Video Article

Next-generation Sequencing of 16S Ribosomal RNA Gene Amplicons

Sylvie Sanschagrin[1], Etienne Yergeau[1]

[1]Energy, Mining and Environment, National Research Council Canada

Correspondence to: Etienne Yergeau at etienne.yergeau@cnrc-nrc.gc.ca

Abstract

One of the major questions in microbial ecology is "who is there?" This question can be answered using various tools, but one of the long-lasting gold standards is to sequence 16S ribosomal RNA (rRNA) gene amplicons generated by domain-level PCR reactions amplifying from genomic DNA. Traditionally, this was performed by cloning and Sanger (capillary electrophoresis) sequencing of PCR amplicons. The advent of next-generation sequencing has tremendously simplified and increased the sequencing depth for 16S rRNA gene sequencing. The introduction of benchtop sequencers now allows small labs to perform their 16S rRNA sequencing in-house in a matter of days. Here, an approach for 16S rRNA gene amplicon sequencing using a benchtop next-generation sequencer is detailed. The environmental DNA is first amplified by PCR using primers that contain sequencing adapters and barcodes. They are then coupled to spherical particles via emulsion PCR. The particles are loaded on a disposable chip and the chip is inserted in the sequencing machine after which the sequencing is performed. The sequences are retrieved in fastq format, filtered and the barcodes are used to establish the sample membership of the reads. The filtered and binned reads are then further analyzed using publically available tools. An example analysis where the reads were classified with a taxonomy-finding algorithm within the software package Mothur is given. The method outlined here is simple, inexpensive and straightforward and should help smaller labs to take advantage from the ongoing genomic revolution.

Video Link

The video component of this article can be found at https://www.jove.com/video/51709/

Introduction

Metagenomic sequencing is a very powerful technology as it targets the entirety of the genetic information contained in an environmental sample. There are different flavors of metagenomic sequencing, including shotgun sequencing, large-insert libraries and amplicon sequencing. Amplicon sequencing offers the advantage of being relatively inexpensive, fast and able to produce reads from a single genomic region that can be generally aligned. In addition, the data analysis workflow for amplicon sequencing is mostly standardized. However, since it is based on PCR, it has all the biases related to incomplete specificity, incomplete coverage and primer biases[1,2], which makes this approach semi-quantitative at best. Several genomic regions can be targeted for amplicon sequencing including functional genes, but the most popular options are to use marker genes such as the 16S rRNA gene to generate a community profile. Traditionally, 16S rRNA gene amplicon sequencing was carried out using labor intensive techniques that included cloning in *E. coli*, colony picking and plasmid extraction followed by Sanger sequencing on the isolated plasmids, and, consequently, most studies analyzed fewer than 100 clones per sample. Next-generation sequencing brought two major advances: massive parallelization of the sequencing reactions and, most importantly, clonal separation of templates without the need to insert gene fragments in a host. This has simplified tremendously the sequencing of 16S rRNA gene amplicons, which is now back as a routine feature of many environmental microbiology studies, resulting in a "renaissance" for 16S rRNA gene amplicon sequencing [3].

Since the advent of Roche 454 sequencing in 2005[4], several other next-generation sequencing technologies have appeared on the market (*e.g.,* Illumina, Solid, PacBio). More recently, the introduction of bench-top sequencers brought to small labs the sequencing capacity once exclusive to large sequencing centers. Five benchtop machines are currently available: the 454 GS Junior, the Ion Torrent Personal Genome Machine (PGM) and Proton, and the Illumina MiSeq and NextSeq 500. While all these sequencers offer less reads per run and fewer bases per dollar than most full-scale sequencers, they are more flexible, rapid, and their low acquisition and run costs makes them affordable for small academic laboratories. Benchtop sequencers are particularly well suited for amplicon, small genome and low-complexity metagenome sequencing in environmental microbiology studies, because this type of studies generally does not require an extreme depth of sequencing. For example, it is generally agreed that for 16S rRNA gene sequencing studies the number of reads per sample is not paramount, as ~1,000 reads can generate the same patterns as multi-million reads datasets[5]. Having said that, benchtop next-generation sequencers still generate large amounts of sequence data, with maximal yields of ~35 Mbp (454 GS Junior), ~2 Gbp (Ion Torrent PGM), ~10-15 Gbp (Ion Torrent Proton), ~10 Gbp (Illumina MiSeq) and ~100 Gbp (Illumina Next Seq 500), which is more than enough for most environmental microbiology studies.

Next-generation sequencing of 16S rRNA amplicons using benchtop sequencers has been recently applied to a wide variety of environments. For example, the Ion Torrent PGM has been used for community analyses of uranium mine tailings that had particularly high pH and low permeability[6], of recirculating aquaculture systems[7], of hydrocarbon-contaminated Arctic soils[8,9], of oil sands mining affected sediments and

biofilms from the Athabasca River[10,11], of the rhizosphere of willows planted in contaminated soils[12], of the human and animal bodies[13-16] and of anaerobic digesters[17].

In this contribution we detail our approach to sequence 16S rRNA gene amplicons in-house using a benchtop next-generation sequencer (the Ion Torrent PGM). After DNA extraction, 16S rRNA genes are amplified using domain-level bacterial primers that contain sequencing adapters and unique, sample-specific sequences (barcodes). The amplicons are purified, quantified and pooled at an equimolar ratio. The pooled samples are then clonally amplified in an emulsion PCR and sequenced. Resulting sequences are analyzed using publicly available bioinformatics tools (*e.g.,* Mothur).

Protocol

1. 16S rRNA Gene Amplicon Library Preparation by the Fusion Method

1. Thaw the primers (Forward, F343-IonA-MIDXX: 5'-**CCA TCT CAT CCC TGC GTG TCT CCG AC***T CAG* XXX XXX XXX XTA CGG RAG GCA GCA G-3'; reverse, R533-IonP1: 5'-**CCT CTC TAT GGG CAG TCG TGT AT**A TTA CCG CGG CTG CTG GC-3'; bold: Ion Torrent specific adapters; italics: sequencing key for calibrating signal intensity at the beginning of the sequencing run; regular font: template-specific primers; X…X: 10 bp barcode, see **Table 1** for some barcode sequences), the 2x HotStartTaq Plus master mix and the samples. Mix thoroughly before use, except for the master mix that should be mixed gently.
2. Prepare a PCR mix (20 µl reactions) with 0.5 µM of the reverse primer, 0.4 mg/ml of Bovine Serum Albumine (BSA) and 1x HotStartTaq Plus Master mix. Add 18.5 µl of the master mix to each PCR tube.
3. Add a forward primer (0.5 µl at 20 mM) with a different barcode for each of the samples that will be sequenced together. Add 1 µl of sample (1-10 ng/µl) to the PCR tube.
4. Place the tubes in a PCR machine and run the following program: initial denaturation at 95 °C for 5 min; 25 cycles of 95 °C for 30 sec, 55 °C for 30 sec and 72 °C for 45 sec; final elongation at 72°C for 10 min. PCR products can be kept at 4 °C O/N or frozen until used.

2. Amplicon Purification, Quantification and Pooling

1. Prepare a 2% agarose gel using the largest combs available. Add 5 µl of 6x gel loading dye to the reaction and load on the gel, separating each sample by an empty well. Run the gel at 70 V for 50 min.
2. Take a picture of the gel and cut the bands (expected size around 250 bp: 190 bp product plus 63 bp adapters and barcodes) using a clean scalpel.
3. Weigh the excised bands and proceed with the gel extraction and purification of the PCR products (with *e.g.,* Qiaquick Gel Extraction kit).
4. Quantify each PCR product with PicoGreen and make dilutions to obtain a stock solution at 5×10^9 molecules/µl (approximately 1 ng/µl). If some samples show lower amplification, adjust the dilutions to a lower concentration, keeping in mind that the lowest concentration suitable for subsequent procedures is 1.57×10^7 molecules/µl (26 pM).
5. Pool 10 µl of each diluted PCR product to obtain an equimolar pool of samples. Prepare one separate pool for each of the planned sequencing run. The pooled PCR products can be kept in the freezer until used.

3. Emulsion PCR and Sequencing

1. Perform emulsion PCR:
    1. Dilute the pooled PCR products to 26 pM in a volume of 25 µl. Thaw reagents from an Ion PGM template kit.
    2. Prepare the emulsion PCR mix (reagents provided in the kit) in a PCR hood and add the pooled PCR products and the sphere particles (provided) on the bench. Mix thoroughly and insert the mixture in a filter cartridge and carefully top with emulsion oil (provided).
    3. Slowly reverse the filter cartridge and load on the automated emulsion PCR apparatus (*e.g.,* Ion One Touch 2 instrument). Select the appropriate program and start the procedure.
2. After the automated emulsion PCR has completed, remove the supernatant from the collection tubes and retrieve the sphere particles at the bottom of the tubes. Resuspend the spheres in 100 µl of wash solution (provided in the kit) and take a 2.0 µl aliquot for library quantification.
3. Perform the library quantification.
    1. Mix 100 µl of SSPE buffer (150 mM NaCl, 10mM NaPO₄, 1mM Na₂-EDTA) with 2 µl of Adapter B'-Fam (5'-*FAM*-CTG AGA CTG CCA AGG CAC ACA GGG GAT AGG-3') probe and 2 µl of Adapter A-Cy5 (5'-*Cy5*-CCA TCT CAT CCC TGC GTG TCT CCG ACT CAG-3') probe. Add 52 µl of this mixture to the 2.0 µl aliquot taken in step 3.2 and add 2 µl of wash solution (from the template kit) to the remaining 52 µl as a blank for quantitation.
    2. Incubate at 95 °C for 2 min, then at 37 °C for 2 min. Transfer the mixtures in 1.5 ml tubes.
    3. Add 1.0 ml of TEX buffer (10 mM Tris, 1 mM EDTA, 0.01% Triton X-100, pH 8.0), mix and centrifuge at 15,500 x g for 3 min at RT. Remove the supernatant and leave 20 µl in each tube. Repeat this washing procedure two times.
    4. Add 180 µl of TEX buffer, resuspend the pelleted particles and transfer into 500 µl Qubit tubes.
    5. Measure fluorescence for FAM and Cy5 using a Qubit apparatus and calculate the ratio of positive spheres using the calculator available on the Ion Community website.
4. Enrich the remainder of the sphere particles for positive spheres (containing amplified DNA) using an automated enrichment system (*e.g.,* Ion One Touch Enrichment System).
    1. Pipet reagents in the appropriate wells of the provided 8-well strip: Well 1: sphere particles from step 3.2; Well 2: pre-washed MyOne Streptavidin C1 beads (13 µl); Well 3, 4 and 5: wash solution (provided in the template kit); Well 6: empty; Well 7: melt-off solution (125 mM NaOH, 0.1% Tween 20); Well 8: empty.

2. Install a tip on the pipetting arm and a 0.2 ml tube for sample collection. Start the instrument. At the end of the enrichment, add 10 µl of neutralizing solution (provided). The enriched beads can be kept a 4 °C for up to 15 days.

5. Initialize the sequencing instrument:
   1. Connect to the sequencing instrument and prepare a sequencing run plan, specifying the details of the sequencing run.
   2. Install the wash solution #2 (provided in the Ion PGM sequencing kit), the wash solution #1 (350 µl 100 mM NaOH) and the auto-pH buffer solution (provided).
   3. Start the initialization procedure and when prompted add the dATP, dGTP dCTP and dTTP nucleotides (provided) in their respective 50 ml tubes. Screw the tubes onto the sequencing machine and continue the initialization procedure. When the initialization is completed, start the run procedure immediately.

6. Prepare samples for sequencing and load the chip:
   1. Collect the enriched spheres from step 3.4 at the bottom of the tube by centrifuging at 15,500 x g for 1.5 min. Remove the supernatant leaving exactly 3 µl.
   2. Add 3 µl of sequencing primer (provided in the sequencing kit), mix thoroughly by pipetting up and down to resuspend the particles. Incubate at 95 °C for 2 min, then at 37 °C for 2 min.
   3. Add 1 µl of polymerase (provided), mix well and incubate at RT for 5 min. Proceed to load the enzyme-sphere mixture on the sequencing chip within 30 min.
   4. During incubation, load the sequencing chip on the sequencer and perform a chip quality check. Remove the liquid from the chip by centrifugation and load the sample in the chip by slowly pipeting down the enzyme-sphere mix (7 µl) and avoiding the introduction of bubbles in the chip.
   5. Centrifuge the chip three times for 1 min in a microcentrifuge. Change the orientation of the chip at each centrifugation and mix by pipetting up and down in between each run.
   6. Load the chip in the instrument and start the run.

4. Basic Sequence Data Analysis

1. Connect to the sequencing data server and download the fastq file. Create a tab-delimited "oligos" file containing the primer and barcode information (see example in **Table 1**). Download the Greengenes reference files from the Mothur website (http://www.mothur.org/wiki/Taxonomy_outline).
2. Launch Mothur and perform analyses:
   1. Convert the fastq file to a fasta and a quality file using the following command: *fastq.info(fastq=test.fastq)*
   2. Determine the group membership of each sequence and trim the sequences using the following command: *trim.seqs(fasta=test.fasta, oligos=barcode.txt, qfile=test.qual, qwindowaverage=20, qwindowsize=50, minlength=150, keepforward=T)*
   3. The various options can be modified according to the stringency desired
   4. Classify the sequences in the greengenes taxonomy using the following command: *classify.seqs(fasta=test.trim.fasta, method=wang, group=test.groups, template=gg_99_otus.fasta, taxonomy=gg_99_otus.tax, cutoff=50)*

Representative Results

After purification on gel, with 25 cycles of PCR amplification, the amplification products are usually at a concentration of 0.2-10.0 ng in 50 µl of water. This may vary widely depending on the starting DNA concentration, the type of sample and the purification kit used. It is recommended to keep the number of PCR cycles to the lowest possible to avoid chimera formation and decrease amplification biases, keeping in mind that all samples should be amplified using the same number of cycles. To minimize the number of polyclonal reads and empty spheres and maximize the number of reads, the Qubit ratio should be between 0.1 and 0.3 and the FAM fluorescence should be above 200. Using a 314 chip on an Ion Torrent PGM, the average output is around 0.3-0.5 M good quality reads after filtering of the results in Mothur. **Table 2** shows a typical breakdown of the number of reads after each step of the procedure for a run containing 36 multiplexed environmental samples amplified with primers targeting the V3-4 region of the 16S and analyzed using Mothur. In Mothur, the trim.seqs procedure generate a "*.trim.fasta" file containing the sequences that passed the quality filters and a "*.scrap.fasta" that contains the sequences that did not pass the quality filters along with the reason for rejection in the sequence header. When supplied with barcodes in the "oligos" file, this command will also generate a "*.groups" file that contains the group membership of every sequence based on the barcode sequence. The classify.seqs procedure generates a ".tax.summary" that can be opened in Excel. This file contains the summary of the taxonomic affiliation (in lines) for each of the samples (in columns). This file can be used for downstream statistical analyses and to visualize community composition at various taxonomic levels. The ".taxonomy" file contains the detailed taxonomic affiliation for each sequence. The average community composition at the phylum/class level across all 36 samples is shown in **Figure 1**.
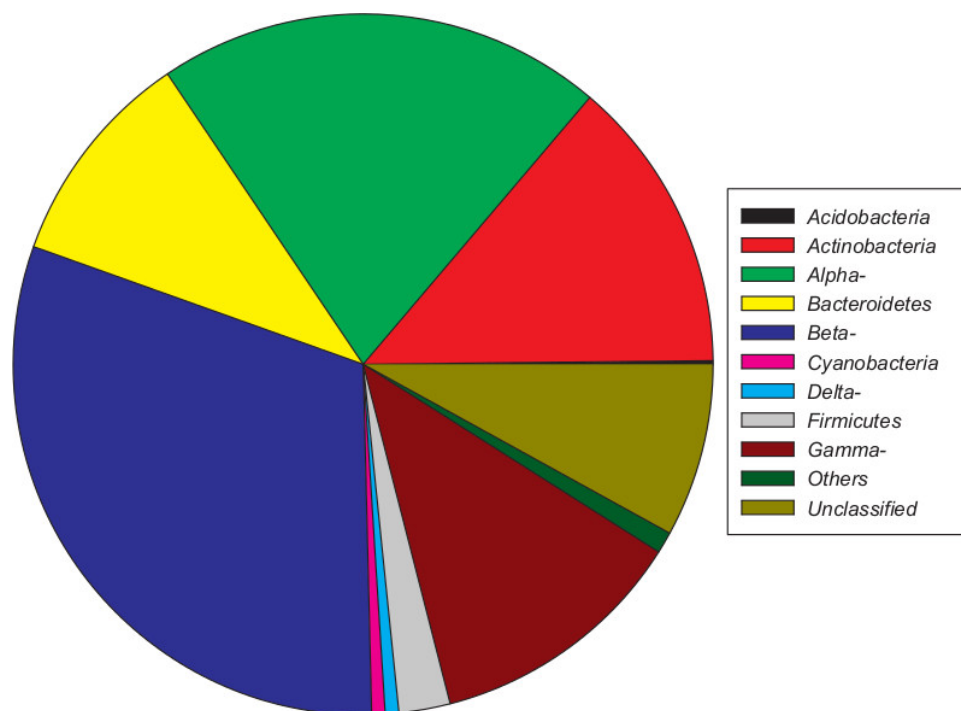
Figure 1. Average community composition at the phylum/class level across all samples.

forward	TACGGRAGGCAGCAG	
barcode	CTAAGGTAAC	Sample01
barcode	TAAGGAGAAC	Sample02
barcode	AAGAGGATTC	Sample03
barcode	TACCAAGATC	Sample04
barcode	CAGAAGGAAC	Sample05
barcode	CTGCAAGTTC	Sample06
barcode	TTCGTGATTC	Sample07
barcode	TTCCGATAAC	Sample08
barcode	TGAGCGGAAC	Sample09
barcode	CTGACCGAAC	Sample10

Table 1. Example "oligos" file for use in Mothur.

	# of reads	% of previous step	Avg. per sample
Number of wells	1,262,519	-	35,070
Wells with beads	1,114,108	88.20%	30,947
Beads with templates	1,112,746	99.90%	30,910
Monoclonal beads	826,805	74.30%	22,967
Good quality reads (Output from the sequencer)	782,204	94.60%	21,728
Pass Mothur filters (min. avg. quality score of 20 over a 50bp window, min. length of 150bp)	372,168	47.60%	10,338
Classified at the phylum level in GreenGenes (50% confidence threshold)	342,171	91.90%	9,505
Classified at the family level in GreenGenes (50% confidence threshold)	316,512	92.50%	8,792

Classified at the genus level in GreenGenes (50% confidence threshold)	289,899	91.60%	8,053

Table 2. Number of reads produced from a typical run for 36 environmental samples multiplexed on one Ion Torrent 314 chip.

Discussion

The method presented here is straightforward and inexpensive, and should allow many laboratories to access the power of metagenomic sequencing. Although it varies depending on the sequencing platform used, once the libraries are constructed very little hands-on time is required, with most of the process being automatized. For the sequencing platform used here (Ion Torrent PGM), the complete procedure can be performed within two days of work. At the moment of writing (September 2013), the reagent costs related to the example detailed above were as follows: PCR amplification of 36 samples: $25, gel purification and PicoGreen DNA quantitation of 36 samples: $125, emulsion PCR for one pooled amplicon sample: $150 and sequencing reagents: $250, for a total of $550 or $15 per sample or $0.0015 per quality-filtered read. This price does not include instrument service contract, instrument depreciation, technician salary and laboratory space usage.

One of the most important steps is to pool all the products in an equimolar ratio, in order to retrieve similar number of reads for each of the samples. PicoGreen quantification was used here, but other methods might be suitable, though less accurate (*e.g.,* UV quantification, gel-based quantification). Even by doing the most accurate quantification and pooling, there is some variability in the number of reads per sample, and in the typical run detailed in **Table 2**, it ranges from 4,380 to 32,750 reads, with an average of 10,338 reads. If processing large number of samples (more than 40-50), single column gel purification can be replaced by gel purification in plate or purification using beads with a stringent size cutoff (*e.g.,* AMPure beads).

To date, the most used next-generation sequencing technology for the 16S rRNA gene is 454. The Ion Torrent sequencing technology used in this protocol is conceptually very similar to 454 and both technologies are prone to the same type of sequencing errors. Not surprisingly, it was shown that Ion Torrent sequencing resulted in sequencing results very similar to 454 sequencing[10]. Recently, many researchers have explored the use of Illumina technology for 16S rRNA gene amplicon sequencing[18,19]. In any case, it would be easy to adapt the current protocol for other benchtop sequencers like the Illumina MiSeq or the 454 GS Junior by changing the fusion primer sequences to match the adapters and barcodes needed for these sequencing technologies, like in the method recently described for the Illumina MiSeq[19]. Alternatively, researchers could follow steps 1 and 2 of the protocol detailed here and send the pooled amplicons to a sequencing center where the emulsion PCR and sequencing would be performed.

The 16S rRNA gene reads were trimmed and classified using Mothur, but many other analyses can be performed on 16S rRNA gene amplicons. For instance, beta diversity can be evaluated by calculating the Unifrac distances between each sample pair using the procedure outlined at http://unifrac.colorado.edu/[20]. Alpha diversity indices and number of operational taxonomical units of each sample can be calculated using tools within QIIME like AmpliconNoise [21] or using the procedure outlined by Huse *et al.*[22] and available within Mothur.

The primers used here amplified the variable regions 3 and 4 from the 16S rRNA gene, but many other regions could be targeted. In present study, 16S rRNA genes were amplified from plant material and the choice of primer was made to avoid amplification of chloroplast 16S rRNA gene[23,24]. There are a wide variety of other primers available that vary in term of the product length, taxonomic power and usefulness[25,26]. However, in all cases 200-400 bp reads of the 16S rRNA gene cannot be reliably classified at the species level, and analyses are limited to the genus and higher taxonomical levels. Other genes could be more appropriate if species level information is needed, like the cpn60 and rpoB genes[27,28]. Future drastic drops in the cost of sequencing and increases in the power of analytical tools might make it feasible to replace 16S rRNA gene sequencing by shotgun metagenomics, but until then 16S rRNA gene sequencing remains the gold standard of environmental microbiology.

Disclosures

The authors have nothing to disclose.

Acknowledgements

References

1. Sipos, R. *et al.* Addressing PCR biases in environmental microbiology studies. *Methods Mol. Biol.* **599**, 37-58, doi:10.1007/978-1-60761-439-5_3 (2010).
2. Schloss, P. D. *et al.* Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. *PLoS ONE.* **6**, e27310, doi:10.1371/journal.pone.0027310 (2011).
3. Tringe, S. G., & Hugenholtz, P. A renaissance for the pioneering 16S rRNA gene. *Curr. Opin. Microbiol.* **11**, 442–446, doi:10.1016/j.mib.2008.09.011 (2008).
4. Margulies, M. *et al.* Genome sequencing in microfabricated high-density picolitre reactors. *Nature.* **437**, 376–380, doi:10.1038/nature03959 (2005).
5. Kuczynski, J. *et al.* Microbial community resemblance methods differ in their ability to detect biologically relevant patterns. *Nat. Methods.* **7**, 813-819, doi:10.1038/nmeth.1499 (2010).

6. Bondici, V. F. *et al.* Microbial communities in low-permeability, high pH uranium mine tailings: characterization and potential effects. *J. Appl. Microbiol.* **113**, 1671-1686, doi:10.1111/jam.12180 (2013).
7. Auffret, M. *et al.* Impact of water quality on the bacterial populations and off-flavours in recirculating aquaculture systems. *FEMS Microbiology Ecology.* **84**, 235-247, doi:10.1111/1574-6941.12053 (2013).
8. Bell, T. H., Yergeau, E., Juck, D., Whyte, L. G., & Greer, C. W. Alteration of microbial community structure affects diesel biodegradation in an Arctic soil. *FEMS Microbiol. Ecol.* **85**, 51-61, doi:10.1111/1574-6941.12102 (2013).
9. Bell, T. H. *et al.* Predictable bacterial composition and hydrocarbon degradation in Arctic soils following diesel and nutrient disturbance. *ISME J.* **7**, 1200-1210, doi:10.1038/ismej.2013.1 (2013).
10. Yergeau, E. *et al.* Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities. *Appl. Environ. Microbiol.* **78**, 7626-7637, doi:10.1128/AEM.02036-12 (2012).
11. Yergeau, E. *et al.* Aerobic biofilms grown from Athabasca watershed sediments are inhibited by increasing bituminous compounds concentrations. *Appl. Environ. Microbiol.* **79**, 7398-7412, doi:10.1128/AEM.02216-13 (2013).
12. Yergeau, E. *et al.* Microbial expression profiles in the rhizosphere of willows depend on soil contamination. *ISME J.* **8**, 344-358, doi:10.1038/ismej.2013.163 (2013).
13. Deagle, B. E. *et al.* Quantifying sequence proportions in a DNA-based diet study using Ion Torrent amplicon sequencing: which counts count? *Mol. Ecol. Resour.* **13**, 620-633, doi:10.1111/1755-0998.12103 (2013).
14. Milani, C. *et al.* Assessing the fecal microbiota: an optimized Ion Torrent 16S rRNA gene-based analysis protocol. *PLoS ONE.* **8**, e68739, doi:10.1371/journal.pone.0068739 (2013).
15. Jünemann, S., Prior, P., Szczepanowski, R., Harks, I., Ehmke, B., Goesmann, A., Stoye, J., & Harmsen, Dag. Bacterial community shift in treated periodontitis patients revealed by Ion Torrent 16S rRNA gene amplicon sequencing. *PLoS ONE.* **7**, e41606, doi:10.1371/journal.pone.0041606 (2012).
16. Petrof, E. *et al.* Stool substitute transplant therapy for the eradication of *Clostridium difficile* infection: 'RePOOPulating' the gut. *Microbiome.* **1**, 3, doi:10.1186/2049-2618-1-3 (2013).
17. Whiteley, A. S. *et al.* Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. *J. Microbiol. Meth.* **91**, 80-88, doi:10.1016/j.mimet.2012.07.008 (2012).
18. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6**, 1621-1624, doi:10.1038/ismej.2012.8 (2012).
19. Kozich, J. J. *et al.* Development of a dual-index sequencing strategy and curation pipeline for analyzing amplicon sequence data on the MiSeq Illumina sequencing platform. *Appl. Environ. Microbiol.* **79**, 5112-5220, doi:10.1128/AEM.01043-13 (2013).
20. Hamady, M. *et al.* Fast UniFrac: facilitating high-throughput phylogenetic analyses of microbial communities including analysis of pyrosequencing and PhyloChip data. *ISME J.* **4**, 17-27, doi:10.1038/ismej.2009.97 (2010).
21. Quince, C. *et al.* Removing noise from pyrosequenced amplicons. *BMC Bioinformatics.* **12**, 38, doi:10.1186/1471-2105-12-38 (2011).
22. Huse, S. M. *et al.* Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environ. Microbiol.* **12**, 1889-1898, doi:10.1111/j.1462-2920.2010.02193.x (2010).
23. Edwards, J. E. *et al.* Characterization of the dynamics of initial bacterial colonization of nonconserved forage in the bovine rumen. *FEMS Microbiol. Ecol.* **62**, 323-335, doi:10.1111/j.1574-6941.2007.00392.x (2007).
24. Rastogi, G. *et al.* A PCR-based toolbox for the culture-independent quantification of total bacterial abundances in plant environments. *J. Microbiol. Methods.* **83**, 127-132, doi:10.1016/j.mimet.2010.08.006 (2010).
25. Baker, G. C. *et al.* Review and re-analysis of domain-specific 16S primers. *J. Microbiol. Methods.* **55**, 541–555, doi:10.1016/j.mimet.2003.08.009 (2003).
26. Schloss, P. D. The effects of alignment quality, distance calculation method, sequence filtering, and region on the analysis of 16S rRNA gene-based studies. *PLoS Comput. Biol.* **6**, e1000844, doi:10.1371/journal.pcbi.1000844 (2010).
27. Links, M. G. *et al.* The chaperonin-60 universal target Is a barcode for bacteria that enables *de novo* assembly of metagenomic sequence data. *PLoS ONE.* **7**, e49755, doi:10.1371/journal.pone.0049755 (2012).
28. Vos, M. *et al.* Comparison of *rpoB* and 16S rRNA as markers in pyrosequencing studies of bacterial diversity. *PLoS ONE.* **7**, e30600, doi:10.1371/journal.pone.0030600 (2012).