Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics

Christian Rinke [Corresp., 1] , Lai Yee Low [1] , Benjamin J Woodcroft [1] , Jean-Baptiste Raina [2] , Adam Skarshewski [1] , Xuyen H Le [1] , Margaret K Butler [1] , Roman Stocker [3] , Justin Seymour [4] , Gene W Tyson [1, 5] , Philip Hugenholtz [Corresp. 1, 6]

[1] Australian Centre for Ecogenomic / School of Chemistry and Molecular Biosciences, University of Queensland, Brisbane, QLD, Australia

[2] Plant Functional Biology and Climate Change Cluster, University of Technology Sydney, Sydney, New South Wales, Australia

[3] Department of Civil, Environmental and Geomatic Engineering, ETH Zurich, Zurich, Switzerland

[4] Climate Change Cluster, University of Technology Sydney, Sydney, New South Wales, Australia

[5] Advanced Water Management Centre, University of Queensland, Brisbane, QLD, Australia

[6] Institute for Molecular Bioscience, University of Queensland, Brisbane, QLD, Australia

Corresponding Authors: Christian Rinke, Philip Hugenholtz
Email address: christian.rinke@gmail.com, phugenholtz@gmail.com

High throughput sequencing libraries are typically limited by the requirement for nanograms to micrograms of input DNA. This bottleneck impedes the microscale analysis of ecosystems and the exploration of low biomass samples. Current methods for amplifying environmental DNA to bypass this bottleneck introduce considerable bias into metagenomic profiles. Here we describe and validate a simple modification of the Illumina Nextera XT DNA library preparation kit which allows creation of shotgun libraries from sub-nanogram amounts of input DNA. Community composition was reproducible down to 100fg of input DNA based on analysis of a mock community comprising 54 phylogenetically diverse Bacteria and Archaea. The main technical issues with the low input libraries were a greater potential for contamination, limited DNA complexity which has a direct effect on assembly and binning, and an associated higher percentage of read duplicates. We recommend a lower limit of 1pg (~100 to 1000 microbial cells) to ensure community composition fidelity, and the inclusion of negative controls to identify reagent-specific contaminants. Applying the approach to marine surface water, pronounced differences were observed between bacterial community profiles of microliter volume samples, which we attribute to biological variation. This result is consistent with expected microscale patchiness in marine communities. We thus envision that our benchmarked, slightly modified low input DNA protocol will be beneficial for microscale and low biomass metagenomics.

1 # Validation of picogram- and femtogram-input DNA libraries

2 # for microscale metagenomics

4 Christian Rinke[a], Lai Yee Low[a], Ben J. Woodcroft[a], Jean-Baptiste Raina[c], Adam Skarshewski[a],

5 Xuyen H. Le[a], Margaret K. Butler[a], Roman Stocker[b], Justin Seymour[c], Gene W. Tyson[ad] and

6 Philip Hugenholtz[ae]

8 Australian Centre for Ecogenomics, School of Chemistry and Molecular Biosciences, The University of

9 Queensland, St. Lucia, Queensland , Australia[a]; Department of Civil, Environmental and Geomatic

10 Engineering, ETH Zurich, Switzerland[b]; Plant Functional Biology and Climate Change Cluster,

11 University of Technology Sydney, Sydney, New South Wales, Australia[c]; Advanced Water Management

12 Centre, The University of Queensland , St. Lucia, Queensland , Australia[d]; Institute for Molecular

13 Bioscience, The University of Queensland , St. Lucia, Queensland, Australia[e]

15 **Corresponding authors:**

16 Christian Rinke (c.rinke@uq.edu.au, +61 7 336 54956); Philip Hugenholtz (p.hugenholtz@uq.edu.au,

17 +61 7 336 53822)

ABSTRACT

High throughput sequencing libraries are typically limited by the requirement for nanograms to micrograms of input DNA. This bottleneck impedes the microscale analysis of ecosystems and the exploration of low biomass samples. Current methods for amplifying environmental DNA to bypass this bottleneck introduce considerable bias into metagenomic profiles. Here we describe and validate a simple modification of the Illumina Nextera XT DNA library preparation kit which allows creation of shotgun libraries from sub-nanogram amounts of input DNA. Community composition was reproducible down to 100fg of input DNA based on analysis of a mock community comprising 54 phylogenetically diverse Bacteria and Archaea. The main technical issues with the low input libraries were a greater potential for contamination, limited DNA complexity which has a direct effect on assembly and binning, and an associated higher percentage of read duplicates. We recommend a lower limit of 1pg (~100 to 1000 microbial cells) to ensure community composition fidelity, and the inclusion of negative controls to identify reagent-specific contaminants. Applying the approach to marine surface water, pronounced differences were observed between bacterial community profiles of microliter volume samples, which we attribute to biological variation. This result is consistent with expected microscale patchiness in marine communities. We thus envision that our benchmarked, slightly modified low input DNA protocol will be beneficial for microscale and low biomass metagenomics.

39 **INTRODUCTION**

40 Over the last decade, advances in high-throughput sequencing technologies have accelerated the

41 exploration of the uncultured microbial majority (Rappe & Giovannoni, 2003). The direct

42 sequencing of environmental samples, termed metagenomics, has revolutionized microbial

43 ecology by providing new insights into the diversity, dynamics and metabolic potential of

44 microorganisms. A remaining limitation of conventional metagenomic library construction is the

45 requirement for relatively large sample amounts, e.g. grams of soil or liters of seawater which

46 comprise millions of microbial cells. However, samples of this size aggregate microbial

47 population heterogeneity and metabolic processes occurring at the microscale.

49 Within natural habitats, specific physiological niches occupied by microorganisms often occur

50 within discrete microenvironments that are several orders of magnitude smaller than typical

51 sample sizes. For example, in the pelagic ocean dissolved and particulate organic matter is often

52 localized within hotspots, ranging in size from tens to hundreds of micrometers (Azam, 1998).

53 These hotspots include marine snow particles, cell lysis and excretions by larger organisms

54 including phytoplankton exudates, which result in microscale chemical gradients that

55 chemotactic bacteria can exploit (Azam & Malfatti, 2007; Stocker, 2012). Specific populations

56 and their associated biogeochemical activities can be restricted to these localized microniches

57 (Paerl & Pinckney, 1996). Therefore, understanding processes occurring at the microscale (μg,

58 μl) is important if we are ever to fully understand ecosystem functionality. Beyond the need for

59 increased ecological resolution, there is a demand for creating metagenomes from small amounts

60 of starting DNA to explore habitats with extremely low biomass such as subseafloor sediments

61   (Kallmeyer et al., 2012), clean-room facilities (Vaishampayan et al., 2013), human skin samples

62   (Probst, Auerbach & Moissl-Eichinger, 2013), and ocean virus samples (Duhaime et al., 2012).

63

64   Template preparation for high throughput sequencing platforms traditionally follows a common

65   workflow, independent of the downstream sequencing chemistry. First the input DNA is sheared

66   to fragments of the desired size by random fragmentation and subsequently platform-specific

67   sequencing adapters are added to the flanking ends in order to attach the library to a solid surface

68   (e.g. flow cell, tagged glass slide, bead) via a complementary sequence. Typically, the input

69   DNA is sheared by sonication, followed by multiple rounds of enzymatic modification to repair

70   the DNA fragments to have blunt ends or A tails and to add the sequencing adapters. This

71   method is labor intensive and requires several tens of nanograms to micrograms of input DNA,

72   making it challenging to prepare sequencing libraries from low yield DNA samples. Linker-

73   amplification comprising ultrasonic shearing, linker ligation and PCR amplification is another

74   approach that has been applied to create low input DNA shotgun libraries from ≤1ng starting

75   DNA (Duhaime et al., 2012; Solonenko et al., 2013). However, this method is time consuming,

76   technically demanding, and known to introduce up to 1.5-fold GC content amplification bias

77   (Duhaime et al., 2012). Multiple displacement amplification (MDA) with phi29 polymerase can

78   increase DNA amounts by nine orders of magnitude allowing femtogram range DNAs to be used

79   for library preparation (Raghunathan et al., 2005). While MDA is successfully applied to obtain

80   single-cell genomes (Clingenpeel et al., 2015), the approach has been shown to significantly

81   skew microbial community profiles (Yilmaz, Allgaier & Hugenholtz, 2010; Probst et al., 2015).

82   The recent development of the Nextera™ technology substantially speeds up Illumina library

83   creation and reduces input DNA requirements down to 1ng (Nextera-XT). In this approach DNA

84   is simultaneously fragmented and tagged ("tagmentation") using *in vitro* transposition (Syed,

85   Grunenwald & Caruccio, 2009; Caruccio, 2011). The resulting tagged fragments undergo a 12-

86   cycle PCR reaction to add sequencing adaptors and sample-specific barcodes, which facilitate

87   sample multiplexing. A number of attempts have been made to push the limits of Nextera library

88   creation into the sub-nanogram range including the creation of unvalidated libraries from 10pg of

89   human DNA (Adey et al., 2010) and validated libraries using as little as 20pg of *E. coli* and

90   mouse genomic DNA (Parkinson et al., 2012). The latter study found that this technique

91   provided deep coverage of the *E. coli* K-12 genome, but also increased the proportion of

92   duplicate reads and resulted in over-representation of low-GC regions. Most recently, the fidelity

93   of picogram-level libraries was assessed using a simple mock microbial community, which

94   found minimal impact of input DNA (down to 1 pg) on community composition estimates using

95   the Nextera-XT kit (Bowers et al., 2015). Here, we extend this approach using a more complex

96   mock community and environmental samples down to the femtogram input DNA range. We find

97   that read-mapping estimates of community composition fidelity are reproducible down to 100fg

98   and infer that variance in community structure between replicate 10 μl marine samples appears

99   to be primarily due to microscale biological differences.

100

101   **MATERIALS AND METHODS**

102   **Mock community construction**

103   Genomic DNA extracted from 40 bacterial and 14 archaeal taxa for which reference genomes are

104   available (54 isolate genomes), were combined to create a mock community (**Table S1**). Purified

105   genomic DNAs from 49 of the 54 mock community members were obtained from collaborators.

106   These DNAs were assessed via gel electrophoresis and quantified with qPCR (Shakya et al.,

107   2013). The amount of DNA and the genome size of each isolate were used to calculate their

108   expected relative abundance in the community. The organisms for which only low amounts of

109 gDNA were available were added in lower abundances to the final mix (**Table S1**). The final

110 DNA concentration of the mock community was 23.1 ng/µl, which was diluted appropriately for

111 low input library construction.

Marine sampling

113 Marine surface seawater samples were obtained from Blackwattle Bay in Sydney Harbor

114 (33°52'S, 151°11'E). Seawater was collected in a 10L sampling container and low volume

115 samples (1 ml, 100 µl, 10 µl) were pipetted individually from the 10L sample and snap frozen

116 directly at the sampling site. For the marine standard operating procedure (SOP), triplicate 10L

117 samples were collected and transported to the laboratory (~30 min travel time). Upon arrival, the

118 samples were pre-filtered through a 10µm filter (Millipore, MA) to remove large particles and

119 subsequently filtered through 0.2 µm Sterivex filters (Millipore, MA). All samples were kept at -

120 80°C until further processing. The entire sample preparation and analysis workflow is shown in

121 **Fig. S1**.

DNA extraction from seawater samples

123 DNA extraction from seawater samples was performed using the UltraClean® Tissue & Cells

124 DNA Isolation Kit following manufacturer's instructions. Minor modifications were made to

125 optimize the DNA extraction for marine samples and the low volume samples ($10^2 - 10^6$

126 cells/ml), and for consistency were also applied to the filtered marine SOP samples. Briefly, 1

127 mL (instead of 700 µL) Solution TD1 was added directly into the low volume seawater samples

128 (1mL, 100µL, 10µl; all in 1.5ml tubes) or directly into the Sterivex filter (10L filtered SOP).

129 Using a Vortex-Genie®, samples were lyzed and homogenized by vortexing the tubes and filters

130 respectively at maximum speed for 1 minute, without adding the recommended beads. Finally

131 20µL (instead of 50µl) of elution buffer was added and incubated at room temperature for 5 min

132 before centrifugation. For the low input libraries 1/4 of the DNA extraction volume (5 µl) was

133 used for library creation. Thereby, the amount of input DNA for library preparation was

134 quantified, using a Qubit-fluorometer (Invitrogen), for the SOP and the 1ml libraries, and was

135 estimated for the 100µl and 10µl samples based on the 1ml sample measurements. The number

136 of cells in the low volume samples was calculated based on an average DNA content of 1-10fg

137 per cell.

138 **DNA library preparation**

139 Libraries were prepared using the Nextera XT DNA Sample Preparation Kit (Illumina Inc., San

140 Diego, CA, USA). The standard protocol of the manufacturer was modified to optimize library

141 preparation from DNA input concentrations of less than 1ng (0.2ng/µL). The amplicon tagment

142 mix (ATM), which includes the enzyme used for tagmentation, was diluted 1 in 10 in nuclease

143 free water. For each sample, a 20 µL tagmentation reaction contained 10 µL TD buffer, 5 µL of

144 input DNA and 5 µL of the diluted ATM. Tagmentation reactions were incubated on a thermal

145 cycler at 55° for 5 min. Subsequently, tagmented DNA was amplified via a limited-cycle PCR

146 whereby the number of amplification cycles was increased from 12 to 20 cycles to ensure

147 sufficient library quantity for the downstream sequencing reaction. Amplified libraries were

148 purified with 1.6x Ampure XP beads and eluted in 20µL of re-suspension buffer. The quality of

149 the purified libraries was assessed using the High Sensitivity DNA kit on the Agilent 2100

150 Bioanalyzer. Successful libraries were quantified through qPCR using the KAPA Library

151 Quantification Kits, according to manufacturer's instructions, prior to pooling and sequencing.

152 The creation of each low input library was performed in triplicate, together with a negative

153 control containing no input DNA.

154 **DNA Sequencing**

155   All libraries were sequenced with an Illumina NextSeq500 platform 2x with 150bp High Output

156   v.1 run chemistry. The replicate SOP, 100pg, 10pg, 1pg, 100fg, negative control, and marine

157   sample libraries were pooled on an indexed shared sequencing run, resulting in 1/37 of a run or

158   ~3.2Gb per sample. The adapter trimmed fastq read files were deposited on the Microscale

159   Ocean webpage (https://microscaleocean.org/data/category/9-low-input-dna-libraries-peerj).

160   **Genome reference database**

161   A genome reference database was created by concatenating the fasta files of the 54 mock

162   community member genomes (**Table S1**), the *M. aerolatum* contaminant genome, the human

163   genome (release GRCh37), and the phiX 174 genome.

164   **Read mapping based mock community profiles**

165   Adapter trimmed sequences were aligned against the genome reference database using BWA

166   MEM 0.7.12 (Li, 2013) through BamM (http://ecogenomics.github.io/BamM/). To improve

167   stringency the seed length was increased to 25 bases in BWA MEM mode (--extras "mem:-k

168   25"). The resulting bam files were evaluated with samtools (Li et al., 2009), using samtools view

169   (http://www.htslib.org/) and a custom script counting the mapped reads per reference genome.

170   **Insert size**

171   The BamM generated bam files (see read mapping above) were randomly subsampled to 1

172   million aligned read pairs and the CIGAR string column 9 (TLEN, observed template length)

173   was extracted, using samtools view and the GNU coreutils command-line programs awk and

174   shuf. Trimmed mean (trim=0.01) and trimmed standard deviation (trim=0.01) were calculated

175   with the Rstudio package (https://www.rstudio.com/). The applied definition of the term "insert

176   size" used throughout this manuscript is the number of bases from the leftmost mapped base in

177   the first read to the rightmost mapped base in the second read.

178 **Read %GC content**

179 Raw FASTQ-format forward reads were converted to FASTA format and the %GC content

180 calculated with a custom perl script for each library replicate. The first 10,000 reads per replicate

181 were used to calculate the average %GC content.

182 **Read duplicates**

183 A custom python script (checkunique_v7.py) was used to estimate the percentage of read

184 duplicates using raw reads as input. The script loads a forward and a reverse read FASTQ file,

185 randomly selects a given number of read pairs, and concatenates the first 30 bp from the forward

186 and reverse reads into a 60 bp sequence. The 60 bp sequences from different read pairs are then

187 compared and the number of unique pairs is recorded. Read duplicates are defined as total

188 counted read pairs minus unique pairs. The script takes increments as optional arguments, and

189 performs subsampling (e.g. in 100,000 read increments) and subsequent counting of unique

190 (duplicate) read pairs per subsample, which allows plotting of a read duplicate rarefaction curve.

191 The cutoff of 30 bases per read and 100% match was chosen after initial trials showing that this

192 cutoff is comparable to read duplicate levels estimated by read mapping to reference genomes of

193 the mock data set (data not shown). For reference-based read mapping, raw reads were

194 subsampled with seqtk (https://github.com/lh3/seqtk) mapped against the reference genome

195 database with BamM and duplicate read pairs were removed with samtools rmdup (Li et al.,

196 2009), which defines a duplicate as a read with the exact same start and stop position as an

197 already mapped read, and compared to the same file before duplicate removal using samtools

198 flagstat.

199 **Taxonomic profiles**

200   16S rRNA gene-based taxonomic profiles of the mock community and seawater samples were

201   generated with GraftM (http://geronimp.github.io/graftM) using the 16S rRNA package

202   (4.06.bleeding_edge_2014_09_17_greengenes_97_otus.gpkg). The pipeline was designed to

203   identify reads encoding 16S rRNA genes based on HMMs and to assign taxonomic

204   classifications by comparing against a reference taxonomy. A detailed feature description, user

205   manual,   and   example   runs   are   available   on   the   GitHub   wiki

206   (https://github.com/geronimp/graftM/wiki). For the heat map, the GraftM output was manually

207   curated, whereby mitochondrial and chloroplast sequences were removed. Taxon counts were

208   trimmed (max >20), and analyzed with DESeq2, a method for differential analysis of count data

209   using shrinkage estimation for dispersions and fold changes (Love, Huber & Anders, 2014), in

210   the software environment R (www.r-project.org). The data were log transformed (rlog), and

211   displayed as a heat-map (pheatmap).

212   **Functional profiles**

213   Reads were searched against uniref100 (Suzek et al., 2007) [accessed 20151020] using

214   DIAMOND v0.7.12 (Buchfink, Xie & Huson, 2015) with the BLASTX option. The top hit of

215   each read (if above 1e-3) was mapped to KEGG Orthology (KO) IDs using the Uniprot ID

216   mapping files. Hits to each KO were summed to produce a count table. Correlations and

217   significance tests were performed with R (www.r-project.org) after applying and a cut-off > 500.

218   **Assembly and binning**

219   Reads were adapter trimmed and subsampled as follows. For the replicate assemblies, forward

220   and reverse reads were subsampled to 5 million reads each. For the combined assemblies, reads

221   from each replicate of a library were combined and then subsampled to 25 million reads.

222   Assemblies were performed with the CLC Genomics Workbench 8.0.2 (http://www.clcbio.com)

223 using default settings and a 1kb minimum contig size. Binning of population genomes was

224 performed with MetaBAT using default settings (--sensitive) as described previously (Kang et

225 al., 2015), and the resulting population genome bins were evaluated and screened with CheckM

226 (Parks et al., 2014).

227 **Sequence logo generation**

228 The mock community library forward reads were subsampled to 1000 reads using seqtk

229 (https://github.com/lh3/seqtk). Reads were trimmed to the first 10 bases with a custom Perl script

230 (trimFasta.pl), and the reads were submitted to weblogo (Crooks et al., 2004) to generate

231 sequence logos (Schneider & Stephens, 1990).

232 **Statistical analysis**

233 The software packages MYSTAT and SYSTAT (http://www.systat.com) and R (www.r-

234 project.org) were used for all statistical analyses. Datasets were analyzed by ANNOVA

235 (parametric) and Tukey's Significance Test, the non-parametric Kruskal-Wallis One-way

236 Analysis of Variance, or the Bonferroni probabilities (p-value) for correlations. Results from the

237 16S rRNA gene based community profiles and the functional profiles were used to calculate the

238 mean coefficient of variation, which is defined as the ratio of the standard deviation to the mean.

239 **Assembly and binning of mock community population genomes**

240 The ten population genomes were obtained by harvesting 10ml of culture medium and extracting

241 the DNA using the PowerSoil® DNA Isolation Kit (MO BIO Laboratories).   Genome

242 sequencing was performed on the Illumina NextSeq 500 platform using the Nextera library

243 protocol. Raw sequencing reads were adapter clipped and quality trimmed with trimmomatic

244 (http://www.usadellab.org/cms/?page=*trimmomatic)* 0.32 using the parameters "LEADING:3

245 TRAILING:3 SLIDINGWINDOW:4:15 CROP:10000 HEADCROP:0 MINLEN:50" with

246   Nextera adapter sequences. BBMerge (version BBMAP: bbmap_34.94;

247   https://sourceforge.net/projects/bbmap) was used to merge overlapping pairs of reads using

248   default parameters. Quality controlled paired reads were assembled with CLC Genomics Cell

249   assembler v4.4 using an estimated insert size of 30-500 bp. Quality controlled paired reads were

250   mapped to the assembled contigs using BamM v1.5.0 (http://ecogenomics.github.io/BamM/),

251   BWA 0.7.12 (Li, 2013) and samtools (http://www.htslib.org/) 0.1.19. The coverage of each

252   contig was determined with BamM 'parse' and the average coverage weighted by contig length

253   estimated. Raw reads were subsampled to provide an estimated 100X coverage using seqtk

254   'sample' (https://github.com/lh3/seqtk), quality controlled with trimmomatic/ BBMerge and

255   reassembled. Contigs less than 2kb were removed and genome quality assessed with CheckM

256   1.0.3 (Parks et al., 2014) using 'lineage_wf'. All genomes were assessed as being >95% complete

257   and <1% contaminated. The resulting ten population genome sequences were deposited in

258   NCBI-BioProject under the BioProject ID PRJNA324744.

259

260 **RESULTS**

261 ***Low input library and sequence data quality***

262 The Nextera-XT Standard Operating Procedure (SOP) is suitable for library preparation of 1ng

263 input DNA, which is equivalent to 10^5-10^6 microbial cells, assuming 1-10fg per cell (Button &

264 Robertson, 2001). We began by testing the Nextera-XT kit with 1pg input DNA from *E.coli* and

265 a mock microbial community, equivalent to 10^2-10^3 cells, a relevant range for localized

266 microenvironments found within natural marine systems. The mock community comprised 54

267 bacterial and archaeal isolates with sequenced genomes and GC contents ranging from 28 to

268 70%, which were pooled at relative abundances ranging from 0.04 to 25% (**Table S1**). For 1pg

269 input DNA, it was necessary to increase the limited cycle PCR from 12 to 20 cycles to ensure

270 sufficient tagmentation amplicon product for library preparation (Bowers et al., 2015) and

271 sequencing (**Fig. S2**). The amplicon tagment mix (ATM) used in the Nextera-XT kit includes a

272 transposase that fragments and tags the input DNA during the tagmentation process, and we

273 hypothesized that over-fragmentation of low input DNA could be avoided by diluting the

274 enzyme and/or decreasing the tagmentation reaction time. We found a direct correlation between

275 ATM dilution and insert size for both the pure and mixed templates, ranging from shorter than

276 desired fragments (<200bp) in the 1:5 dilutions to fragment sizes equal to the 1ng SOP in the

277 1:50 dilutions (~300bp; **Fig. 1 & Fig. S3**). Surprisingly, the 1pg undiluted ATM controls

278 produced insert sizes similar to the 1:10 dilutions (~240bp) in contrast to the anticipated over-

279 fragmentation observed in the 1:5 dilutions (see *Discussion*).

280

281 Next we investigated the percentage of read duplicates in each library, a known artifact of the

282 limited cycle PCR step (Kozarewa et al., 2009), which may compromise *de novo* assembly (Xu

283 et al., 2012). In general, the percentage of read duplicates was much higher in the 1pg libraries

284 than the 1ng SOP libraries; ~50% vs ~2% respectively at a sampling depth of 5 million read pairs

285 (**Fig. 1**). The percentage of read duplicates increased with increasing ATM dilution (**Fig. 1**), so

286 to achieve an insert size of >200bp (recommended by Illumina for Nextera-XT libraries) while

287 minimizing read duplicates, we proceeded with the 1:10 dilution. This ATM dilution gave

288 similar results to the undiluted samples in terms of insert size and read duplicates (**Fig. 1**), but

289 provides the potential to create multiple libraries from the same ATM starting volume.

290

291 We next tested a range of low input DNA concentrations (100pg to 100fg) using the mock

292 community and our modified protocol (20 cycles, 1:10 ATM dilution) prepared in triplicate, and

293 evaluated by comparison to 1ng SOP libraries. Library creation was reproducibly successful

294 down to 1pg as assessed by Bioanalyzer and qPCR profiles (**Fig. S4**). Two of the three initial

295 100fg libraries were successful, but only slightly above the Bioanalyzer detection limit and

296 showed a lag in qPCR amplifications of about two cycles compared to the 100pg library (**Fig.**

297 **S4**). Therefore, an additional three 100fg libraries were created, all of which progressed to

298 sequencing. To assess possible contamination we included two types of negative controls by

299 substituting ddH$_2$O for input DNA in the DNA extraction and library construction steps. We

300 selected the UltraClean DNA extraction kit for the DNA extraction control as this was the kit

301 used to extract marine samples (see *Application to environmental samples* below). Negative

302 controls showed DNA detectable via the Bioanalyzer and/or qPCR assays and were therefore

303 sequenced (**Fig. S5**). The sequencing runs were successful for all low input libraries (with the

304 exception of the one 100fg library), producing between 12 and 22 million adaptor-trimmed

305 reads, which slightly less than the 1ng SOP libraries (30.5 mil; SD ±6.9 mil), with a trend

306 towards decreasing sequence yield with decreasing library input DNA (**Fig. 2**). The no input

307   DNA negative controls resulted in a higher yield for the DNA extraction plus library kit control

308   (13.8 mil; SD ±3.8 mil) compared to low read numbers for the library only kit control (5.7 mil;

309   SD ±3.7 mil, **Fig. 2**).

310

311   The average insert size of the low input mock community libraries, as estimated by read mapping

312   to the reference genomes, was significantly smaller ($p < 0.01$) than the SOP library, with

313   307±115 bp observed in the SOP libraries vs 204 - 257 bp in the low input libraries (**Fig. 2;**

314   **Table S2**). Although the average read GC content was not significantly different ($p < 0.05$)

315   between the 1ng SOP and low input libraries, an appreciable drop was recorded for the 100fg

316   libraries (**Fig. 2; Fig. S6**). Read duplicates increased with decreasing input DNA from an

317   average of 1.2% for the 1ng SOP up to 78.8% for 100fg, at a sampling depth of five million read

318   pairs (**Fig. 2 & Fig. S7a**). This is consistent with the higher levels of read duplicates observed in

319   the initial 1pg *E.coli* and mock community libraries (**Fig. 1**). We noted that read duplicates are

320   library-specific as combining reads from the 1pg library replicates sampled to the same depth

321   reduced the proportion of duplicates and improved assembly statistics (**Fig.S7b, Fig. S8**). This

322   library specificity suggests that successfully tagmented DNA is a random subset of input DNA,

323   which in the case of low input samples increases the proportion of duplicates by further limiting

324   the template for limited cycle PCR. We also explored the relationship between percentage read

325   duplicates and the limited cycle PCR step, whereby we raised the number of cycles from 12 to

326   20. We predicted that lower cycles should produce a lower proportion of read duplicates.

327   Additional replicated low input libraries were prepared using from 12 to 20 cycles, according to

328   Bioanalyzer detectability thresholds, for a given amount of input DNA (**Fig. S9**). In contrast to

329   expectations, the percentage of read duplicates did not change appreciably as a function of PCR

330   cycle for a given DNA input concentration (**Fig. S10**), suggesting that they are mostly created

331   within the first 12 cycles.

332

333   ***Sequencing contaminants***

334   To determine the identity of possible sequencing contaminants, we assembled the reads from all

335   negative control libraries. A substantially complete *Methylobacterium* genome (87.9% according

336   to CheckM; (Parks et al., 2014) was assembled which had 100% identity over 758 bp to the 16S

337   rRNA gene of *M. aerolatum* strain 92a-8 (Weon et al., 2008). Members of the genus

338   *Methylobacterium* are recognized reagent contaminants (Salter et al., 2014). This genome was

339   used to aid in identifying contaminant reads in the mock community libraries (**Fig. 2**). We

340   aligned the sequence reads of the mock community and negative controls against the reference

341   database containing all 54 microbial genomes, the *M. aerolatum* contaminant genome, the

342   human genome to detect possible operator contamination and the phiX 174 genome, which is

343   used as an internal run quality control during Illumina sequencing. For the 1ng SOP and 100pg,

344   10pg, and 1pg low input libraries, 95.6 to 97.7% of reads mapped to the microbial mock

345   community and small percentages mapped to the human reference (0.1 to 1.9%) or were

346   unmapped (2.1 to 3.9%**; Fig. 2**). We found more variable results for the five 100fg libraries: two

347   libraries produced similar results to the higher input libraries (rep1 & rep2 in **Fig. 2**) and the

348   other three had high human contamination (30.4 to 45.0% of reads; rep3, rep4 & rep5 in **Fig. 2**).

349   Reads from the negative control libraries were mostly human contamination (11.0 to 60.5%), *M.*

350   *aerolatum* (0.4 to 65.5%), or unmapped (16.0 - 48.6%). The three UltraClean kit negative control

351   libraries accounted for 98.5% of all reads mapping to *M. aerolatum* suggesting that this DNA

352   extraction kit is the primary source of this contaminant. A small proportion of the negative

353   control reads (0.3 to 0.9%) mapped to the mock community, which we attribute to cross sample

354   contamination with mock community libraries due to false index pairings of the multiplex

355   sequencing runs (Kircher, Sawyer & Meyer, 2012). A closer examination of the unmapped

356   negative control reads (16.0 - 48.6%) revealed the presence of mostly Firmicutes (*Bacilli and*

357   *Clostridia*), but also *Proteobacteria*, *Actinobacteria* and *Elusimicrobia* (**Fig. S11**). We did not

358   detect phiX 174 contamination (*data not shown*).

359

360   ***Fidelity of community composition***

361   We evaluated community composition fidelity by comparing the relative abundance of the 54

362   mock community members between the 1ng SOP library and the low input libraries, based on

363   read mapping to the mock community genome database. Community composition of low-input

364   libraries averaged across replicates (excluding library 100fg_S12) was strongly ($R^2 \geq 0.94$) and

365   significantly ($p < 0.001$) correlated to standard input libraries (**Fig.3**; **Table S3a**), indicating that

366   reducing input DNA resulted in minimal representational bias. This significant correlation was

367   upheld even when the five most abundant community members were excluded, although the

368   100fg libraries began to show slightly higher variance (**Fig.3**; **Table S3b**). Since reference

369   genome sets are usually unavailable for environmental samples, we also assessed community

370   composition using 16S rRNA-based taxonomic profiling and functional profile analysis using the

371   KEGG database. The 16S rRNA-based taxonomic profiles matched the genome-based read

372   mapping analysis with all libraries being strongly ($R^2 \geq 0.98$) and significantly ($p < 0.001$)

373   correlated to each other (**Fig. S12**). Functional profiles between SOP and low input libraries

374   down to 1 pg were also strongly ($R^2 \geq 0.99$) and significantly ($p < 0.001$) correlated, with lower

375   but still significant correlations to the 100 fg libraries ($R^2 \geq 0.91$; $p < 0.01$; **Fig. S13**). We

376   therefore suggest that community composition can be reliably assessed using low input libraries

377   down to 100fg despite higher proportions of read duplicates (**Fig. S7a**) provided contaminants

378   are accounted for by including negative controls. To determine if the consistently reduced

379   correlation between the 100fg and SOP libraries was a systematic effect of the observed

380   difference in average read %GC (**Fig. 2**), we investigated the relative abundance of community

381   members with high, median and low genomic GC content. No significant differences (p > 0.45)

382   were observed between the relative abundances of reads aligning to high, medium or low GC

383   genomes among all input libraries compared to the SOP (**Table S4**). This suggests that there is

384   no substantial bias against high GC organisms among our mock community members with

385   decreasing input DNA, although there was a slight trend in this direction (**Fig. S14**). An analysis

386   of the regions immediately flanking the transposase insertion sites (first 10 bases of each read)

387   indicated a slight preference for insertion into AT-richer regions relative to the mean GC content

388   (**Fig. S15**). However, this preference was consistent between the different input DNA

389   concentrations and should have had no net effect on average read %GC content.

390

391   ***Assembly and binning***

392   We normalized each replicate dataset to five million read pairs and assembled them using CLC

393   Genomics workbench (*see Materials and Methods*). Assembly quality metrics - maximum contig

394   length, total assembly size and number of contigs - were found to deteriorate with decreasing

395   input DNA (**Fig. 4**). The increasing percentage of read duplicates with decreasing input DNA

396   (**Fig.S7a**) may be primarily responsible for this observed drop in assembly statistics. However,

397   when duplicates were removed, assembly statistics did not improve relative to the corresponding

398   assemblies using all reads (**Fig.4**) suggesting that read duplicates *per se* did not hamper the

399   assembly. Instead, we noted a strong correlation between decreasing percentage of unique reads

400   and loss of assembly performance for the mock community datasets (**Fig.S16**). Despite this loss

401   in performance, datasets down to 1 pg still produced thousands of contigs >1 kb (up to a

402   maximum of 260 kb) from five million read pairs suitable for population genome binning. To

403   assess maximum binning potential for the mock community datasets, we combined and co-

404   assembled replicates (50 million reads) except for the 100 fg libraries which did not meet this

405   sequence read threshold. A drop in assembly performance was also noted with decreasing input

406   DNA for the larger datasets and the removal of read duplicates had no effect on assembly

407   statistics (**Fig. S17**). However, nine moderately complete (>50%) to near (>90%) complete

408   genomes with low contamination (<10%) were still recovered from the 1 pg library assembly,

409   compared to 24 for the SOP library (**Table 1**).

410

411   ***Application to environmental samples***

412   Based on the observation that the mock community profiles were consistent between the SOP

413   and low-input DNA library protocols, we applied our approach to marine surface water samples

414   (Sydney Harbor, Australia). In order to obtain sufficient DNA for the Nextera XT SOP, we

415   filtered 10 liters of surface water obtaining >100 ng bulk DNA. Small unfiltered volumes,

416   obtained from a 10L surface water sample, comprising 1ml, 100µl and 10µl, were used to create

417   low-input DNA libraries. DNA was extracted from replicated samples using the UltraClean kit,

418   applying a minor modification to accommodate marine samples and the smaller unfiltered

419   volumes (*see Materials and Methods*). The bulk DNA from the 10L filtered sample was used for

420   the SOP libraries and diluted to create low input DNA libraries equivalent to the low sample

421   volume libraries. Library creation was successful for all samples and yielded between 14.8 to

422   30.2 million adaptor-trimmed reads for the low input samples compared to 24.6 ±1.2 million

423   reads for the equimolar pooled 10L filtered SOP samples (**Fig.5**). To detect possible

424   contamination, we aligned the reads against the reference genome database (*see Materials and*

425   *Methods).* No *phiX* contamination was found, but varying degrees of human and *M. aerolatum*

426   contamination were detected, with the general trend of increasing contamination with decreasing

427   input DNA (**Fig.5**) as seen in the mock community samples (**Fig. 2**). Substantial contamination

428   was noted in four of the low volume libraries ($\leq 27.3\%$ reads) highlighting the importance of i)

429   running negative control libraries to identify contaminants, some of which may be kit or reagent

430   specific, and ii) screening low-input DNA shotgun datasets for identified and known

431   contaminants.

432

433   After removing contaminant reads, we used 16S rRNA-based taxonomic and KO-based

434   functional profile analyses to evaluate community reproducibility between the different libraries

435   in the absence of reference genomes for these samples. The inferred taxonomic profiles were

436   consistent with those expected within surface marine water samples including a dominance of

437   populations belonging to the *Pelagibacteraceae, Flavobacteriaceae,* and *Synechococcaceae*

438   families (**Fig. 6; Fig. S18**). Profiles averaged across replicates of both the filter dilution and low

439   volume marine communities were strongly correlated to the SOP libraries down to 5 pg and 10

440   µl respectively (**Fig. 7, Fig. S19, Fig. S20**). Replicate profiles were also highly correlated with

441   the exception of the 10 µl libraries (**Fig. 8, Fig. S18**). This exceeded the expected technical

442   variation observed in the corresponding 5 pg filter dilutions according to comparisons based on

443   the mean coefficient of variation (**Fig. 9**).

444

445   Assembly and binning were also assessed for the marine samples. We repeated the approach

446   used for the mock community and found that the assembly and binning of the 10L filtered

447   dilution libraries (5 and 50 pg) produced similar results to the 10L filtered SOP (**Figs. S21;**

448 **Table S5)**. However, the low volume libraries representing between ~3 pg (10 µl) and ~300 pg

449 (1 ml) input DNA had poorer assembly metrics (**Fig. S21**) and did not produce any population

450 genomes with >50% completeness as compared to the 10L filtered datasets which produced five

451 such genomes (**Table S5**).

452

453 **Discussion**

454 The typical requirement of microgram quantities of DNA for metagenome library creation often

455 necessitates the use of bulk samples comprising millions to billions of individual microbial cells

456 in macroscale quantities (mls, grams). Low input DNA library creation protocols provide the

457 opportunity to dissect microbial communities into microscale volumes (µl, mg), containing

458 hundreds to a few thousands of cells comprising picogram quantities of DNA. Such protocols

459 will also aid efforts to characterize microbial communities in very low biomass environments.

460 Currently, the most widely used commercially available low input DNA kit is Nextera XT

461 (Illumina Inc., San Diego, CA, USA), which uses transposase insertion to fragment and tag the

462 DNA with sequencing adapters. Using *E. coli* genomic DNA, a mock community of 54

463 phylogenetically diverse bacterial and archaeal isolates, and environmental samples obtained

464 from coastal seawater, we evaluated a modified Nextera XT low input DNA library protocol for

465 sequence yield and contamination, community composition fidelity, and assembly and

466 population genome binning.

467

468 The manufacturer recommended Nextera XT library insert size range is 200 bp to 1 kb. We

469 hypothesized that lowering the amount of input DNA may decrease average insert size due to a

470 higher ratio of transposase to DNA resulting in more frequent tagmentations (cutting and

471 attaching of adapters) per unit length of DNA. Indeed, a recent study by Bowers et al. (2015)

472   found that insert size did indeed decrease as a function of input DNA for the Nextera XT kit. In

473   theory, the insert size would only be restricted to a lower boundary of ~20 nt due to steric

474   interference between adjacent transposase complexes, which require at least a 19 bp binding site

475   (Steiniger et al., 2006). The issue of over-tagmentation may be addressable by reducing reaction

476   time or diluting the transposase. We observed a direct correlation between amplicon tagment mix

477   (ATM) dilutions (1:5 down to 1:50) and insert size, with the largest insert sizes at the highest

478   dilutions (**Fig. 1**). It was surprising, however, that the 1pg undiluted ATM controls produced

479   larger than expected insert sizes, similar to the 1:10 dilutions (**Fig. 1**). A possible explanation is

480   that the ATM includes a factor that inhibits the hyperactive Tn*5* transposase and subsequently a

481   dilution of the mix would also result in a dilution of the inhibitor, allowing an increase in enzyme

482   efficiency. Inhibition of the Tn*5* transposase activity is known for *E. coli*, in which an inhibitor

483   of the transposition protein (a Tn*5* transposase variant which lacks the N-terminal 55 amino acids

484   and thus does not possess DNA-binding activity) forms a complex with Tn*5* and interferes with

485   transposition (de la Cruz et al., 1993). Based on these findings we prepared low input libraries

486   using a 1:10 ATM dilution.

487

488   Average GC content is an often-reported metric in regard to Illumina libraries because of

489   observed biases of GC-poor or GC-rich genomes and regions (Chen et al., 2013). We observed a

490   slight decrease in the average read GC content with lower amounts of input DNA in the mock

491   community libraries (**Fig. 2**). This was associated with a slight trend towards over-representation

492   and under-representation of low and high GC organisms, respectively, with lower input DNA

493   (**Fig. S14**). A slightly reduced coverage in high but also low GC regions was observed using the

494   Nextera protocol (50ng SOP) for virus genomes and the GC coverage trend was attributed to an

495   amplification bias that occurred during the Nextera limited-cycle PCR (Marine et al., 2011). The

496   opposite trend was observed by Bowers et al. (2015), which they attributed to organism specific

497   differences. Another possible contributor to GC bias could be transposase insertion, as suggested

498   previously (Lamble et al., 2013). An analysis of the regions immediately flanking the transposase

499   insertion sites revealed a GC content ~2% lower than the average read GC content, despite the

500   known preference of Tn*5* to insert at a guanine (Goryshin et al., 1998) (**Fig. S15**). However, this

501   drop in GC content was consistent between the different input DNA concentrations suggesting

502   that the limited cycle PCR is the main source of the observed slight shift in average GC content.

503   This conclusion is in agreement with previous results who point to the PCR as the most

504   important cause for GC-content bias during library preparation (Risso et al., 2011).

505

506   The largest difference in sequencing metrics between the Nextera XT 1 ng SOP and low input

507   libraries, however, was the proportion of read duplicates, which increased dramatically with

508   decreasing input DNA (**Fig. 2**). This is consistent with previous studies that found high levels of

509   read duplicates for low input DNA libraries, whereby duplication levels of over 60% have been

510   observed for 50pg libraries at a sampling depth of 5 million reads, and up to 74% duplication

511   occurring in 1pg libraries with a sampling depth of 15 million reads (Chafee, Maignien &

512   Simmons, 2015; Bowers et al., 2015). Read duplicate rates were positively correlated with ATM

513   dilutions (**Fig. 1**), which is presumably the result of fewer transposase insertions producing fewer

514   DNA fragments and subsequently less unique templates, increasing the likelihood of read

515   duplicates during PCR. The number of limited cycle PCR cycles (12 to 20) had no consistent

516   effect on the percentage of read duplicates (**Fig. S10**), thus at a given ATM dilution the amount

517   of input DNA is the primary factor determining the fraction of read duplicates (**Fig. S7**). This is

518   consistent with previous findings that PCR duplicates arise from a lack of DNA complexity

519   (unique templates) due to low levels or quality of input DNA (Smith et al., 2014). While it is also

520   possible that read duplicates are "community derived" and created by the fragmentation of two

521   identical DNA molecules that are tagmented at exactly the same location, the probability of such

522   an event in metagenomic samples is extremely low (Gomez-Alvarez, Teal & Schmidt, 2009). By

523   subsampling reads we could show that read duplicates are library specific and their proportion is

524   reduced by combining library replicates (**Fig. S7b**). This is likely due to the random loss of up to

525   90% of the initial DNA during library creation (Parkinson et al., 2012; Zhou et al., 2014), which

526   in the case of Nextera XT libraries would be due to tagmentations that do not produce limited

527   cycle PCR-amplifiable products. Thus a repeat of this random process from the same starting

528   material will result in reads covering different regions of the input DNA. Reducing read

529   duplicates would appear to be a useful endeavor as they have been reported to bias coverage-

530   based quality validation and hamper assemblies (Xu et al., 2012; Ekblom & Wolf, 2014).

531   However, we found that the observed differences in percentage duplicates, GC content, and

532   insert size (**Fig. 2**) had no impact on either our taxonomic or functional-based community

533   profiles as evidenced by significant correlations between the 1ng SOP and the low input libraries

534   of the mock community (**Fig.3, Fig. S12 & S13).** Furthermore, read duplicates *per se* did not

535   affect assemblies (**Fig.4**), which are instead limited by the number of unique reads in a sequence

536   dataset (**Fig.S16**). Therefore combining library replicates improves assembly outcomes because

537   it increases the number of unique reads (**Fig. S8)**. Based on these results, we suggest that low

538   input DNA libraries are reproducible down to 100 fg using the minimally modified Nextera XT

539   protocol of a 1:10 ATM dilution and 20 cycles of limited cycle PCR, regardless of the observed

540   increase in duplicate reads. However, $\geq$ 1pg libraries gave results more consistent with the SOP

541   than the 100fg libraries in terms of insert size, read duplicates, GC content, and community

542   composition fidelity **(Fig. 2)**. We therefore recommend 1pg (~10^2 - 10^3 cells) as the lowest DNA

543   input amount using the modified protocol.

544

545   An important consideration when preparing low input DNA libraries is contamination. Small

546   amounts of contaminating DNA are common in nucleic acid extraction kits and other laboratory

547   reagents as indicated by no input DNA negative controls (Salter et al., 2014). Documented

548   contaminants include representatives of the Proteobacteria, Actinobacteria, Firmicutes,

549   Bacteroidetes, fungi, green plants and animals (Lusk, 2014; Strong et al., 2014). Such

550   contamination is typically negligible in the context of standard input DNA libraries, but becomes

551   an issue with decreasing input DNA, for example, with low biomass samples (Lusk, 2014; Salter

552   et al., 2014). Therefore, we ran negative controls for both DNA extraction and library

553   preparation in parallel with the low input DNA samples, where we substituted input DNA with

554   ultrapure water. We recovered a near complete genome closely related to *Methylobacterium*

555   *aerolatum* from the negative control libraries, which our data indicates can be mostly attributed

556   to the UltraClean DNA extraction kit **(Fig. 2)**. The genus *Methylobacterium* is a known

557   contaminant introduced during sample preparation, possibly through molecular biology grade

558   water, PCR reagents, or DNA extraction kits (Salter et al., 2014). One way to reduce potential

559   reagent contamination is UV irradiation, which has been successfully used to decontaminate

560   reagents for single-cell genomics (Woyke et al., 2011). However, since *M. aerolatum*, a strictly

561   aerobic alphaproteobacterium, was originally isolated from air samples (Weon et al., 2008),

562   aerial contamination cannot be entirely ruled out. The other major contaminant in the negative

563   control libraries was human DNA, presumably attributable to the operator, though reagents may

564   also be a source. Mining the remaining fraction of unmapped negative control reads, we found

565   contaminants from a wide diversity of bacterial taxa (**Fig. S11**) matching previous reports (Lusk,

566   2014; Salter et al., 2014). Notably the family *Staphylococcaceae* was dominant in the library

567   preparation negative controls with up to 100% relative abundance (**Fig. S11**). Since this family

568   includes the genus *Staphylococcus*, a known member of the human skin and mucus microbiome

569   (Otto, 2010), we attribute its presence to operator introduced contamination. We therefore

570   recommend running ultrapure water controls with and without the DNA isolation procedure to

571   identify reagent-specific contaminants, and to include them in contaminant screens of low input

572   DNA sequencing libraries.

573

574   Having addressed contamination issues and established the preservation of community

575   composition for the low input DNA libraries in the mock community, we applied the modified

576   Nextera XT protocol to marine samples, one of the most intensively studied ecosystems using

577   metagenomics (Gilbert & Dupont, 2011; Reddy et al., 2015). We extracted DNA from samples

578   having a range of volumes, representing standard (10 L filtered control) down to microscale

579   (10µl) samples. We also prepared libraries from dilutions of the 10L control to match the DNA

580   inputs of the low volume samples. Ten microliter samples were the lowest volumes used as they

581   were estimated to comprise ~3pg of DNA, which is above our recommended 1pg library input

582   DNA threshold (**Fig.5**). Community profiles based on 16S rRNA sequences identified in the

583   samples were consistent with those of previously described marine surface waters (**Fig. 6**). As

584   was observed with the mock community, both the taxonomic and functional profiles of the low

585   volume and filtered dilution control samples were strongly correlated to the 10L filtered SOP

586   libraries when comparing replicate averages (**Fig. 7, Fig. S19 & S20**). However, we noticed a

587   high degree of variance between replicates of the 10 µl libraries (**Fig. 8, Fig. S18)**, above the

588   expected technical variation observed in the corresponding 5 pg filtered dilution replicates of the

589 taxonomic profiles (**Fig. 9**). We attribute this elevated variation to microscale biological

590 differences between replicates of the same sample volume. Indications of such microscale

591 patchiness have been reported previously for bacterioplankton with significant variation in

592 bacterial species richness amongst microliter samples using denaturing gradient gel

593 electrophoresis (Long & Azam, 2001). Therefore, our results constitute the first metagenomic

594 data of microscale heterogeneity in marine surface waters, which has mostly been overlooked to

595 date (Azam & Malfatti, 2007; Stocker & Seymour, 2012). Indeed, the observed

596 microheterogeneity may be an underestimate of *in-situ* conditions as the low volume samples

597 were taken from a 10L subsample in which localized mixing may have occurred.

598

599 *De novo* assembly and binning are important bioinformatic steps in metagenomic workflows

600 (Leung et al., 2013), which we assessed in relation to low input DNA libraries using the mock

601 community and marine datasets. It is important to consider that 1pg of DNA only comprises ~1.2

602 Gb of potentially sequenceable template (cf. ~1.2 Tb template for the 1 ng SOP), therefore

603 expectations of assembly and binning should be appropriately calibrated. Both steps are known

604 to be dependent on community composition with the general rule of thumb that increasingly

605 complex communities produce lower quality assembly and binning results (Mavromatis et al.,

606 2007; Dröge & McHardy, 2012). Therefore, we did not directly compare the mock to marine

607 results, instead focusing on within community type comparisons. As expected, both assembly

608 and binning deteriorated with decreasing input DNA, although substantial assemblies (kb range)

609 were still obtained down to 1pg libraries for both community types (**Fig.4; Fig. S21**), which are

610 suitable for e.g. gene neighborhood analyses. Approximately a third of the population bins

611 (>50% completeness) obtained for the mock community SOP were still recoverable from the 1pg

612 datasets (**Table 1**), in contrast to a recent report using a simpler mock community in which no

613   bins of this quality were recovered from 1pg Nextera XT libraries (Bowers et al., 2015). The low

614   volume marine datasets yielded no bins, although only five bins were obtained from the marine

615   SOP in total (**Table S5**), consistent with the higher complexity of this community, again

616   emphasizing the case-by-case nature of assembly and binning.

617

Conclusions

618

619   We demonstrate that it is possible to successfully prepare and sequence low input metagenome

620   libraries down to 100fg of DNA using a slightly modified version of the Nextera XT protocol,

621   with the important caveat that negative controls are included to detect possible reagent and other

622   contaminants. Community composition was highly reproducible down to 1pg despite a

623   pronounced increase in the proportion of read duplicates with decreasing input DNA indicating

624   that duplicate formation is random. Assembly and binning are both compromised by lowering

625   input DNA due to decreasing sequenceable template and associated number of unique reads.

626   However, both assembly and binning are still possible in the pg range depending on community

627   complexity. Applying the approach to surface marine waters, we found evidence for microscale

628   community composition heterogeneity in 10 µl volumes, demonstrating the utility of applying

629   metagenomics at spatial scales relevant to microorganisms.

630

Acknowledgements

631

632   We thank Mircea Podar and Stuart Denman (CSIRO) for generously providing DNA for

633   construction of the mock community, David Wood and Pierre-Alain Chaumeil for providing

634   bioinformatics support, Yun Kit Yeoh for contributing reference population genomes bins, Paul

635 Evans for assistance with the mock community, and the ACE team (http://ecogenomic.org) for

636 stimulating discussions.

637

638 **References**

639 Adey A., Morrison HG., Asan., Xun X., Kitzman JO., Turner EH., Stackhouse B., MacKenzie AP., Caruccio

640   NC., Zhang X., Shendure J. 2010. Rapid, low-input, low-bias construction of shotgun fragment

641   libraries by high-density in vitro transposition. *Genome Biology* 11:R119. DOI: 10.1186/gb-2010-

642   11-12-r119.

643 Azam F. 1998. Microbial Control of Oceanic Carbon Flux: The Plot Thickens. *Science* 280:694–696. DOI:

644   10.1126/science.280.5364.694.

645 Azam F., Malfatti F. 2007. Microbial structuring of marine ecosystems. *Nature Reviews Microbiology*

646   5:782–791. DOI: 10.1038/nrmicro1747.

647 Bowers RM., Clum A., Tice H., Lim J., Singh K., Ciobanu D., Ngan CY., Cheng J-F., Tringe SG., Woyke T.

648   2015. Impact of library preparation protocols and template quantity on the metagenomic

649   reconstruction of a mock microbial community. *BMC Genomics* 16:856. DOI: 10.1186/s12864-

650   015-2063-6.

651 Buchfink B., Xie C., Huson DH. 2015. Fast and sensitive protein alignment using DIAMOND. *Nature*

652   *Methods* 12:59–60. DOI: 10.1038/nmeth.3176.

653 Button DK., Robertson BR. 2001. Determination of DNA Content of Aquatic Bacteria by Flow Cytometry.

654   *Applied and Environmental Microbiology* 67:1636–1645. DOI: 10.1128/AEM.67.4.1636-

655   1645.2001.

656 Caruccio N. 2011. Preparation of Next-Generation Sequencing Libraries Using Nextera™ Technology:

657   Simultaneous DNA Fragmentation and Adaptor Tagging by In Vitro Transposition - Springer. In:

658   Kwon YM, Ricke SC eds. *High-Throughput Next Generation Sequencing*. Methods in Molecular

659   Biology. Humana Press,.

660 Chafee M., Maignien L., Simmons SL. 2015. The effects of variable sample biomass on comparative

661   metagenomics. *Environmental Microbiology*:n/a-n/a. DOI: 10.1111/1462-2920.12668.

662 Chen Y-C., Liu T., Yu C-H., Chiang T-Y., Hwang C-C. 2013. Effects of GC Bias in Next-Generation-

663   Sequencing Data on De Novo Genome Assembly. *PLoS ONE* 8. DOI:

664   10.1371/journal.pone.0062856.

665 Clingenpeel S., Clum A., Schwientek P., Rinke C., Woyke T. 2015. Reconstructing each cell's genome

666   within complex microbial communities—dream or reality? *Microbial Physiology and Metabolism*

667   5:771. DOI: 10.3389/fmicb.2014.00771.

668 Crooks GE., Hon G., Chandonia J-M., Brenner SE. 2004. WebLogo: a sequence logo generator. *Genome*

669   *Research* 14:1188–1190. DOI: 10.1101/gr.849004.

670 de la Cruz NB., Weinreich MD., Wiegand TW., Krebs MP., Reznikoff WS. 1993. Characterization of the

671   Tn5 transposase and inhibitor proteins: a model for the inhibition of transposition. *Journal of*

672   *Bacteriology* 175:6932–6938.

673 Dröge J., McHardy AC. 2012. Taxonomic binning of metagenome samples generated by next-generation

674   sequencing technologies. *Briefings in Bioinformatics* 13:646–655. DOI: 10.1093/bib/bbs031.

675 Duhaime MB., Deng L., Poulos BT., Sullivan MB. 2012. Towards quantitative metagenomics of wild

676   viruses and other ultra-low concentration DNA samples: a rigorous assessment and optimization

677   of the linker amplification method. *Environmental Microbiology* 14:2526–2537. DOI:

678   10.1111/j.1462-2920.2012.02791.x.

679 Ekblom R., Wolf JBW. 2014. A field guide to whole-genome sequencing, assembly and annotation.

680   *Evolutionary Applications* 7:1026–1042. DOI: 10.1111/eva.12178.

681   Gilbert JA., Dupont CL. 2011. Microbial Metagenomics: Beyond the Genome. *Annual Review of Marine*

682   *Science* 3:347–371. DOI: 10.1146/annurev-marine-120709-142811.

683   Gomez-Alvarez V., Teal TK., Schmidt TM. 2009. Systematic artifacts in metagenomes from complex

684   microbial communities. *The ISME Journal* 3:1314–1317. DOI: 10.1038/ismej.2009.72.

685   Goryshin IY., Miller JA., Kil YV., Lanzov VA., Reznikoff WS. 1998. Tn5/IS50 target recognition. *Proceedings*

686   *of the National Academy of Sciences of the United States of America* 95:10716–10721.

687   Kallmeyer J., Pockalny R., Adhikari RR., Smith DC., D'Hondt S. 2012. Global distribution of microbial

688   abundance and biomass in subseafloor sediment. *Proceedings of the National Academy of*

689   *Sciences of the United States of America* 109:16213–16216. DOI: 10.1073/pnas.1203849109.

690   Kang DD., Froula J., Egan R., Wang Z. 2015. MetaBAT, an efficient tool for accurately reconstructing

691   single genomes from complex microbial communities. *PeerJ* 3:e1165. DOI: 10.7717/peerj.1165.

692   Kircher M., Sawyer S., Meyer M. 2012. Double indexing overcomes inaccuracies in multiplex sequencing

693   on the Illumina platform. *Nucleic Acids Research* 40:e3. DOI: 10.1093/nar/gkr771.

694   Kozarewa I., Ning Z., Quail MA., Sanders MJ., Berriman M., Turner DJ. 2009. Amplification-free Illumina

695   sequencing-library preparation facilitates improved mapping and assembly of GC-biased

696   genomes. *Nature methods* 6:291–295. DOI: 10.1038/nmeth.1311.

697   Lamble S., Batty E., Attar M., Buck D., Bowden R., Lunter G., Crook D., El-Fahmawi B., Piazza P. 2013.

698   Improved workflows for high throughput library preparation using the transposome-based

699   nextera system. *BMC Biotechnology* 13:104. DOI: 10.1186/1472-6750-13-104.

700   Leung HCM., Wang Y., Yiu SM., Chin FYL. 2013. Next-Generation Sequencing on Metagenomic Data:

701   Assembly and Binning. In: Nelson KE ed. *Encyclopedia of Metagenomics*. Springer New York, 1–7.

702   Li H., Handsaker B., Wysoker A., Fennell T., Ruan J., Homer N., Marth G., Abecasis G., Durbin R.,

703   Subgroup 1000 Genome Project Data Processing. 2009. The Sequence Alignment/Map format

704   and SAMtools. *Bioinformatics* 25:2078–2079. DOI: 10.1093/bioinformatics/btp352.

705   Li H. 2013. Aligning sequence reads, clone sequences and assembly contigs with BWA-MEM.

706   *arXiv:1303.3997 [q-bio]*.

707   Long R., Azam F. 2001. Microscale patchiness of bacterioplankton assemblage richness in seawater.

708   *Aquatic Microbial Ecology* 26:103–113.

709   Love MI., Huber W., Anders S. 2014. Moderated estimation of fold change and dispersion for RNA-seq

710   data with DESeq2. *Genome Biology* 15:1–21. DOI: 10.1186/s13059-014-0550-8.

711   Lusk RW. 2014. Diverse and Widespread Contamination Evident in the Unmapped Depths of High

712   Throughput Sequencing Data. *PLoS ONE* 9:e110808. DOI: 10.1371/journal.pone.0110808.

713   Marine R., Polson SW., Ravel J., Hatfull G., Russell D., Sullivan M., Syed F., Dumas M., Wommack KE.

714   2011. Evaluation of a Transposase Protocol for Rapid Generation of Shotgun High-Throughput

715   Sequencing Libraries from Nanogram Quantities of DNA. *Applied and Environmental*

716   *Microbiology* 77:8071–8079. DOI: 10.1128/AEM.05610-11.

717   Mavromatis K., Ivanova N., Barry K., Shapiro H., Goltsman E., McHardy AC., Rigoutsos I., Salamov A.,

718   Korzeniewski F., Land M., Lapidus A., Grigoriev I., Richardson P., Hugenholtz P., Kyrpides NC.

719   2007. Use of simulated data sets to evaluate the fidelity of metagenomic processing methods.

720   *Nature Methods* 4:495–500. DOI: 10.1038/nmeth1043.

721   Otto M. 2010. Staphylococcus colonization of the skin and antimicrobial peptides. *Expert review of*

722   *dermatology* 5:183–195. DOI: 10.1586/edm.10.6.

723   Paerl HW., Pinckney JL. 1996. A mini-review of microbial consortia: Their roles in aquatic production and

724   biogeochemical cycling. *Microbial Ecology* 31:225–247. DOI: 10.1007/BF00171569.

725   Parkinson NJ., Maslau S., Ferneyhough B., Zhang G., Gregory L., Buck D., Ragoussis J., Ponting CP.,

726   Fischer MD. 2012. Preparation of high-quality next-generation sequencing libraries from

727   picogram quantities of target DNA. *Genome Research* 22:125–133. DOI: 10.1101/gr.124016.111.

728 Parks DH., Imelfort M., Skennerton CT., Hugenholtz P., Tyson GW. 2014. CheckM: assessing the quality

729   of microbial genomes recovered from isolates, single cells, and metagenomes. *PeerJ PrePrints*

730   2:e554v1. DOI: 10.7287/peerj.preprints.554v1.

731 Probst AJ., Weinmaier T., DeSantis TZ., Santo Domingo JW., Ashbolt N. 2015. New Perspectives on

732   Microbial Community Distortion after Whole-Genome Amplification. *PLoS ONE* 10. DOI:

733   10.1371/journal.pone.0124158.

734 Probst AJ., Auerbach AK., Moissl-Eichinger C. 2013. Archaea on Human Skin. *PLoS ONE* 8. DOI:

735   10.1371/journal.pone.0065388.

736 Raghunathan A., Ferguson HR., Bornarth CJ., Song W., Driscoll M., Lasken RS. 2005. Genomic DNA

737   Amplification from a Single Bacterium. *Appl. Environ. Microbiol.* 71:3342–3347. DOI:

738   10.1128/AEM.71.6.3342-3347.2005.

739 Rappe MS., Giovannoni SJ. 2003. THE UNCULTURED MICROBIAL MAJORITY. *Annual Review of*

740   *Microbiology* 57:369–394. DOI: 10.1146/annurev.micro.57.030502.090759.

741 Reddy TBK., Thomas AD., Stamatis D., Bertsch J., Isbandi M., Jansson J., Mallajosyula J., Pagani I., Lobos

742   EA., Kyrpides NC. 2015. The Genomes OnLine Database (GOLD) v.5: a metadata management

743   system based on a four level (meta)genome project classification. *Nucleic Acids Research*

744   43:D1099–D1106. DOI: 10.1093/nar/gku950.

745 Risso D., Schwartz K., Sherlock G., Dudoit S. 2011. GC-Content Normalization for RNA-Seq Data. *BMC*

746   *Bioinformatics* 12:480. DOI: 10.1186/1471-2105-12-480.

747 Salter SJ., Cox MJ., Turek EM., Calus ST., Cookson WO., Moffatt MF., Turner P., Parkhill J., Loman NJ.,

748   Walker AW. 2014. Reagent and laboratory contamination can critically impact sequence-based

749   microbiome analyses. *BMC Biology* 12:87. DOI: 10.1186/s12915-014-0087-z.

750 Schneider TD., Stephens RM. 1990. Sequence logos: a new way to display consensus sequences. *Nucleic*

751   *Acids Research* 18:6097–6100. DOI: 10.1093/nar/18.20.6097.

752   Shakya M., Quince C., Campbell JH., Yang ZK., Schadt CW., Podar M. 2013. Comparative metagenomic

753   and rRNA microbial diversity characterization using archaeal and bacterial synthetic

754   communities. *Environmental Microbiology* 15:1882–1899. DOI: 10.1111/1462-2920.12086.

755   Smith EN., Jepsen K., Khosroheidari M., Rassenti LZ., D'Antonio M., Ghia EM., Carson DA., Jamieson CH.,

756   Kipps TJ., Frazer KA. 2014. Biased estimates of clonal evolution and subclonal heterogeneity can

757   arise from PCR duplicates in deep sequencing experiments. *Genome Biology* 15. DOI:

758   10.1186/s13059-014-0420-4.

759   Solonenko SA., Ignacio-Espinoza JC., Alberti A., Cruaud C., Hallam S., Konstantinidis K., Tyson G., Wincker

760   P., Sullivan MB. 2013. Sequencing platform and library preparation choices impact viral

761   metagenomes. *BMC Genomics* 14:320. DOI: 10.1186/1471-2164-14-320.

762   Steiniger M., Adams CD., Marko JF., Reznikoff WS. 2006. Defining characteristics of Tn5 Transposase

763   non-specific DNA binding. *Nucleic Acids Research* 34:2820–2832. DOI: 10.1093/nar/gkl179.

764   Stocker R. 2012. Marine Microbes See a Sea of Gradients. *Science* 338:628–633. DOI:

765   10.1126/science.1208929.

766   Stocker R., Seymour JR. 2012. Ecology and Physics of Bacterial Chemotaxis in the Ocean. *Microbiology*

767   *and Molecular Biology Reviews : MMBR* 76:792–812. DOI: 10.1128/MMBR.00029-12.

768   Strong MJ., Xu G., Morici L., Splinter Bon-Durant S., Baddoo M., Lin Z., Fewell C., Taylor CM., Flemington

769   EK. 2014. Microbial Contamination in Next Generation Sequencing: Implications for Sequence-

770   Based Analysis of Clinical Samples. *PLoS Pathog* 10:e1004437. DOI:

771   10.1371/journal.ppat.1004437.

772   Syed F., Grunenwald H., Caruccio N. 2009. Next-generation sequencing library preparation:

773   simultaneous fragmentation and tagging using in vitro transposition. *Nature Methods* 6. DOI:

774   10.1038/nmeth.f.272.

775   Vaishampayan P., Probst AJ., La Duc MT., Bargoma E., Benardini JN., Andersen GL., Venkateswaran K.

776   2013. New perspectives on viable microbial communities in low-biomass cleanroom

777   environments. *The ISME Journal* 7:312–324. DOI: 10.1038/ismej.2012.114.

778   Weon H-Y., Kim B-Y., Joa J-H., Son J-A., Song M-H., Kwon S-W., Go S-J., Yoon S-H. 2008.

779   Methylobacterium iners sp. nov. and Methylobacterium aerolatum sp. nov., isolated from air

780   samples in Korea. *International Journal of Systematic and Evolutionary Microbiology* 58:93–96.

781   DOI: 10.1099/ijs.0.65047-0.

782   Woyke T., Sczyrba A., Lee J., Rinke C., Tighe D., Clingenpeel S., Malmstrom R., Stepanauskas R., Cheng J-

783   F. 2011. Decontamination of MDA Reagents for Single Cell Whole Genome Amplification. *PLoS*

784   *ONE* 6:e26161. DOI: 10.1371/journal.pone.0026161.

785   Xu H., Luo X., Qian J., Pang X., Song J., Qian G., Chen J., Chen S. 2012. FastUniq: A Fast De Novo

786   Duplicates Removal Tool for Paired Short Reads. *PLoS ONE* 7. DOI:

787   10.1371/journal.pone.0052249.

788   Yilmaz S., Allgaier M., Hugenholtz P. 2010. Multiple displacement amplification compromises

789   quantitative analysis of metagenomes. *Nat Meth* 7:943–944. DOI: 10.1038/nmeth1210-943.

790   Zhou W., Chen T., Zhao H., Eterovic AK., Meric-Bernstam F., Mills GB., Chen K. 2014. Bias from removing

791   read duplication in ultra-deep sequencing experiments. *Bioinformatics* 30:1073–1080. DOI:

792   10.1093/bioinformatics/btt771.

793

Dilution series evaluation for low input DNA libraries.

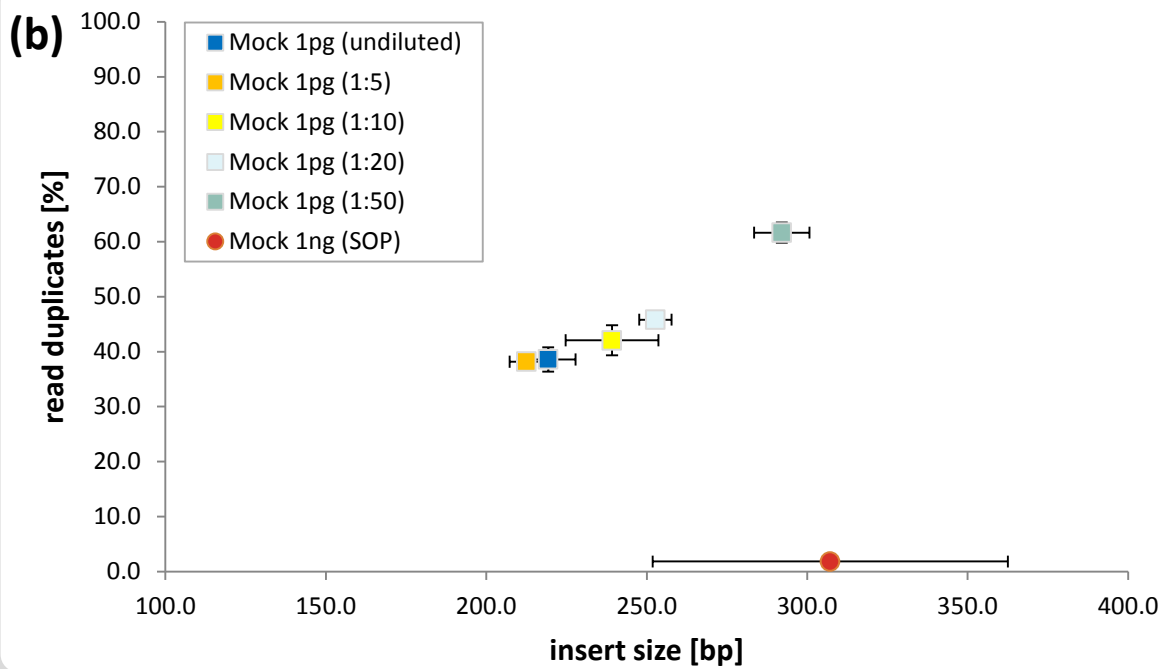
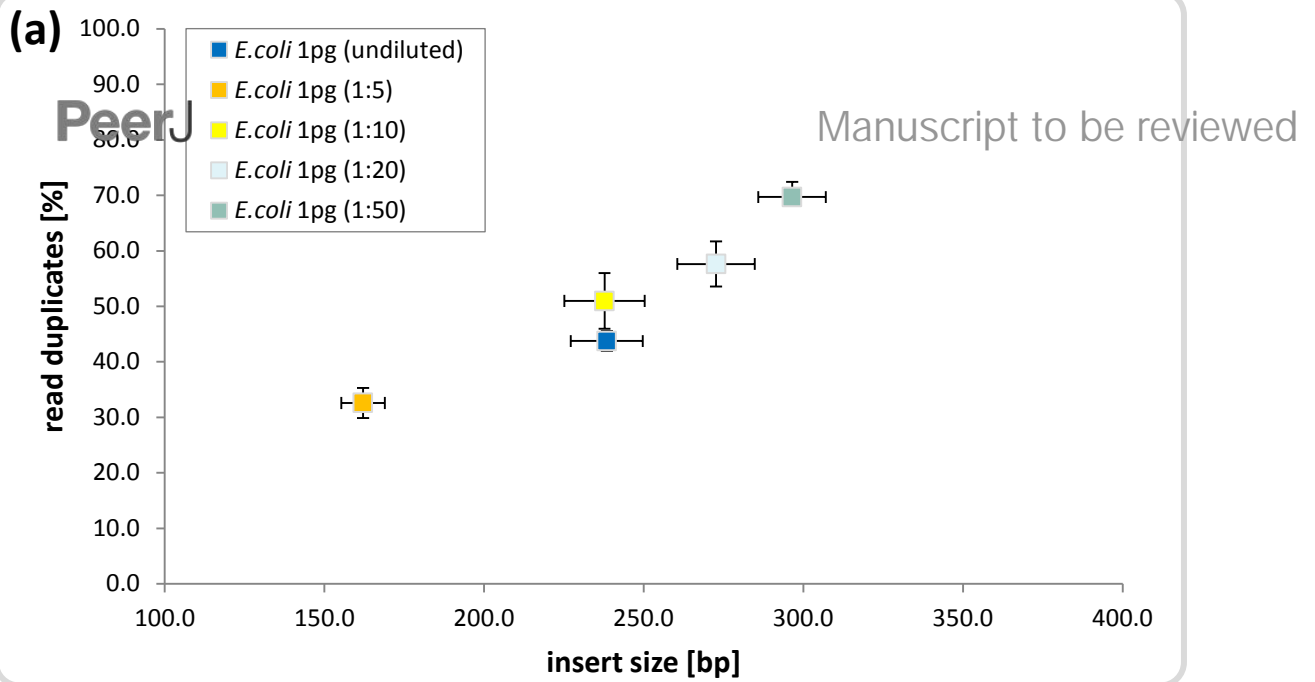
Figure 1 | Dilution series evaluation for low input DNA libraries. Dilutions down to 1:50 of amplicon tagment mix (ATM) for low input DNA libraries of 1pg DNA are shown in comparison to the 1ng SOP. The trimmed mean insert size (length of DNA fragments in bases without adaptors, determined via read mapping) is plotted against the relative number of read duplicates. Libraries were created with 1pg of (a) E. coli DNA, (b) Mock community DNA. Reads were subsampled to 5 million read pairs. Note that E.coli was not sequenced with the 1ng SOP.

Yield and quality assessment of low input libraries

Figure 2 | Yield and quality assessment of low input libraries. The bar graph shows the absolute number of reads for all replicates of the 1ng SOP (left), the low input libraries (middle) and the negative controls (right, grey background). Negative controls are comprised of the library prep kit control (NegLib) and the DNA extraction kit + library prep kit control (NegExt), see Methods for details. Reads are colour coded based on the reference they aligned to, including the bacterial and archaeal mock community (green) and the human genome (blue). The remaining reads are shown as unmapped (orange) or mapped against the contaminant Methylobaterium aerolatum (red). The calculated cell number range (~ no. cells) is based on the amount of input DNA and an estimated 1-10fg DNA per microbial cell. The sequence yield is provided as million reads (read yield). The average insert size (insert size), the average percent GC content (%GC), and the average number of read duplicates (% duplicates) was calculated as a mean of all replicates. The bar above the figure indicates when the standard protocol (SOP) or our modified protocol was used for library creation. The bar below the figure provides the average expected reads per sample, based on a NextSeq500 2x 150bp High Output v. 1 run with 1/37 sequence allocation per library. Sample replicate numbers are given in parenthesis.

	SOP	Modified Protocol					NegLib	NegExt
	1ng SOP	**100pg**	**10pg**	**1pg**	**100fg**		**NegLib**	**NegExt**
~ no. cells	$10^5 - 10^6$	$10^4 - 10^5$	$10^3 - 10^4$	$10^2 - 10^3$	$10^1 - 10^2$		-	-
read yield	30.5 ±6.9	19.4 ±2.0	17.8 ±1.2	18.4 ±2.6	14.3 ±2.0		5.9 ±3.8	13.9 ±3.9
insert size	307 ±55	257 ±08	254 ±22	219 ±29	204 ±30		197 ±119	176 ±11
%GC	55.5 ±0.13	55.3 ±1.36	53.2 ±0.41	53.5 ±0.65	50.3 ±3.26		45.8 ±7.41	56.2 ±5.66
%duplicates	1.2 ±0.1	5.3 ±0.4	21.0 ±1.2	52.1 ±3.5	78.8 ±4.8			

Legend:
- **unmapped**
- *Methylobacterium aerolatum*
- **human**
- **mock**

read Yield [million reads]

Expected reads per sample ~ 21 million/ sequence allocation 1/37th

Mock community profile comparisons

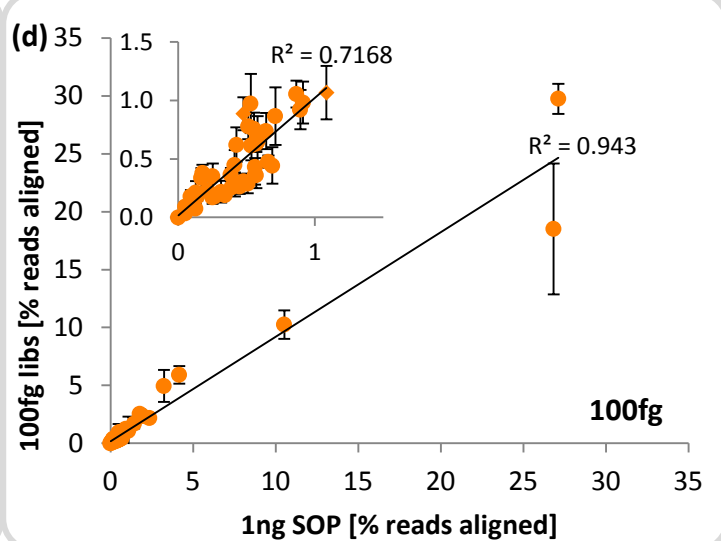
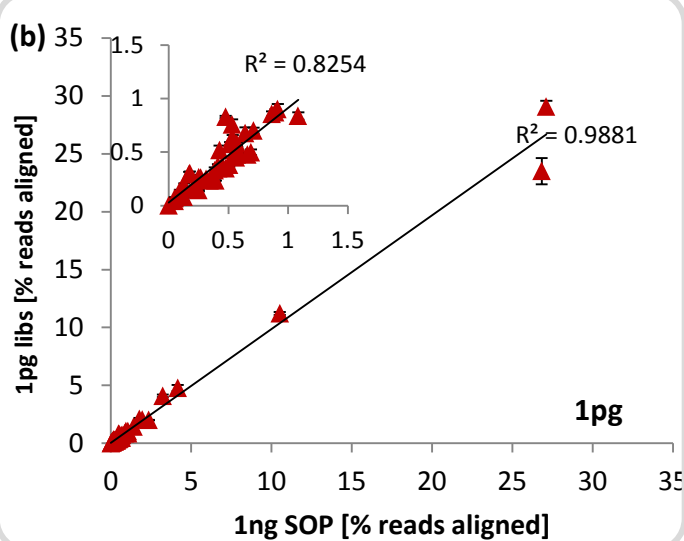
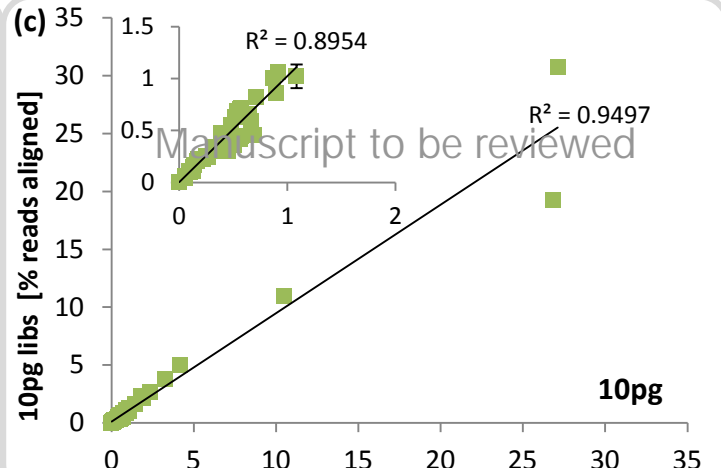
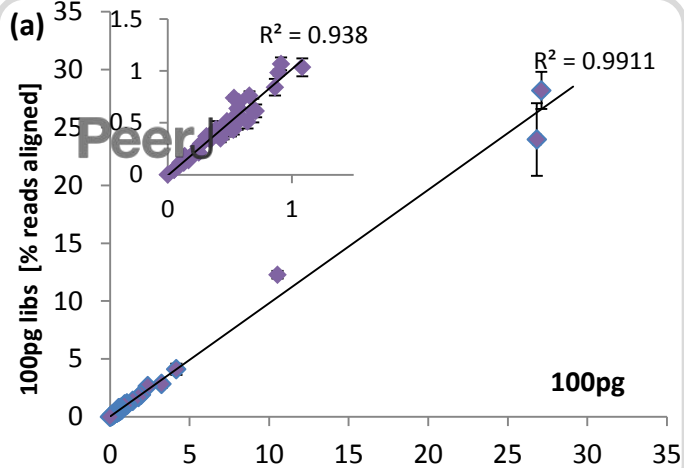
Figure 3 | Mock community profile comparisons. Correlation between the 1ng SOP libraries (X-axes) and the low input DNA libraries (100pg, 10pg, 1pg, 100fg; y-axes). Shown is the mean relative abundance of the 54 mock community members, based on reads aligned to the respective reference genomes.Inserts: show a subset of the relative abundances excluding the five most dominant organisms of the mock community. The mean standard deviation for each library is provided as error bars. The 100fg libraries include four replicates (1, 2, 4, 5) out of five, omitting replicate 3 which was highly contaminated.

Mock community assembly statistics

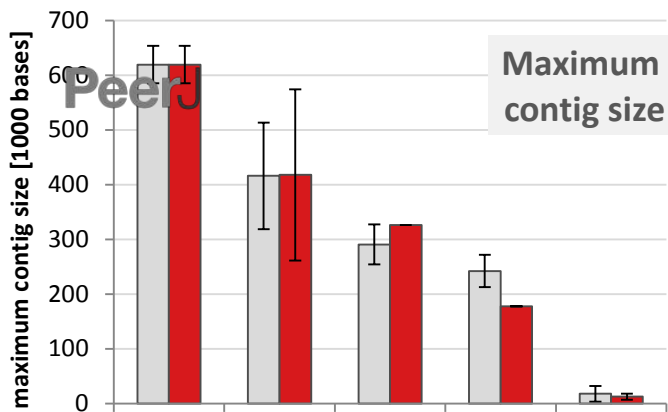
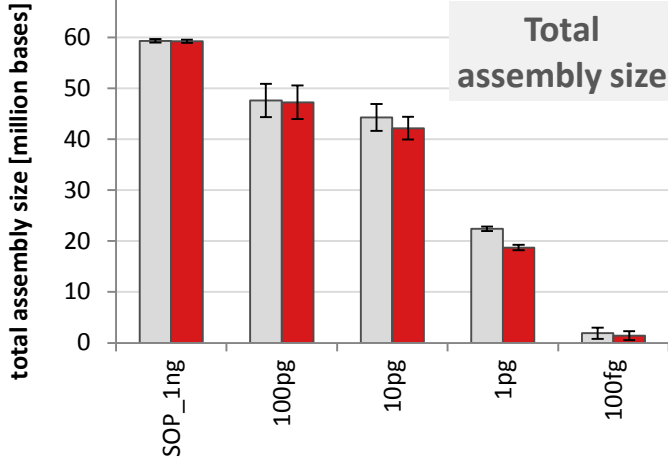
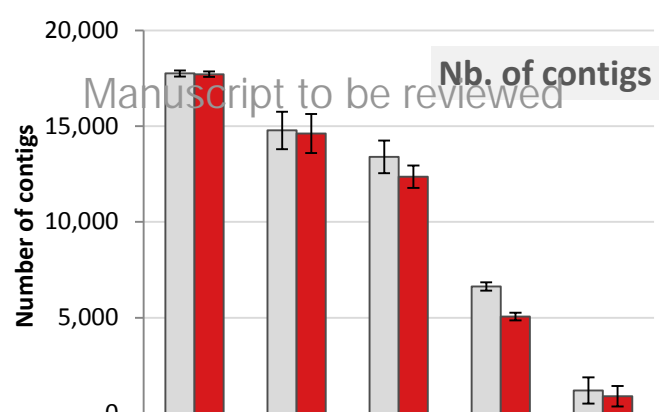
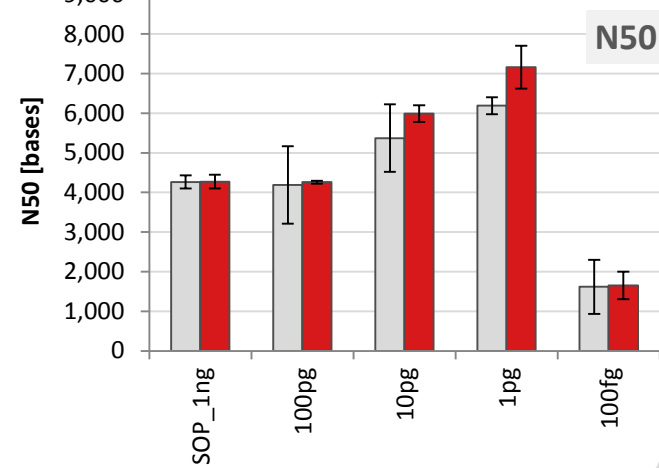
Figure 4 | Mock community assembly statistics. (a) Maximum contig size, (b) total assembly size, (c) number of contigs, and (d) N50 of the SOP and low input mock community libraries. Read files were subsample to 5 million read pairs. Gray bars show assemblies of all reads, red bars show assemblies after read duplicates were removed. Only contigs ≥ 1kb were included in the analysis. All values are given as mean and standard deviation. ? 3~? ���U

Yield and quality assessment of marine samples

Figure 5 | Yield and quality assessment of marine samples. Reads are color coded based on the reference they aligned to, including the known contaminant Methylobacterium aerolatum (red) and the human genome (blue). The remaining reads are shown as unmapped (orange). The amount of DNA extracted with the modified extraction protocol is given as total DNA in 20μl elution buffer (DNA extract). Number of cells (~no. cells) was calculated based on an average DNA content of 1-10fg per cell. The amount of input DNA for library preparation was measured for the SOP and the 1ml libraries, and was estimated for the 100μl and 10μl samples based on the 1ml sample measurements. The bar above the figure indicates when the standard protocol (SOP) or our modified protocol was used to create the libraries. All libraries were sequenced at an allocation of 1/37 of an Illumina NextSeq500 2x 150bp High Output v. 1 run.

	10L SOP	Low volume samples			10L dilutions	
	SOP	**1ml**	**100µl**	**10µl**	**50pg**	**5pg**
DNA extract	116ng	1.25ng	120pg	12pg	-	-
~ no. cells	10^7-10^8	10^5-10^6	10^4-10^5	10^3-10^4	10^4-10^5	10^3-10^4
DNA input	1ng	~300pg	~ 30pg	~3pg	50pg	5pg

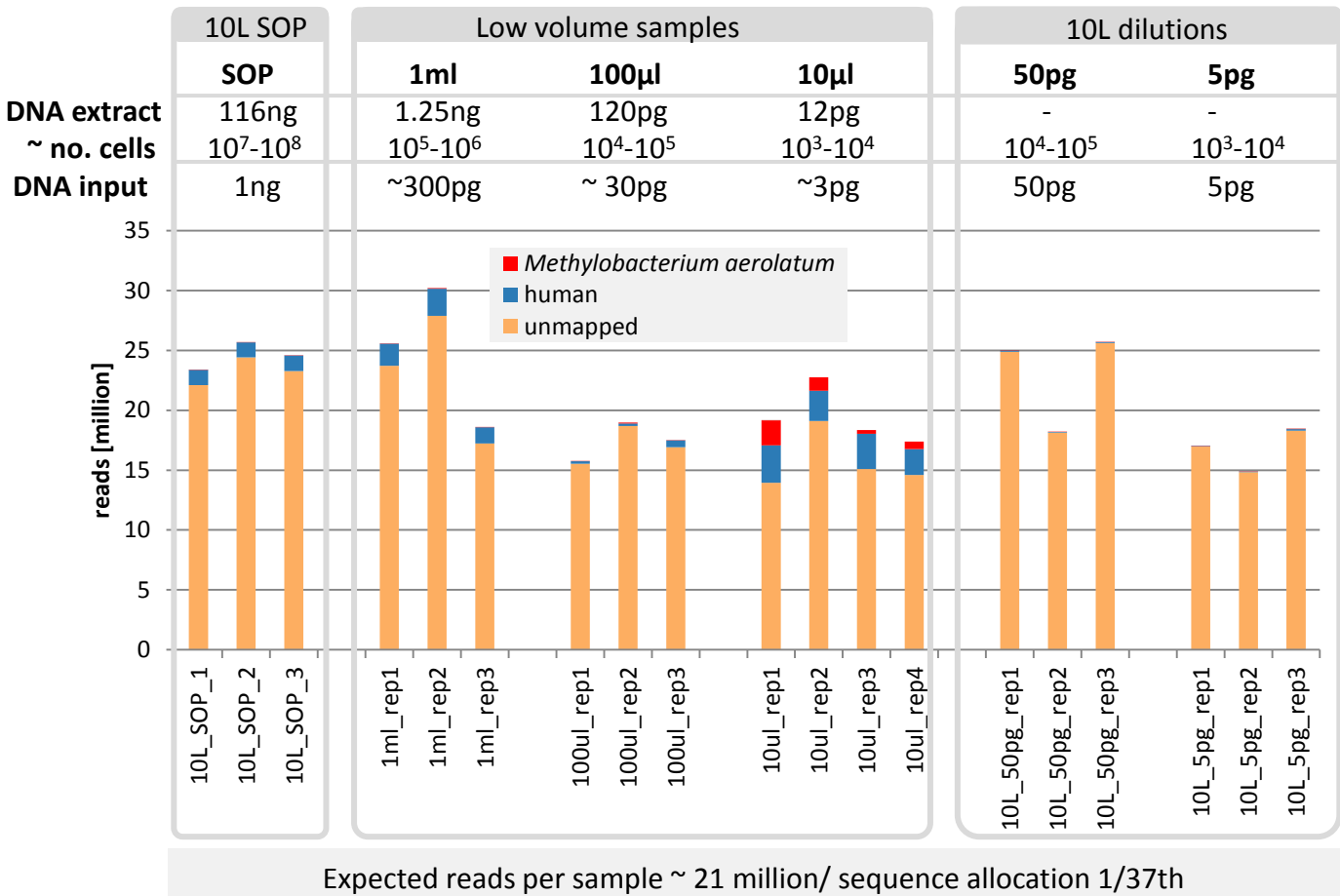
Expected reads per sample ~ 21 million/ sequence allocation 1/37th

Abundance profiles of the marine microbial samples

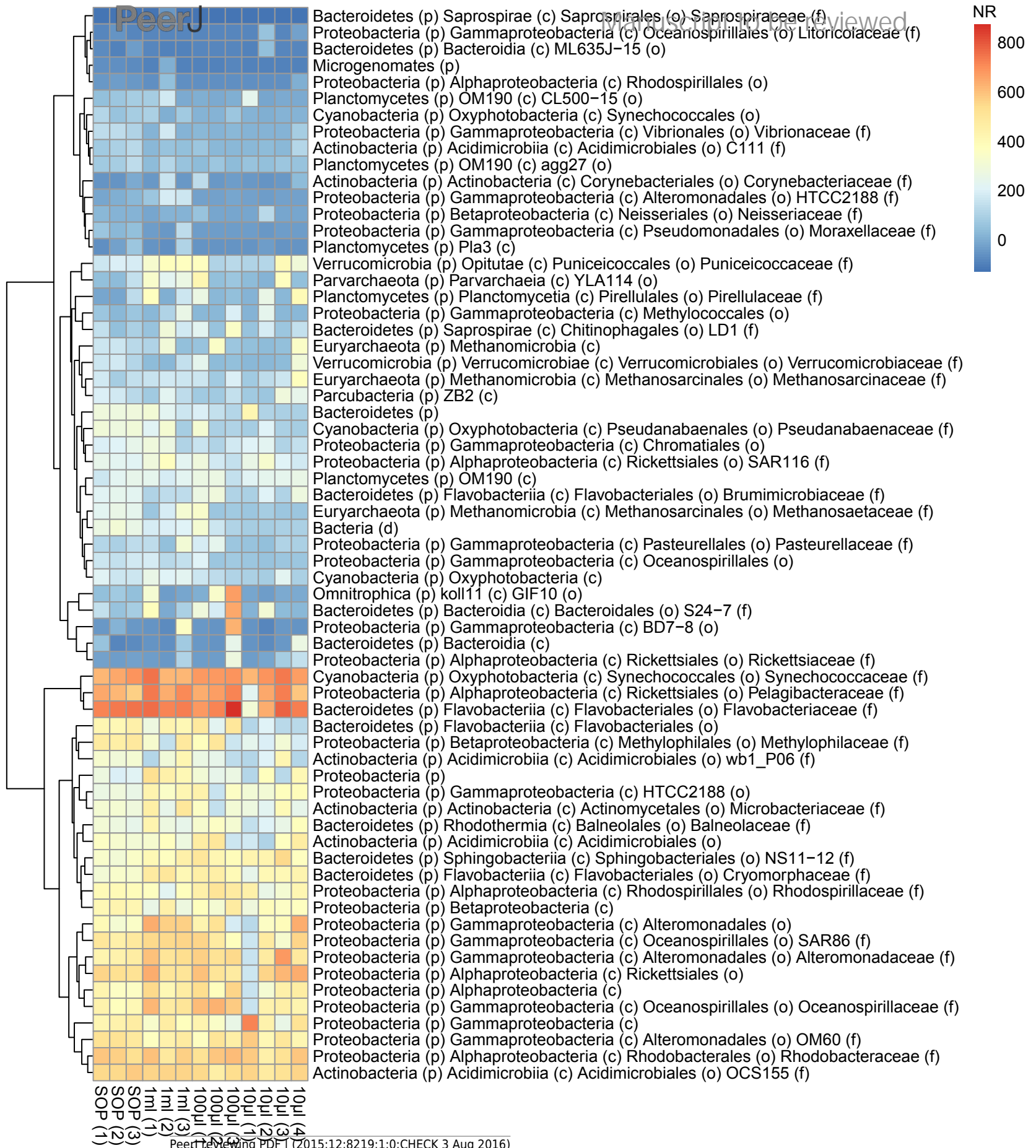
Figure 6 | Abundance profiles of the marine microbial samples. Bacterial OTUs were assigned based on 16S rRNA gene sequence detection of shotgun sequencing reads (graftM; see Methods). The normalized abundance is shown after square root transformation for all OTUs above the abundance threshold, resulting in a normalized read count (NR) from 0 to 800. The taxonomic assignment for each OTU is provided down to the family level if available, otherwise the best available taxonomic rank is given.

Marine communities profile correlations

Figure 7 | Marine communities profile correlations. Correlations coefficients are shown for the marine 10L SOP, the 10L filtered dilution, and the low input DNA libraries. The upper right panel shows the 16S rRNA based taxonomic profile correlations, and the lower left panel the KO-based functional profile correlations. The Pearson correlation coefficient is colour coded from zero (white) to 1 (dark blue).

16S rRNA gene taxonomic profile

KO-based functional profile

Profile analyses of marine sample replication

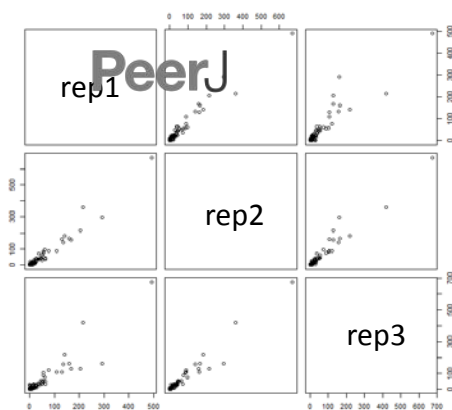
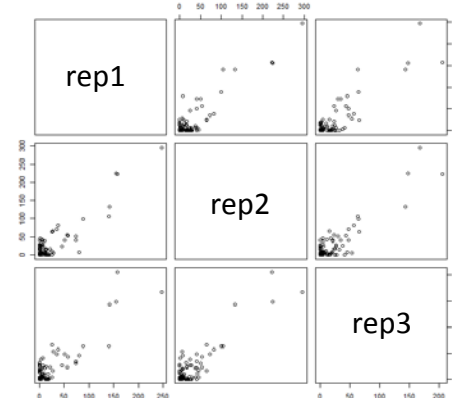
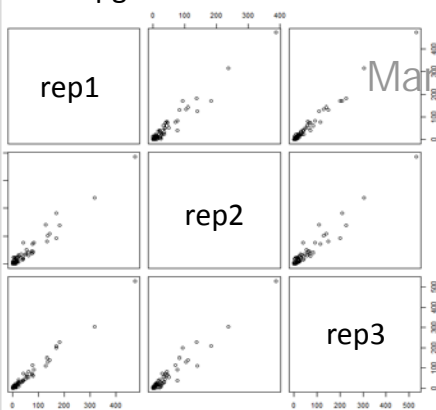
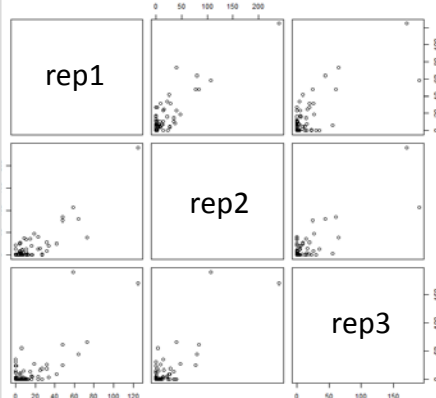
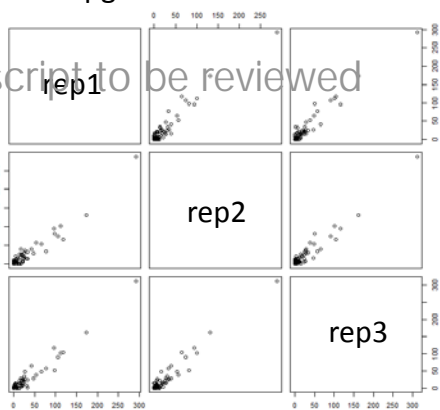
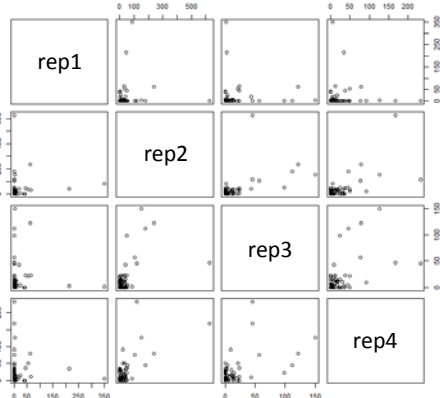
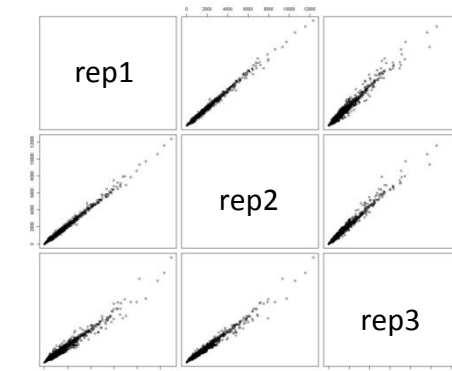
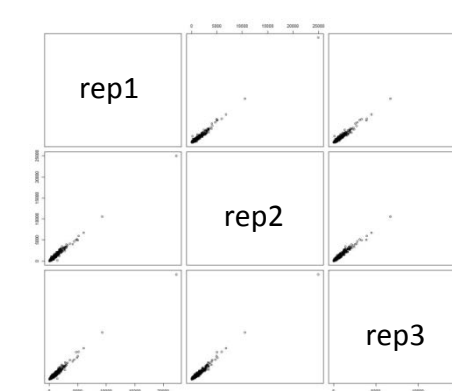
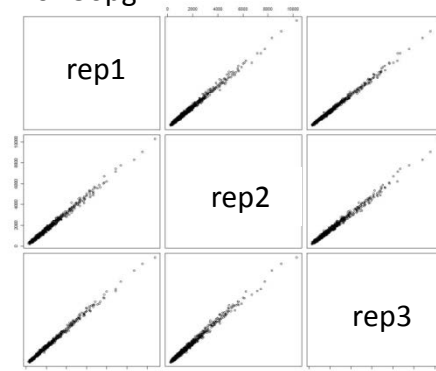
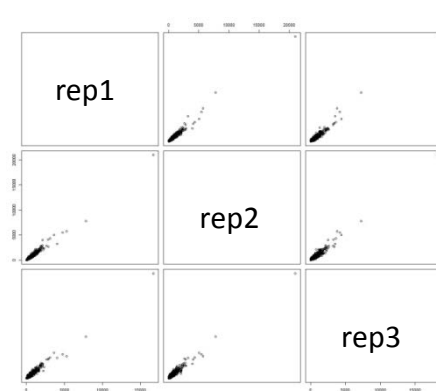
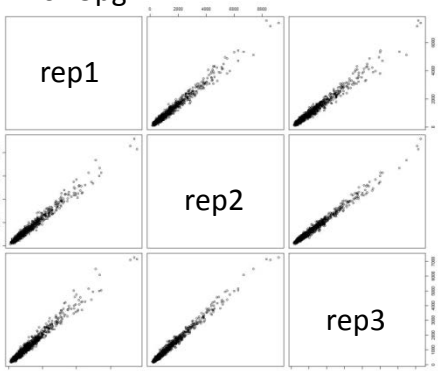
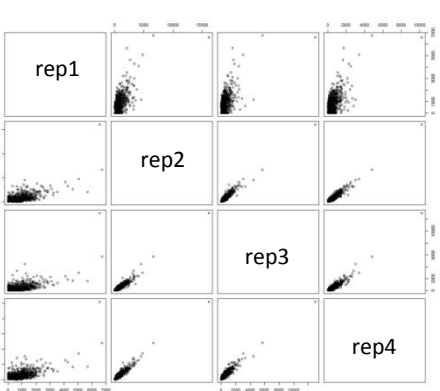
Figure 8 | Profile analyses of marine sample replication. Replicate correlation plots of (a) 16S rRNA gene based taxonomic profiles and (b) functional KO based profiles. Samples with comparable DNA input amounts are connected via a grey box.

Mean coefficient of variation for taxonomic marine community profiles

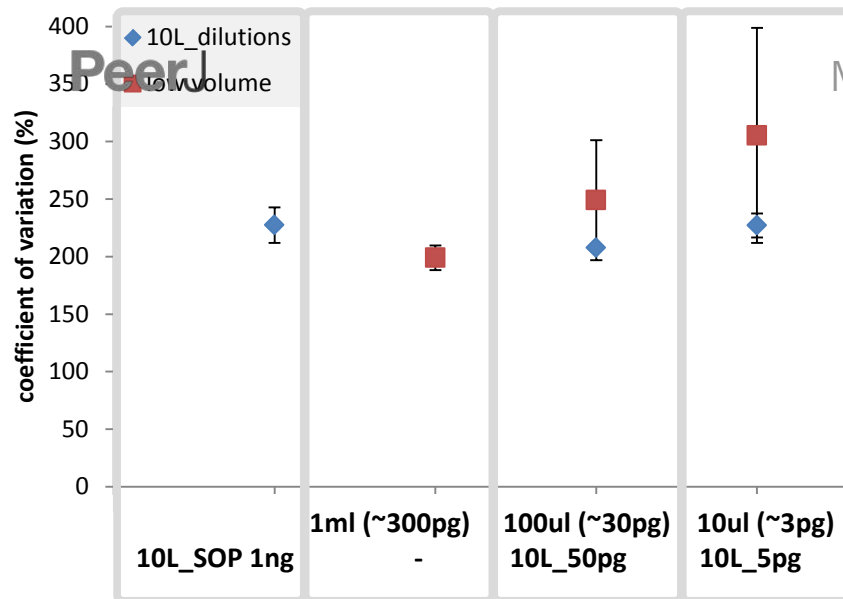
Figure 9 | Mean coefficient of variation for taxonomic marine community profiles. The mean coefficient of variation is applied to compare the 10L dilutions (SOP 1ng, 50pg, 5pg) against the low volume (1ml, 100ul, 10ul) samples using 16S based taxonomic profiles. The X-axis shows the different amounts of input DNA and volumes for the low volume samples (upper row) and the 10L filtration (SOP, and dilutions; lower row).

Mock community population genome bins

Table 1 | Mock community population genome bins. Read were subsample to 50 million reads, and only contigs ≥ 1kb were used for population genome binning. Completeness and contamination were estimated based on marker genes, see Material and Methods.

	assembly			Completeness			Contamination		
	size	max contig	bins	mean	max	min	mean	max	min
mock_SOP_1ng	148,892,162	952,162	24	83.116	99.63	54.62	0.898	6.62	0
mock_100pg	132,682,229	866,524	17	87.504	99.68	53.07	1.136	2.64	0
mock_10pg	115,499,380	852,918	13	83.87	99.45	54.41	1.812	6.9	0
mock_1pg	54,332,814	522,839	9	85.994	98.86	57.55	0.978	2.03	0