegoldia	4744
	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Fodinicurvata	1
	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Acidocella	1

Table 4 (Continued)

Lake of sample	Phylum	Class	Order	Familly	Number of sequences
East	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	4144
	Planctomycetacia	Planctomycetales	Planctomycetaceae	Gemmata	721
	Actinobacteria	Actinomycetales	Micrococcaceae	Kocuria	468
	Planctomycetacia	Planctomycetales	Planctomycetaceae	Zavarzinella	103
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Alkanindiges	5
	Actinobacteria	Actinomycetales	Microbacteriaceae	Microterricola	4
	Actinobacteria	Actinomycetales	Microbacteriaceae	Microcella	4
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Polynucleobacter	3
	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	3
	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sandaracinobacter	2
	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Asticcacaulis	2
	Planctomycetacia	Planctomycetales	Planctomycetaceae		2
	Actinobacteria	Actinomycetales	Micrococcaceae	Renibacterium	2
	Thermodesulfobacteria	Thermodesulfobacteriales	Thermodesulfobacteriaceae	Thermodesulfatator	1
	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	1
	Betaproteobacteria	Neisseriales	Neisseriaceae	Laribacter	1
	Betaproteobacteria	Burkholderiales	Burkholderiales	Thiobacter	1
Indian	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Stenotrophomonas	1212
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	631
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Variovorax	149
	Gammaproteobacteria	Alteromonadales	Shewanellaceae	Shewanella	148
	Sphingobacteria	Sphingobacteriales	Cytophagaceae	Adhaeribacter	142
	Acidobacteria_Gp6	unclassified	unclassified	unclassified	92
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Pelomonas	61
	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Prosthecomicrobium	59
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Psychrobacter	44
	Actinobacteria	Actinomycetales	Micrococcaceae	Kocuria	28
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Flavobacterium	23
	Gemmatimonadetes	Gemmatimonadales	Gemmatimonadaceae	Gemmatimonas	21
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Agromonas	19
	Fusobacteria	Fusobacteriales	Fusobacteriaceae	Cetobacterium	18
	Fusobacteria	Fusobacteriales	Fusobacteriaceae	Psychrilyobacter	17
	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Arenimonas	16
	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Pseudomonas	16
	Gammaproteobacteria	Alteromonadales	Ferrimonadaceae	Paraferrimonas	16
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	16
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Afipia	16
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae		12
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Afipia	9
	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Phenylobacterium	8
	Fusobacteria	Fusobacteriales	Fusobacteriaceae		8
	Actinobacteria	Actinomycetales	Micrococcaceae	Arthrobacter	8
	Betaproteobacteria	Burkholderiales	Alcaligenaceae	Derxia	7
	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Thiorhodospira	5
	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Thermomonas	4
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	4
	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Rahnella	4
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Afipia	4
	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Aspromonas	3
	Betaproteobacteria	Burkholderiales	Oxalobacteraceae	Oxalicibacterium	3
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Burkholderia	3
	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Prosthecomicrobium	3
	Alphaproteobacteria	Rhizobiales	Brucellaceae	Daeguia	3
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Chryseobacterium	3
	Actinobacteria	Acidimicrobiales	Iamiaceae	Iamia	3
	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Luteimonas	2

Table 4 (Continued)

Lake of sample	Phylum	Class	Order	Familly	Number of sequences
	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae		2
	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Leminorella	2
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Caenimonas	2
	Betaproteobacteria	Burkholderiales	Comamonadaceae		2
	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Tistrella	2
	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Caenispirillum	2
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	2
	Actinobacteria	Actinomycetales	Intrasporangiaceae	Phycicoccus	2
	Gammaproteobacteria	Xanthomonadales	Xanthomonadaceae	Aquimonas	1
	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Azomonas	1
	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	Pseudospirillum	1
	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Modicisalibacter	1
	Gammaproteobacteria	Methylococcales	Methylococcaceae	Methylosphaera	1
	Gammaproteobacteria	Methylococcales	Methylococcaceae	Methylococcus	1
	Gammaproteobacteria	Legionellales	Legionellaceae	Legionella	1
	Gammaproteobacteria	Gammaproteobacteria	Gammaproteobacteria	Simiduia	1
	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Pragia	1
	Gammaproteobacteria	Alteromonadales	Moritellaceae	Paramoritella	1
	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Ferribacterium	1
	Betaproteobacteria	Rhodocyclales	Rhodocyclaceae	Denitratisoma	1
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Schlegelella	1
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Pseudacidovorax	1
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Ottowia	1
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Comamonas	1
	Betaproteobacteria	Burkholderiales	Burkholderiales	Mitsuaria	1
	Betaproteobacteria	Burkholderiales	Burkholderiales	Aquincola	1
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Wautersia	1
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Ralstonia	1
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Polynucleobacter	1
	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Zymomonas	1
	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	1
	Alphaproteobacteria	Sphingomonadales	Erythrobacteraceae	Croceicoccus	1
	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Albidovulum	1
	Alphaproteobacteria	Rhizobiales	Xanthobacteraceae	Starkeya	1
	Alphaproteobacteria	Rhizobiales	Rhizobiaceae	Sinorhizobium	1
	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Microvirga	1
	Alphaproteobacteria	Rhizobiales	Hyphomicrobiaceae	Filomicrobium	1
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodopseudomonas	1
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Rhodoblastus	1
	Alphaproteobacteria	Rhizobiales	Beijerinckiaceae	Methylovirgula	1
	Alphaproteobacteria	Alphaproteobacteria	Alphaproteobacteria	Elioraea	1
	Fusobacteria	Fusobacteriales	Fusobacteriaceae	Propionigenium	1
	Bacilli	Bacillales	Staphylococcaceae	Staphylococcus	1
	Sphingobacteria	Sphingobacteriales	Chitinophagaceae	Filimonas	1
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Mesoflavibacter	1
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Microlunatus	1
	Actinobacteria	Actinomycetales	Microbacteriaceae	Glaciibacter	1
	Actinobacteria	Actinomycetales	Cellulomonadaceae	Paraoerskovia	1
	Actinobacteria	Actinobacteria	Acidimicrobidae	Ilumatobacter	1
Témiscouata	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Escherichia/Shigella	1496
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter	656
	Actinobacteria	Actinomycetales	Corynebacteriaceae	Corynebacterium	151
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	140
	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	80
	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Salmonella	4
	Gammaproteobacteria	Oceanospirillales	Litoricolaceae	Litoricola	1

Table 4 (Continued)

Lake of sample	Phylum	Class	Order	Familly	Number of sequences
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Pelomonas	1
	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Roseicyclus	1
	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Pannonibacter	1
Indian	Anaerolineae	Anaerolineales	Anaerolineaceae	Levilinea	23
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter	20
	Alphaproteobacteria	Caulobacterales	Caulobacteraceae	Brevundimonas	12
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Cloacibacterium	9
	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Paracoccus	4
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	3
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Pelomonas	3
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Ralstonia	3
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	3
	Actinobacteria	Actinomycetales	Micrococcaceae	Kocuria	3
	Clostridia	Thermoanaerobacterales	Thermoanaerobacteraceae	Gelria	2
	Actinobacteria	Actinomycetales	Microbacteriaceae		2
	Gammaproteobacteria	Pseudomonadales	Pseudomonadaceae	Azomonas	1
	Gammaproteobacteria	Enterobacteriales	Enterobacteriaceae	Enterobacter	1
	Betaproteobacteria	Burkholderiales	Burkholderiaceae	Polynucleobacter	1
	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Zymomonas	1
	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	1
	Alphaproteobacteria	Rhodospirillales	Acetobacteraceae	Acidocella	1
	Alphaproteobacteria	Rhodobacterales	Rhodobacteraceae	Catellibacterium	1
	Alphaproteobacteria	Rhizobiales	Phyllobacteriaceae	Aminobacter	1
	Clostridia	Clostridiales	Clostridiaceae	Sarcina	1
	Anaerolineae	Anaerolineales	Anaerolineaceae	Bellilinea	1
	Sphingobacteria	Sphingobacteriales	Flammeovirgaceae	Reichenbachiella	1
	Sphingobacteria	Sphingobacteriales	Chitinophagaceae	Terrimonas	1
East	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Bradyrhizobium	5064
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	1069
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Agromonas	14
	Alphaproteobacteria	Rhizobiales	Methylobacteriaceae	Microvirga	1
	Alphaproteobacteria	Rhizobiales	Bradyrhizobiaceae	Afipia	1
Cliff	Clostridia	Clostridiales	Clostridiaceae	Clostridium	1488
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Cloacibacterium	715
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	92
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	51
	Clostridia	Clostridiales	Clostridiaceae	Sarcina	48
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	36
	Clostridia	Clostridiales	Clostridiaceae	Sarcina	34
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	34
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	33
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	32
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	18
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	17
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter	11
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	11
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	10
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	10
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	10
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	10
	Clostridia	Clostridiales	Clostridiaceae	Anaerobacter	10
	Clostridia	Clostridiales	Clostridiaceae		10
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	9
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Riemerella	8
	Clostridia	Clostridiales	Clostridiaceae	Clostridium	6
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Cloacibacterium	6

Table 4 (Continued)

Lake of sample	Phylum	Class	Order	Familly	Number of sequences
	Clostridia	Clostridiales	Clostridiaceae	Sarcina	4
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae		3
	Clostridia	Clostridiales	Clostridiaceae	Sarcina	3
	Bacilli	Bacillales	Bacillaceae	Salirhabdus	3
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Chryseobacterium	3
	Actinobacteria	Actinomycetales	Microbacteriaceae	Microbacterium	3
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Alkanindiges	2
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Comamonas	2
	Clostridia	Clostridiales	Clostridiaceae	Geosporobacter	2
	Bacilli	Bacillales	Bacillaceae		2
	Flavobacteria	Flavobacteriales	Flavobacteriaceae	Kaistella	1
	Actinobacteria	Actinomycetales	Propionibacteriaceae	Propionibacterium	1
East	Betaproteobacteria	Burkholderiales	Comamonadaceae	Pelomonas	2550
	Gammaproteobacteria	Chromatiales	Chromatiaceae	Thiohalocapsa	357
	Betaproteobacteria	Burkholderiales	Burkholderiales	Mitsuaria	13
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Tepidicella	8
	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Telmatospirillum	6
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Verminephrobacter	4
	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Rhodovibrio	4
	Gammaproteobacteria	Chromatiales	Chromatiaceae	Thiohalocapsa	3
	Gammaproteobacteria	Chromatiales	Chromatiaceae	Thiophaeococcus	3
	Betaproteobacteria	Burkholderiales	Comamonadaceae	Verminephrobacter	2
	Alphaproteobacteria	Sphingomonadales	Sphingomonadaceae	Sphingomonas	2
	Alphaproteobacteria	Alphaproteobacteria	Alphaproteobacteria	Elioraea	2
	Gammaproteobacteria	Gammaproteobacteria	Gammaproteobacteria	Simiduia	1
	Gammaproteobacteria	Chromatiales	Ectothiorhodospiraceae	Natronocella	1
	Gammaproteobacteria	Acidithiobacillales	Thermithiobacillaceae	Thermithiobacillus	1
	Betaproteobacteria	Neisseriales	Neisseriaceae	Uruburuella	1
	Alphaproteobacteria	Rhodospirillales	Rhodospirillaceae	Tistrella	1
	Clostridia	Clostridiales	Ruminococcaceae	Butyricicoccus	1
	Clostridia	Clostridiales	Ruminococcaceae	Anaerotruncus	1
	Clostridia	Clostridiales	Clostridiaceae	Caminicella	1
Cliff	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Psychrobacter	1262
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Acinetobacter	603
	Clostridia	Clostridiales	Ruminococcaceae	Ruminococcus	306
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Psychrobacter	197
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Alkanindiges	22
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Enhydrobacter	14
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Alkanindiges	10
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae		9
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Alkanindiges	4
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae	Psychrobacter	4
	Gammaproteobacteria	Pseudomonadales	Moraxellaceae		3
	Bacilli	Lactobacillales	Enterococcaceae	Pilibacter	3
	Gammaproteobacteria				2
	Gammaproteobacteria	Oceanospirillales	Oceanospirillaceae	Bermanella	1
	Gammaproteobacteria	Oceanospirillales	Halomonadaceae	Cobetia	1
	Clostridia	Clostridiales	Ruminococcaceae	Subdoligranulum	1
	Clostridia	Clostridiales	Ruminococcaceae	Faecalibacterium	1
	Clostridia	Clostridiales	Lachnospiraceae	Anaerosporobacter	1
	Bacilli	Lactobacillales	Enterococcaceae	Catellicoccus	1
	Bacilli	Bacillales	Bacillaceae	Thalassobacillus	1

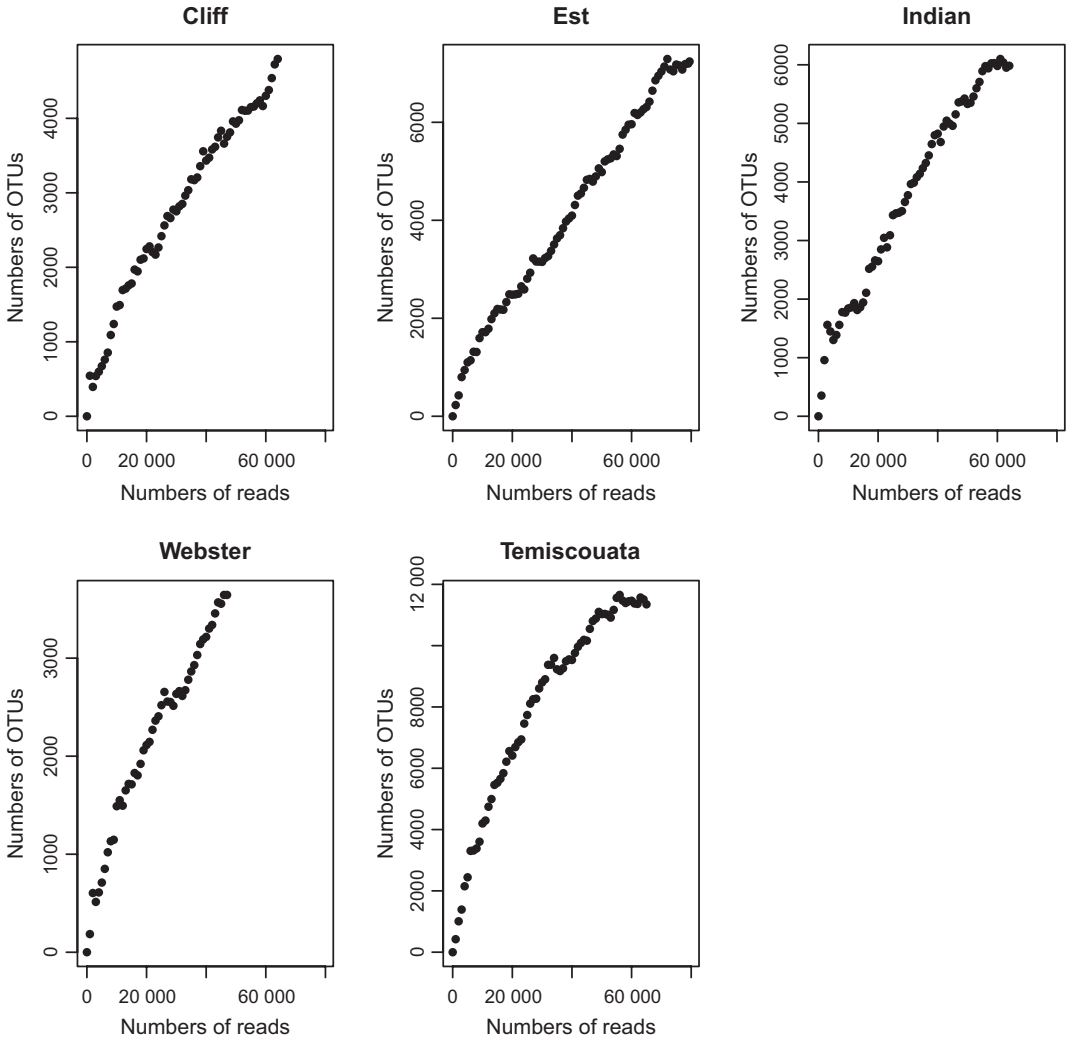

Fig. 3 Chao curve: number of obtained sequences according to the theoretical sequence numbers. This Chao curve was calculated with the totality of the sample.

Acknowledgements

We thank Brian Boyle for his help with the 454 sequencing and Eric Normandeau for his bioinformatic skills and support. This work was supported by a National Sciences and Engineering Council of Canada (NSERC) Discovery grant and Canadian Research Chair to LB. It is also a contribution to the research programmes of Québec-Océan and RAQ. SB PhD project was financially supported by a Partnership grant (Strategic programme) of the National Sciences and Engineering Council of Canada (NSERC) to LB and ND, and the 'Collaborative Research and Training Experience Program' (CREATE).

References

Amann R, Ludwig W (2000) Ribosomal RNA-targeted nucleic acid probes for studies in microbial ecology. *FEMS Microbiology Reviews*, **24**, 555–565.

Amend AS, Seifert KA, Bruns TD (2010) Quantifying microbial communities with 454 pyrosequencing: does read abundance count? *Molecular Ecology*, **19**, 5555–5565.

Ampe F, Miambi E (2000) Cluster analysis, richness and biodiversity indexes derived from denaturing gradient gel electrophoresis fingerprints of bacterial communities demonstrate that traditional maize fermentations are driven by the transformation process. *International Journal of Food Microbiology*, **60**, 91–97.

Ampe F, ben Omar N, Moizan C *et al.* (1999) Polyphasic study of the spatial distribution of microorganisms in mexican pozol, a fermented maize dough, demonstrates the need for cultivation-independent methods to investigate traditional fermentations. *Applied and Environmental Microbiology*, **65**, 5464–5473.

Bottger EC (1990) Frequent contamination of Taq polymerase with DNA. *Clinical Chemistry*, **36**, 1258–1259.

Burggraf S, Stetter KO, Rouviere P *et al.* (1991) *Methanopyrus kandleri*: an archaeal methanogen unrelated to all other known methanogens. *Systematic and Applied Microbiology*, **14**, 346–351.

Cahill MM (1990) Bacterial flora of fishes: a review. *Microbial Ecology*, **19**, 21–41.

Chakravorty S, Helb D, Burday M *et al.* (2007) A detailed analysis of 16s ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *Journal of Microbiological Methods*, **69**, 330–339.

Cole J, Chai B, Farris R *et al.* (2005) The ribosomal database project (RDP-II): sequences and tools for high-throughput rRNA analysis. *Nucleic Acids Research*, **33**, D294–D296.

Costello EK, Lauber CL, Hamady M *et al.* (2009) Bacterial community variation in human body habitats across space and time. *Science*, **326**, 1694–1697.

Dionne M, Miller KM, Dodson JJ *et al.* (2009) MHC standing genetic variation and pathogen resistance in wild Atlantic salmon. *Philosophical Transactions of the Royal Society B-Biological Sciences*, **364**, 1555–1565.

Droege M, Hill B (2008) The genome sequencer FLX™ system—longer reads, more applications, straight forward bioinformatics and more complete data sets. *Journal of Biotechnology*, **136**, 3–10.

Engelbrektson A, Kunin V, Wrighton KC *et al.* (2010) Experimental factors affecting PCR-based estimates of microbial species richness and evenness. *The ISME Journal*, **4**, 642–647.

Haas BJ, Gevers D, Earl AM *et al.* (2011) Chimeric 16s rRNA sequence formation and detection in sanger and 454-pyrosequenced PCR amplicons. *Genome Research*, **21**, 494–504.

Hughes JB, Hellmann JJ, Ricketts TH *et al.* (2001) Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology*, **67**, 4399–4406.

Humblot C, Guyot J-P (2009) Pyrosequencing of tagged 16S rRNA gene amplicons for rapid deciphering of the microbiomes of fermented foods such as pearl millet slurries. *Applied and Environmental Microbiology*, **75**, 4354–4361.

Kulakov LA, McAlister MB, Ogden KL *et al.* (2002) Analysis of bacteria contaminating ultrapure water in industrial systems. *Applied and Environment Microbiology*, **68**, 1548–1555.

Lane DJ, Pace B, Olsen GJ *et al.* (1985) Rapid determination of 16s ribosomal RNA sequences for phylogenetic analyses. *Proceedings of the National Academy of Sciences of the United States of America*, **82**, 6955–6959.

Larsen MH, Blackburn N, Larsen JL *et al.* (2004) Influences of temperature, salinity and starvation on the motility and chemotactic response of *Vibrio anguillarum*. *Microbiology*, **150**, 1283–1290.

Lozupone C, Knight R (2005) Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, **71**, 8228–8235.

Marchesi JR, Sato T, Weightman AJ *et al.* (1998) Design and evaluation of useful bacterium-specific PCR primers that amplify genes coding for bacterial 16s rRNA. *Applied and Environmental Microbiology*, **64**, 795–799.

Petrosino JF, Highlander S, Luna RA *et al.* (2009) Metagenomic pyrosequencing and microbial identification. *Clinical Chemistry*, **55**, 856–866.

Rand KH, Houck H (1990) Taq polymerase contains bacterial DNA of unknown origin. *Molecular and Cellular Probes*, **4**, 445–450.

Schloss P, Westcott S, Ryabin T *et al.* (2009) Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology*, **75**, 7537–7541.

Schmidt TM, Pace B, Pace NR (1991) Detection of DNA contamination in Taq polymerase. *BioTechniques*, **11**, 176–177.

Stoeck T, Behnke A, Christen R *et al.* (2009) Massively parallel tag sequencing reveals the complexity of anaerobic marine protistan communities. *BMC Biology*, **7**, 72.

Takahiro K (2003) Bias and artifacts in multitemplate polymerase chain reactions (PCR). *Journal of Bioscience and Bioengineering*, **96**, 317–323.

Turner S, Pryer KM, Miao VPW *et al.* (1999) Investigating deep phylogenetic relationships among cyanobacteria and plastids by small subunit rRNA sequence analysis. *Journal of Eukaryotic Microbiology*, **46**, 327–338.

Uhland CF, Martineau DL, Mikaelian I, Eds. (2000). *Maladies des poissons d'eau douce du québec: Guide de diagnostic*. Les Presses de l'Université de Montréal, Montréal.

Von Wintzingerode F, Gobel U, Stackebrandt E (1997) Determination of microbial diversity in environmental samples: pitfalls of PCR-based rRNA analysis. *FEMS Microbiological Review*, **21**, 213–229.

Yoon J-H, Lee ST, Park Y-H (1998) Inter- and intraspecific phylogenetic analysis of the genus Nocardioides and related taxa based on 16S rDNA sequences. *International Journal of Systematic Bacteriology*, **48**, 187–194.

Yourno J (1992) A method for nested PCR with single closed reaction tubes. *PCR Methods and Applications*, **2**, 60–65.

Zhang H, Parameswaran P, Badalamenti J *et al.* (2011) Integrating high-throughput pyrosequencing and quantitative real-time PCR to analyze complex microbial communities. *Methods in Molecular Biology*, **733**, 107–128.

S.B. conceived the experimental design and the lab protocol. M.S. performed the laboratory work. M.S. and S.A.P helped to develop the experimental design and the lab protocol. S.B. and M.S. participated in the design of the analysis, performed research and analyzed the data. L.B. and N.D. conceived of the collaborative study and participated in the design of the analysis. S.B., M.S. and N.D. wrote the manuscript. S.B., M.S., S.A.P., L.B. and N.D. all provided comments on the manuscript. All authors read and approved the final manuscript.