The following supplement accompanies the article

Protist diversity along a salinity gradient in a coastal lagoon

Sergio Balzano*, Elsa Abs, Sophie C. Leterme

*Corresponding author: sergio.balzano@nioz.nl

Aquatic Microbial Ecology 74: 263–277 (2015)

Supplement 1.

Nutrient analyses

For nutrient analyses 100 mL water were filtered in triplicate through bonnet syringe Minisart filters (0.45 μm pore size, Sartorius Stedim, Dandenong, Australia) to remove large particles, transported on ice and stored at -20 °C until analyses. The concentration of dissolved silica, ammonium, orthophosphate and the combined concentrations of nitrate and nitrite (nitrate/nitrite) were measured simultaneously every two weeks using a Lachat Quickchem Flow Injection Analyser (FIA) and carried out following published methods (Hansen & Koroleff 2007).

Flow Cytometry

Total bacteria, virus-like particles (VLPs), cyanobacteria (*Prochlorococcus* and *Synechococcus* populations) and photosynthetic picoeukaryotes were enumerated using flow cytometry. From each sampling site, 1 mL water was collected in 6 replicates, 3 of them were fixed with 0.5 % gluteraldehyde (Proscitech, thuringowa, Australia) for bacteria and VLPs enumeration and 3 replicates were fixed with 2 % paraformaldehyde (Proscitech) for picophytoplankton. The samples were then flash frozen in liquid nitrogen and stored at -80 °C until analysis. For the enumeration of total bacteria and VLPs the samples were diluted 1:10 in TE buffer (10 mM Tris-HCl, 1mM EDTA, pH 8, National Diagnostics, Atlanta, USA) and DNA from the cells was stained with 2.5 % (w/v) SYBR I Green (Invitrogen, Carlsbad, USA). The samples were incubated at 80 ^{0}C for 10 min and fluorescent marker beads (1 μL, Molecular probes, Eugene, USA) were added afterwards to all samples as an internal size and concentration standard prior to analyses which were performed using a FacsCanto (Becton Dickinson, San José, USA). Cyanobacteria and picoeukaryotes were enumerated from the

samples based on Chl-a autofluorescence, marker beads were added to samples and samples were analyzsed on a FacsCanto as described previously (Marie et al. 1997).

Phytoplankton identification

For phytoplankton identification and enumeration samples were filtered through 5 µm pore size Sterlitech mixed cellulose ester membranes (Sterlitech, Kent, USA). The cells on the filters were resuspended in a smaller volume of filtrate from the same sample. This suspension (1 mL) was then pipetted into a Sedgewick Rafter counting chamber and counted using a Zeiss Axiolab upright microscope equipped with bright-field and phase contrast optics (Carl Zeiss Microscopy, Thornwood, USA). The cells were identified up to the genus or species level based on their key taxonomic features (Tomas 1997, Hallegraef et al. 2010).

Sample collection

For small (< 10 µm) eukaryotes, 1 L water samples were collected in sterile polyethylene bottles and pre-filtered through 10 µm isopore membrane filters (Merck Millipore, Kilsyth, Australia) using a Millivac vacuum pump (Millipore, Billerica, USA). Samples were subsequently filtered using sterile syringes through 0.45 µm Sterivex filters (Merck Millipore). Then, 2 mL of Lysis Buffer (0.75 M sucrose, 50mM Tris-HCl, 40mM EDTA; pH 8.3) were added to the Sterivex filters prior to flash freezing in liquid nitrogen as described before (Marie et al. 2010). Filters were then stored at - 80 °C until analyses.

DNA extraction

For DNA extraction, Sterivex filters were thawed on ice and agitated for 20 min on a horizontal shaker. The Lysis Buffer containing DNA was then removed from the Sterivex filters using sterile syringes. DNA was then extracted as described previously (Balzano et al. 2012). In brief, lysozyme (2 mL, 20 mg/mL) and proteinase K (289 µL, 50 µM) were added to 2 mL of samples which were then incubated at 55 °C for 30 min. The Proteinase K was then inactivated at 70 °C for 10 min. 1.9 mL of 100 % molecular grade ethanol (Chem-supply, Gillman, Australia) was then added to the samples and DNA was precipitated from the samples using a cell and tissue DNA extraction kit and following the instruction provided by the supplier (Macherey-Nagel, Düren, Germany).

Sequencing

Sequencing was performed in two distinct Ion Torrent runs, for November 2012 and June 2013 samplings, respectively. Samples were shipped to Australian Genome Research Facility (AGRF, Brisbane, Australia) where they were pooled in equimolar amounts before sequencing. Sequencing was performed at AGRF using an Ion Torrent Personal Genome Machine (PGM, Life Technology, Mulgrave, Australia) provided with a 318 chip (Life Technology) and adapted for a maximum read length of 400 bp.

Bioinformatic pipeline

Sequence data were downloaded as 2 fastq files (one for run) and analysed using the Quantitative Insight In Microbial Ecology (QIIME) pipeline (Caporaso et al. 2010). Reads were filtered, attributed to the different samples based on their barcode sequences, and their barcodes were removed using the script *split_library.py*. Since the Ion Torrent quality scores underestimate the base accuracy (Bragg et al. 2013) we used a phred quality threshold of 20 instead of 25 to filter our reads, similarly to a previous studies based on this sequencing platform (Frank-Fahle et al. 2014). Reads were also filtered according to their length and primer matching as described previously (Behnke et al. 2011, Logares et al. 2012, Brown et al. 2013). Overall only sequences with an exact match to the forward primer, a length between 230 and 470 bp from the 3' end of the forward primer, an average phred quality score > 20 on a sliding windows of 50 bp, no ambiguous bases, a number of homopolymers not exceeding 8, were considered for further analyses. Reads that passed the quality filtering were then separated in different samples according to the barcodes. Sequences were then truncated to 230-bp length after removal of both barcode and forward primer. Chimeras were identified by comparison with the 18S rDNA gene sequences present on the Protist Ribosomal Database (Guillou et al. 2013) and removed from the dataset using the command *identify_chimeric_seqs.py* and the UCHIME algorithm (Edgar et al. 2011). Reads were then clustered using *pick_otus.py* with uclust algorithm (Edgar 2010). Reads were first grouped at 100 % similarity to remove singletons (every read that did not share 100 % identity with at least another read), and then clustered in 1321 distinct Operational Taxonomic Units (OTUs) based on 97 % similarity. The dataset was then rarefied to 16 855 reads using *single_rarefaction.py* and subsequently the number of OTUs decreased to 1204. The most abundant sequences were extracted from each OTU using *pick_rep_set.py* and then aligned using muscle (Edgar 2004). The scripts *filter_alignment.py* and *make_phylogeny.py* were then used to cure the alignment and construct a phylogenetic tree, respectively with the

method fastTree (Price et al. 2010). The taxonomic affiliation of our representative set was inferred by comparison with the Protist Ribosomal Database (Guillou et al. 2013) using *assign_taxonomy.py* and the uclust algorithm (Edgar 2010) and taxa were summarised using *make_otu_table.py* and *summarise_taxa.py* scripts.

Rarefaction analyses were carried out using *alpha_rarefaction.py* with 100 steps and a maximum number of 16 855 reads per samples. Shannon, Simpson-Gini, Chao1 and phylogenetic diversity (PD) indices were calculated from 10 replicate OTU tables subsampled from our dataset using *multiple_rarefaction_even_depth.py* and our results were then averaged using *collate_alpha.py*.

To compare the distribution of the different OTUs in our samples we compiled a code to count the number of OTU shared between the same station, the same sampling season, the same salinity condition (low brackish < 10, high brackish 10-35, hypersaline >35), the same region (North and South lagoon). The code was compiled using R software (www.r-project.org).

We used the script *beta_diversity.py* to calculate the Bray-Curtis dissimilarities among our different samples as well as their phylogenetic distances based on weighted (relative abundance) UniFrac (Lozupone & Knight 2005). Principal Coordinate Analysis (PCoA) was then carried out on our Bray-Curtis dissimilarities and UniFrac distances using *principal_coordinates.py*. We computed a similarity matrix based on Bray-Curtis similarities as well as weighted UniFrac distances and clustered the results with the unweighted pair group method with arithmetic mean (UPGMA) using the script *jackknifed_beta_diversity.py*.

Statistical analyses

To investigate the influence of environmental conditions and geographic distance on the β-diversity found for the small eukaryotes parameter we performed aMantel test using the vegan package. Small eukaryote dissimilarities (Bray-Curtis and weighted UniFrac) between our samples were compared with five matrices reflecting differences in temperature, salinity, geographic distance, environmental distance and large phytoplankton. Similarly to a previous study (Lepere et al. 2013), the environmental data included temperature, salinity, pH (Table 1), nutrient concentrations (Table 2) as well as the distribution of bacteria, VLP, *Prochlorococcus* and picoeukaryotes measured by flow cytometry. Temperature, salinity, geographic distance and environmental data were standardised with the standard normal deviate equivalents (SNDE) as described previously (Pagaling et al. 2009), with the equation

SNDE = (x – mean of the raw data)/standard deviation of the raw data, where x represents the raw data for one sampling site. The Euclidean distances between our different samples were then calculated for these five parameters and compiled in distance matrices. Simple Mantel tests were then performed between the different matrices using the R package vegan (www.cran.r-project.org/web/packages/vegan/index.html) using 9999 permutations.

Ion Torrent data

Ion Torrent data consisted in 477,303 good reads > 230 bp with a phred quality score > 20. The reads were truncated at 230 bp rather than using the entire V4 fragment length (~400) to recover a higher number of reads allowing a more detailed investigation of the microbial diversity in the Coorong. Both chimera and singletons were then removed from the dataset and overall 343,321 reads were recovered for the Coorong (Table 2).

Methodological consideration

In the present study small eukaryotes (0.45 to 10 µm) were filtered from the Coorong Lagoon and analysed by PCR on the 18S rDNA gene followed by Ion Torrent sequencing. For each run, a different barcode was applied to each sample. Overall > 90 % of our reads were < 250 bp and a < 1 % of our reads were > 400 bp although an Ion torrent sequencing run based on 400 bp chemistry had been performed. Since our amplicons, which also included the barcodes and the adaptors, were too long (> 450 bp), a high proportion (> 90 %) of them could not be sequenced properly resulting in reads < 200 bp. These errors likely occurred during emulsion PCR (AGRF personal communication).

We selected 230 bp as a compromise between recovering a high number of reads and recovering reads long enough for taxonomic analyses. Although the V4 region is 400 bp long (Logares et al. 2012) and most studies use longer reads, V4 sequences as short as 200 bp have been successfully used for analyses previously (Behnke et al. 2011). Only 40 953 reads were > 350 bp whereas using a threshold of 230 bp we recovered 343 321 reads. The truncation of our reads at 230 bp might have led to an underestimation of our microbial diversity, especially at lower taxonomy levels, since rDNA sequences from distinct but closely related taxa might share higher identical DNA sequences over short fragments. To investigate the effect of the truncation on our taxonomic results we extracted from our initial dataset our reads > 350 bp and analysed them following the same steps as for our > 230 bp reads. These steps consist in (1) removal of chimera, homopolymers-containing sequences and singleton,

(2) OTU picking at 97 % identity, and (3) taxonomy analyses against Protist Ribosomal Database using uclust (Edgar 2010). The taxonomic composition inferred from our reads > 350 bp does not differ significantly from that obtained in our study (Supplementary Figure S3), at least at the phylum level.

Ion Torrent, as well as other high-throughput sequencing platforms are usually more prone to errors compared to Sanger sequencing. As a consequence most studies based on these platforms analyse diversity at identity threshold lower than 100 % and remove singletons, which are more likely to results from sequencing errors. Insertions and deletions rates as high as 1 % were recently reported for Ion Torrent (Bragg et al. 2013). As a consequence we first removed all our singletons based on 100 % identity and then grouped our reads in OTUs based on 97 % similarity. We did not investigate the diversity of our sequences at an identity > 97 % because above this threshold some 'unique' sequences might result from sequencing errors.

In environmental studies Ion Torrent amplicon sequencing has been previously applied on the ITS operon of the rDNA to investigate fungal communities in soil (Brown et al. 2013, Geml et al. 2014) and plant tissues (Kemler et al. 2013), as well as to investigate prokaryotic 16S rDNA diversity in permafrost (Frank-Fahle et al. 2014), rhizosphaere (Yergeau et al. 2014) and mines (Khan et al. 2013). In all these studies the authors used 100 bp to 300 bp Ion Torrent chemistry since they were targeting fragments shorter than the V4.

Ion Torrent has been also recently used to sequence the metagenome of a saline desert (Pandit et al. 2014).

References for the supplementary information

Balzano S, Gourvil P, Siano R, Chanoine M, Marie D, Lessard S, Sarno D, Vaulot D (2012) Diversity of cultured photosynthetic flagellates in the northeast Pacific and Arctic Oceans in summer. Biogeosciences 9:4553-4571

Behnke A, Engel M, Christen R, Nebel M, Klein RR, Stoeck T (2011) Depicting more accurate pictures of protistan community complexity using pyrosequencing of hypervariable SSU rRNA gene regions. Environmental Microbiology 13:340-349

Bragg LM, Stone G, Butler MK, Hugenholtz P, Tyson GW (2013) Shining a Light on Dark Sequencing: Characterising Errors in Ion Torrent PGM Data. Plos Computational Biology 9

Brown SP, Callaham MA, Jr., Oliver AK, Jumpponen A (2013) Deep Ion Torrent sequencing identifies soil fungal community shifts after frequent prescribed fires in a southeastern US forest ecosystem. Fems Microbiology Ecology 86:557-566

Caporaso JG, Kuczynski J, Stombaugh J, Bittinger K, Bushman FD, Costello EK, Fierer N, Pena AG, Goodrich JK, Gordon JI, Huttley GA, Kelley ST, Knights D, Koenig JE, Ley RE, Lozupone CA, McDonald D, Muegge BD, Pirrung M, Reeder J, Sevinsky JR, Tumbaugh PJ, Walters WA, Widmann J, Yatsunenko T, Zaneveld J, Knight R (2010) QIIME allows analysis of high-throughput community sequencing data. Nature Methods 7:335-336

Edgar RC (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research 32:1792-1797

Edgar RC (2010) Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26:2460-2461

Edgar RC, Haas BJ, Clemente JC, Quince C, Knight R (2011) UCHIME improves sensitivity and speed of chimera detection. Bioinformatics 27:2194-2200

Frank-Fahle BA, Yergeau E, Greer CW, Lantuit H, Wagner D (2014) Microbial Functional Potential and Community Composition in Permafrost-Affected Soils of the NW Canadian Arctic. Plos One 9

Geml J, Pastor N, Fernandez L, Pacheco S, Semenova TA, Becerra AG, Wicaksono CY, Nouhra ER (2014) \ Large-scale fungal diversity assessment in the Andean Yungas forests reveals strong community turnover among forest types along an altitudinal gradient. Molecular Ecology 23:2452-2472

Guillou L, Bachar D, Audic S, Bass D, Berney C, Bittner L, Boutte C, Burgaud G, de Vargas C, Decelle J, del Campo J, Dolan JR, Dunthorn M, Edvardsen B, Holzmann M, Kooistra WHCF, Lara E, Le Bescot N, Logares R, Mahé F, Massana R, Montresor M, Morard R, Not F, Pawlowski J, Probert I, Sauvadet A-L, Siano R, Stoeck T, Vaulot D, Zimmermann P, Christen R (2013) The Protist Ribosomal Reference database (PR2): a catalog of unicellular eukaryote Small Sub-Unit rRNA sequences with curated taxonomy. Nucleic Acids Research 41:D597-D604

Hallegraef GM, Bolch CJS, Hill DRA, Jameson I, Leroi JM, McMinn A, Murray S, de Salas MF, Saunders KM (2010) Algae of australia: phytoplankton of temperate coastal waters, Vol 432. CSIRO publishing, Melbourne

Hansen H, Koroleff F (2007) Determination of nutrients. In: Grasshoff K, Kremling K, Ehrhardt M (eds) Methods of seawater analysis

Kemler M, Garnas J, Wingfield MJ, Gryzenhout M, Pillay K-A, Slippers B (2013) Ion Torrent PGM as Tool for Fungal Community Analysis: A Case Study of Endophytes in Eucalyptus grandis Reveals High Taxonomic Diversity. Plos One 8

Khan NH, Bondici VF, Medihala PG, Lawrence JR, Wolfaardt GM, Warner J, Korber DR (2013) Bacterial Diversity and Composition of an Alkaline Uranium Mine Tailings-Water Interface. J Microbiol 51:558-569

Lepere C, Domaizon I, Taib N, Mangot JF, Bronner G, Boucher D, Debroas D (2013) Geographic distance and ecosystem size determine the distribution of smallest protists in lacustrine ecosystems. Fems Microbiology Ecology 85:85-94

Logares R, Audic S, Santini S, Pernice MC, de Vargas C, Massana R (2012) Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. Isme Journal 6:1823-1833

Lozupone C, Knight R (2005) UniFrac: a new phylogenetic method for comparing microbial communities. Appl Environ Microbiol 71:8228-8235

Marie D, Partensky F, Jacquet S, Vaulot D (1997) Enumeration and cell cycle analysis of natural populations of marine picoplankton by flow cytometry using the nucleic acid stain SYBR Green I. Appl Environ Microbiol 63:186-193

Marie D, Shi XL, Rigaut-Jalabert F, Vaulot D (2010) Use of flow cytometric sorting to better assess the diversity of small photosynthetic eukaryotes in the English Channel. FEMS Microbiol Ecol 72:165-178

Pagaling E, Wang H, Venables M, Wallace A, Grant WD, Cowan DA, Jones BE, Ma Y, Ventosa A, Heaphy S (2009) Microbial Biogeography of Six Salt Lakes in Inner Mongolia, China, and a Salt Lake in Argentina. Appl Environ Microbiol 75:5750-5760

Pandit AS, Joshi MN, Bhargava P, Ayachit GN, Shaikh IM, Saiyed ZM, Saxena AK, Bagatharia SB (2014) Metagenomes from the saline desert of kutch. Genome announcements 2

Price MN, Dehal PS, Arkin AP (2010) FastTree 2-Approximately Maximum-Likelihood Trees for Large Alignments. Plos One 5

Tomas CR (1997) Identifying marine phytoplankton, Vol

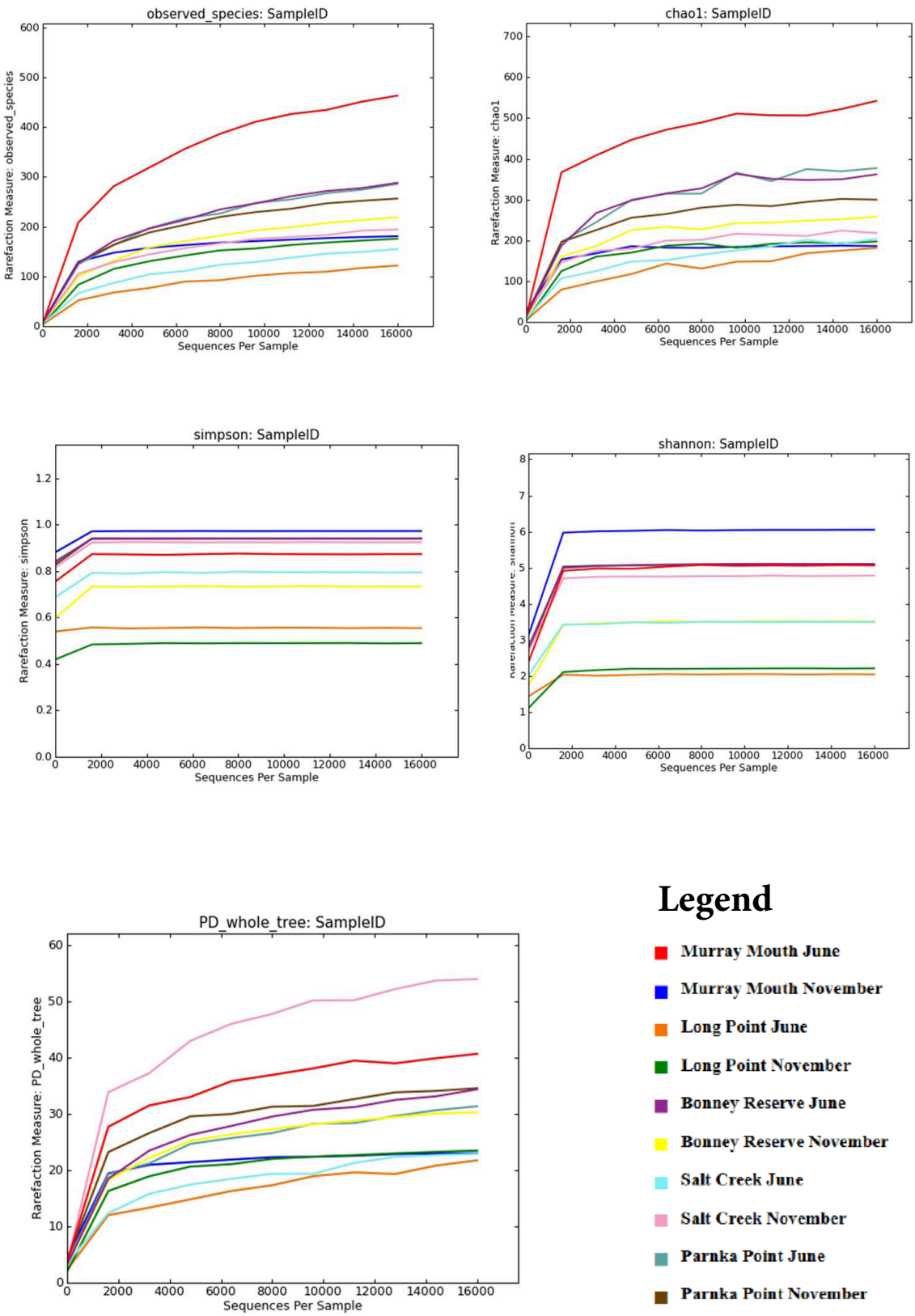
Yergeau E, Sanschagrin S, Maynard C, St-Arnaud M, Greer CW (2014) Microbial expression profiles in the rhizosphere of willows depend on soil contamination. Isme Journal 8:344-358

Supplementary Figure Legends

Supplementary Figure S1. Rarefaction curves based on Ion Torrent sequencing of the V4 region of the 18S rRNA gene, for the different samples analysed in the present study. The rarefaction analyses were carried out on the number of operational taxonomic units (OTUs, based on 97 % sequence identity), Chao1, Simpson-Gini and Shannon indices as well as the phylogenetic diversity (PD).

Supplementary Figure S2. Comparison between the taxonomic assignments inferred to our Ion Torrent data, after truncating the reads at 230 and 350 bp, respectively. Please note the higher number of reads on the upper panel.

Supplementary Figure S1



Supplementary Figure S2