Hidden heterogeneity and co-occurrence networks of soil prokaryotic communities revealed at the scale of individual soil aggregates

Running title: Microbiota at soil aggregate-level

Márton Szoboszlay and Christoph C. Tebbe[1]

Thünen Institut für Biodiversität, Bundesforschungsinstitut für Ländliche Räume, Wald und Fischerei, Braunschweig, Germany

Corresponding author: christoph.tebbe@thuenen.de

Address of corresponding author:

Prof. Dr. Christoph Tebbe, Thünen Institut für Biodiversität, Bundesallee 65, D-38116 Braunschweig, Germany

Declaration by the authors:

There are no competing financial interests in relation to the work described.

Abstract

Sequencing PCR-amplified gene fragments from metagenomic DNA is a widely applied method for studying the diversity and dynamics of soil microbial communities. Typically DNA is extracted from 0.25 to 1 g of soil. These amounts, however, neglect the heterogeneity of soil present at the scale of soil aggregates; and thus, ignore a crucial scale for understanding the structure and functionality of soil microbial communities. Here we show with an arable soil the impact of reducing the amount of soil used for DNA extraction from 250 mg to approx. 1 mg in order to access spatial information on the prokaryotic community structure as indicated by 16S rRNA-gene amplicon analyses. Furthermore, we demonstrate that individual aggregates from the same soil differ in their prokaryotic communities. The analysis of 16S rRNA gene amplicon sequences from individual soil aggregates allowed us, in contrast to 250 mg soil samples, to construct a co-occurrence network that provides insight into the structure of microbial associations in the studied soil. Two dense clusters were apparent in the network, one dominated by *Thaumarchaeota*, *Verrucomicrobia*, and *Actinobacteria*, and the other by *Acidobacteria* subgroup 6, linked by three putative keystone taxa (candidatus *Udaeobacter*, Verrucomicrobia; a member of Actinobacteria and Acidobacteria, respectively).

Introduction

One of the major challenges in microbial ecology is to elucidate the principles, patterns, and interactions which lead to the formation of highly diverse microbial communities as found for example in soil [1, 2]. To gain a predictive understanding of soil microbial diversity, it is essential to consider spatial scale information linked to soil structure [3, 4].

Soil structure develops as primary particles of different sizes and mineral composition, i.e., clay, silt, and sand, interact with each other, and with organic material to build microaggregates and macroaggregates, both distinguished by their diameter, i.e., below or above 250 µm [5]. Most bacterial cells occur inside aggregates rather than on their surfaces [6]. Biogeochemical cycles, which are key ecosystem services driven by an interacting microbial community [7], are considered to mainly take place within aggregates [8]. Aggregates have been regarded as "massively concurrent evolutionary incubators" [9], or as "microbial villages" [8], that represent small communities separated by distance and physical barriers and connected only periodically, e.g. during wetting events.

DNA-based methods to assess microbial diversity typically start with extracting 250 mg to 1 g of soil [10]. This strategy has been useful to investigate the overall microbial diversity of soils [11, 12], its variation across geographical regions [13, 14], or its response to land use change at a continental scale [15]. However, in order to understand the processes and interactions taking place within soil microbial communities, it would be rewarding to increase the spatial resolution to individual aggregates and investigate microbial diversity in these spatial entities. This would make it possible to assess the micro-heterogeneity of soil microbial communities and increase the likelihood of detecting interacting microbial partners.

54 The potential impact of the heterogeneous soil constituents on structuring microbial
55 communities at a microscale has already been demonstrated with pooled soil primary particles,
56 where the majority of abundant bacterial and fungal taxa exhibited particular preferences for
57 clay, silt, or sand-sized fractions [16, 17]. Comparing pooled samples of micro- and
58 macroaggregates revealed that they also differ in microbial community composition [18-20],
59 diversity [21, 22] and their response to stress [23]. However, information on the heterogeneity
60 of microbial communities of individual aggregates within specific aggregate fractions is still
61 lacking. A major limitation of analyzing individual aggregates is the difficulty of obtaining a
62 sufficient quantity of nucleic acids from small amounts of soil for molecular methods. Attempts
63 made so far, therefore, either pooled several aggregates for DNA extraction [21, 22, 24],
64 sampled very large aggregates of 20 – 70 mg [25], or applied whole genome amplification [26].
65 These solutions, however, do not deliver data on individual aggregates, provide coarse spatial
66 resolution, or generate substantial bias in the results [27, 28], respectively. To our knowledge,
67 the only study that reported the bacterial community composition in smaller (2 mm) individual
68 aggregates without applying whole genome amplification utilized taxonomic microarrays; a
69 method of relatively low resolution, and focused solely on linking enzyme activity profiles with
70 community structure [29]. Furthermore, applying molecular methods to small samples that yield
71 very low amounts of nucleic acids require validation to prove the consistent performance of the
72 methods and rule out the possibility of contamination and stochastic effects influencing the
73 results.

74 The tremendous scientific potential that individual soil aggregate-based microbial community
75 analysis should have for characterizing the heterogeneity of soil microbial communities at a

76 biologically and ecologically more meaningful scale motivated us testing the following
77 hypotheses in this study:

78 1. Metagenomic DNA of sufficient quantity and quality for PCR-based analyses can be
79 extracted from soil samples as small as 1 mg
80 2. Increasing spatial resolution reveals heterogeneity in soil bacterial and archaeal
81 community structure and abundance
82 3. A higher heterogeneity seen among small soil samples is not a result of contamination or
83 sub-optimal performance of molecular methods.
84 4. Comparing individual aggregates from the same soil unveils patterns of microbial co-
85 occurrence within the soil microbial community not seen with the commonly used
86 250 mg soil sample size.

87 **Materials and methods**

88 *Overview of the experiments*

89 Three experiments were conducted in this study. In the 1st experiment, samples decreasing in
90 size from 250 mg to 1 mg taken from the same soil were subjected to DNA extraction. To
91 address the first two hypotheses, qPCR and high-throughput amplicon sequencing targeting the
92 16S rRNA gene were conducted to characterize the bacterial, archaeal, and fungal abundance
93 and the prokaryotic diversity in these samples. The 2nd experiment addressed the third
94 hypothesis by comparing 250 mg soil samples and aliquots of a homogenized soil slurry. The
95 volumes of the aliquots were chosen to contain the amount of DNA expected from 1 mg, 5 mg,

96 and 25 mg soil samples. Since all aliquots were taken from the same thoroughly homogenized
97 soil slurry, differences in prokaryotic community structure between these soil homogenate
98 samples must be results of contamination, stochastic effects, or sub-optimal performance of the
99 DNA extraction and PCR. In the 3rd experiment, DNA was extracted from individual aggregates
100 and 250 mg soil samples taken from the same soil to address the fourth hypothesis. All
101 experiments included several control samples to test for the presence of contamination.

102 *Soil sampling and DNA extraction*

103 The soil used in all experiments was a loam topsoil of a Haplic Chernozem (FAO classification)
104 from the Bad Lauchstädt experimental research station of the Helmholtz Centre for
105 Environmental Research in Germany (51°24'N 11°53'E). It was collected in December from a plot
106 under long term sugar beet – potato – winter wheat – barley rotation without fertilization. The
107 soil had pH 7.1 (in 0.01 M $CaCl_2$) and 17.7 g kg^{-1} organic C. It was sieved (2 mm) and stored at
108 4°C until use.

109 Before sampling, approximately 100 g soil was incubated at room T in the dark for 24 h. The soil
110 was then spread out in a sterile petri dish and samples were taken with sterilized spatulas
111 directly into the bead-beating tubes of the DNA extraction kit. Control samples were included in
112 all experiments. They received no soil but were handled together with and the same way as the
113 soil samples. DNA was extracted with the Quick-DNA Fecal/Soil Microbe Microprep Kit (Zymo
114 Research, Freiburg, Germany) including two 45 s bead beating cycles in an MP FastPrep-24 5G
115 Instrument (MP Biomedicals, Eschwege, Germany) at 6.5 m/s speed with a 300 s break in
116 between. The DNA extracts were eluted in 30 µl elution buffer. All work was done in a biosafety

117 cabinet decontaminated with UV light to minimize the chance of contamination. Measurement
118 of the DNA yield was attempted with Quant-iT PicoGreen dsDNA Assay Kit (Molecular Probes,
119 Life Technologies, Eugene, Or) but accurate quantification was not possible for many of the
120 small samples due to results close to the background fluorescence in the blank controls. In
121 preparation of this study, several DNA extractions methods were tested but were not found
122 suitable. This included the Fast DNA Spin kit for soil (MP Biomedicals, LLC, Illkrich, France) and
123 variations of the phenol-chloroform protocol [30].

124 In the 1st experiment, five different size-groups of soil samples were collected: 250 mg, 125 mg,
125 25 mg, 5 mg, and 1 mg, respectively. Eight samples were taken from each size-group along with
126 six control samples. Sample weights from all experiments are listed in Table 1. In the 2nd
127 experiment, ten samples of 250 mg soil and six control samples were taken. Eight of the soil
128 samples were processed normally in the DNA extraction, while for the other two, the DNA
129 extraction was interrupted after the centrifugation following the bead beating. By this point, the
130 samples had been turned into homogenized slurry by the bead beating, and the soil debris had
131 been separated from the supernatant that contained the metagenomic DNA. The supernatant
132 from the two samples were pooled and homogenized by vortexing. The mass of the resulting
133 suspension was 1 112 mg and it originated from 497 mg soil in total. Accordingly, a 55.9 mg
134 aliquot of this suspension contained the amount of DNA extractable from 25 mg soil, an 11.2 mg
135 aliquot the amount from 5 mg soil, and a 2.2 mg aliquot the amount from 1 mg soil. Eight
136 aliquots from each of these sizes, hereafter referred to as 25 mg, 5 mg, and 1 mg soil
137 homogenate samples, were taken and each mixed with 350 µl bashing bead buffer from the
138 DNA extraction kit to continue the DNA extraction. In the 3rd experiment, 37 individual soil

139 aggregates, weighing 5.3 mg on average, were taken for DNA extraction along with 35 samples
140 of 250 mg soil and nine control samples.

141 *Abundances of microbial groups assessed by qPCR*

142 The abundance of *Bacteria*, *Archaea*, and *Fungi* were assessed by qPCR targeting the 16S rRNA
143 gene and the ITS region respectively following the protocol of Hemkemeyer *et al.* [31]. Archaeal
144 and fungal abundance was only investigated in the 1st experiment. Reactions were run in a Bio-
145 Rad CFX96 real-time PCR system in duplicates from different dilutions of the DNA extracts. In
146 case of the 250 mg and 125 mg soil samples, 50- and 100-fold dilutions were taken; from the 25
147 mg samples 10- and 20-fold dilutions; and from the 5 mg, 1 mg, individual aggregate, and
148 control samples undiluted DNA extracts and two-fold dilutions were used. PCR efficiencies were
149 95.9 – 104.6% for the bacterial 16S rRNA gene, 94.2 – 96.2% for the archaeal 16S rRNA gene,
150 and 83.3 – 84.3% for the fungal ITS with $R^2 \geq 0.995$ in all cases. Results were compared with
151 Tukey's HSD tests, or Welch's t-test in case of the data from the 3rd experiment. The analysis
152 was carried out in R 3.4.4 (www.r-project.org). One of the 250 mg samples from the 1st
153 experiment yielded a magnitude higher copy number in the fungal ITS qPCR assay than the
154 others. It was deemed to be an outlier and excluded from the analysis.

155 *High-throughput sequencing of 16S rRNA gene amplicons and data processing*

156 To characterize the bacterial and archaeal communities in the samples, DNA extracts were
157 subjected to high-throughput amplicon sequencing of the V4 region of the 16S rRNA gene
158 following the protocol of Kozich *et al.* [32] with primers updated to match the modified 515f
159 and 806r sequences according to Walters *et al.* [33]. In case of the small soil samples and

160 control samples, due to the low DNA yield, 10 µl DNA extract was used as template in the PCRs.
161 Paired-end sequencing was done on Illumina MiSeq instruments at StarSEQ (Mainz, Germany).
162 Samples from the same experiment were sequenced in the same run. All data is available at the
163 European Nucleotide Archive under the accession numbers PRJEB36881 (1st experiment),
164 PRJEB36883 (2nd experiment), and PRJEB36887 (3rd experiment).

165 The sequencing data from the three experiments were analyzed separately. Raw reads were
166 processed with the dada2 (version 1.6.0) pipeline [34] in R 3.4.4. Forward and reverse reads
167 were truncated at positions 240 and 90, respectively. Reads with any ambiguous bases were
168 discarded as well as forward reads with over two and reverse reads with over one expected
169 error. The data from the 2nd experiment had higher quality allowing the reverse reads to be
170 truncated at position 130 and keeping those with two or less expected errors. Error models
171 were constructed from 10^6 randomly selected reads. Sequence variants (SVs) were inferred
172 using the pool option. Forward and reverse SVs were merged trimming overhangs, and the
173 removeBimeraDenovo function was employed to detect chimeras. The SVs were classified
174 according to the SILVA reference version 132 [35] accepting only results with ≥70% bootstrap
175 support. SVs shorter than 220 nt or longer than 275 nt, or identified as chimeric, mitochondrial,
176 or chloroplast sequences, or not classified into *Bacteria* or *Archaea* were deleted. SVs with
177 ≥0.1% relative abundance in any of the control samples of an experiment were regarded as
178 potentially contaminant and removed from the dataset of that experiment.

179 *Analysis of the sequencing results*

180 Principal component analysis (PCA) plots were created in R using the rda function of the vegan
181 package version 2.5-2 [36]. To decrease the sparsity of the data, SVs not reaching 0.1% relative
182 abundance in any of the samples included in the PCA were removed. Zeroes were replaced with
183 the count zero multiplicative (CZM) method using the zCompositions package version 1.1.1 [37]
184 and centered log-ratio (CLR) transformation was applied to the data to correct for compositional
185 effects and differences in sequencing depth [38].

186 Aitchison distances between the samples were calculated as Euclidean distances in the CLR
187 transformed dataset [39]. SVs that didn't have at least 0.1% relative abundance in any of the
188 compared samples were removed and zeroes were replaced with the CZM method to allow CLR
189 transformation before calculating Euclidean distances with the 'vegdist' function of the vegan
190 package.

191 Plots illustrating the prevalence of abundant SVs among the soil samples were prepared in
192 Cytoscape 3.7.1 (www.cytoscape.org). CoNet 1.1.1 beta [40] in Cytoscape was used to construct
193 co-occurrence networks from the data from the 3rd experiment. To limit the number of parallel
194 significance tests and the sparsity of the data, only SVs with ≥0.2% relative abundance in at least
195 one of the samples were included. Separate networks were constructed for the 250 mg soil
196 samples and the soil aggregates. However, the selection of SVs was done on the joint data
197 matrix to ensure that both networks include the same SVs. The data was relativized to the total
198 sequence count of each sample. Pearson and Spearman correlations, mutual information (jsl
199 setting), and Bray-Curtis and Kullback-Leibler (with a pseudo count of 10^{-8}) dissimilarities were
200 calculated and the 1 000 highest and 1 000 lowest scoring edges from each of the five metrics

201 were kept. The ReBoot method [41], which mitigates compositional effects, was used to assess
202 the significance of the edges based on 1 000 permutations with renormalization and 1 000
203 bootstrap iterations. In the network of the soil aggregates, edges with scores below the 2.5th
204 and over the 97.5th percentile of the bootstrap distribution or not supported by at least three of
205 the five metrics were considered unstable and removed. Brown's method of p-value merging
206 was applied followed by Benjamini-Hochberg correction. Only edges with $q \leq 0.05$ were
207 included in the final network. In the network of the 250 mg samples, unstable edges were not
208 removed and the Benjamini-Hochberg correction was not applied as otherwise no edges were
209 retained. The networks were visualized in Cytoscape using the compound spring embedder
210 layout. Topological parameters were calculated with NetworkAnalyzer version 2.7 [42].

211 **Results**

212 *Metagenomic DNA can be extracted from soil samples as small as 1 mg*

213 DNA could be extracted from all soil samples including aggregates and amounts as little as 0.87
214 mg. In all cases, the extracted DNA was sufficient for 16S rRNA gene amplicon sequencing and
215 the qPCR assays. Accurate quantification of the DNA yield with PicoGreen assays was not
216 possible because the fluorescence readings from many of the small samples were close to the
217 background fluorescence in the blank controls.

218 To assess whether the DNA extraction could recover microbial DNA with similar efficiency from
219 small quantities of soil as from 250 mg samples, estimates of the abundances of *Bacteria*,
220 *Archaea*, and *Fungi* in one gram of soil were calculated from the qPCR results (Figure 1). Similar

221 estimates were obtained from the 250 mg and 5 mg soil samples from the 1st experiment.
222 Estimates from the 1 mg samples tended to be lower but were not significantly different. In
223 contrast, the estimates of fungal abundance in a gram of soil were significantly lower from the 1
224 and 5 mg than from the 250 mg samples. Estimates of bacterial abundance obtained from the
225 25 mg and 125 mg samples and archaeal abundance from the 25 mg samples were significantly
226 higher than from the 250 mg samples. Bacterial abundance estimates from the single
227 aggregates and 250 mg soil samples of the 3rd experiment covered the same range (Figure S1),
228 but the mean of the estimates from the single aggregates (2.27×10^9 copies/g soil) was lower (p
229 < 0.001) than from the 250 mg samples (4.69×10^9 copies/g soil).

230 *Increasing spatial resolution reveals heterogeneity in soil bacterial and archaeal*
231 *community structure but not in their abundance*

232 The yield of high-quality 16S rRNA gene amplicon sequences did not differ significantly between
233 small soil samples or single aggregates and the 250 mg samples in any of the three experiments
234 (Figure 2A). However, in the 1st experiment, the sequencing yield was somewhat lower from the
235 1 mg samples than from the 5 mg and 25 mg samples. The number of SVs detected in the
236 samples from the 1st experiment was not significantly different between the 250 mg, 125 mg,
237 and 25 mg samples, but decreased significantly in the 5 mg and even further among the 1 mg
238 samples (Figure 2B). Similarly, significantly lower numbers of SVs were detected in the single
239 aggregates than in the 250 mg soil samples of the 3rd experiment.

240 Principle component analysis (PCA) from the sequencing results from the 1st experiment
241 arranged all 250 mg and 125 mg samples, and most 25 mg samples into a single, tight group,

242 indicating high similarity in their prokaryotic community structures (Figure 3A). In contrast,
243 samples from the 5 mg and more so from the 1 mg categories, showed higher heterogeneity in
244 community structure. Similarly, PCA indicated heterogeneity between individual aggregates that
245 was not seen among the 250 mg samples in the 3rd experiment (Figure 3B).

246 The abundant SVs ($\geq$ 0.1% relative abundance in at least one of the samples) in the 250 mg, 125
247 mg, and 25 mg samples from the 1st experiment were almost all detectable in each individual
248 sample, showing that the composition of the soil prokaryotic community appears very uniform
249 when investigated at such a coarse spatial resolution (Figure 4). In contrast, 172 of the abundant
250 SVs detected in the 1 mg samples were unique to just one or two of these samples. Among
251 these SVs, representatives of *Planctomycetes*, *Proteobacteria*, and *Acidobacteria* were especially
252 numerous, while *Thaumarchaeota* and *Actinobacteria* were dominant among the SVs present in
253 all samples. The 5 mg samples represented a level of spatial resolution at which some
254 heterogeneity in the prevalence of the abundant SVs was already apparent with 22 of them
255 detectable in two or only in a single sample.

256 In total, 5 620 SVs were detected in the eight 1 mg soil samples of the 1st experiment. Of these,
257 4 764 (85%) were also present in at least half of the 250 mg samples. The remaining 856 SVs had
258 low relative abundance in the 1 mg samples with only 59 reaching more than 0.1% relative
259 abundance in any of them. The 5 mg samples together contained 8 010 SVs, of which 6 443
260 (80%) were also detectable in at least half of the 250 mg samples. Of the remaining 1 567 SVs,
261 only 26 reached more than 0.1% relative abundance in any of the 5 mg samples.

262 The qPCR results did not confirm our hypothesis that increasing spatial resolution would reveal
263 heterogeneity in microbial abundance. Bacterial and archaeal 16S rRNA gene, and fungal ITS

264 copy numbers did not show a larger variation among the 1 and 5 mg samples than between the
265 250 mg samples from the 1st experiment (Figure 1). Similarly, in the 3rd experiment, bacterial
266 abundance did not vary more in the single aggregates than in the 250 mg samples (Figure S1).

267 *Impact of contamination, inconsistent performance of the methods, and stochastic*
268 *effects*

269 The control samples amplified in the qPCR assay targeting the bacterial 16S rRNA gene but
270 yielded only 3 to 356 copies per µl DNA extract. In comparison, the 1 mg soil samples had 8 777
271 to 45 534 copies per µl DNA extract. The control samples from the 1st experiment had 0 to 13
272 fungal ITS copies per µl DNA extract and none of them showed amplification in the archaeal 16S
273 rRNA gene qPCR assays. It was possible to generate sequencing results from all control samples.
274 They were very similar to each other in their prokaryotic community structures but very
275 different from the soil samples (Figure S2). Every SV that reached 0.1% relative abundance in
276 any of the control samples of an experiment was considered as a potential contaminant and
277 removed from the datasets. There were 450, 77, and 591 such SVs in the datasets of the 1st, 2nd,
278 and 3rd experiments respectively. In the data from the 1st experiment, these SVs together
279 covered 98.9 – 99.8% of the sequences obtained from the control samples and 2.5 – 6.6% of the
280 sequences from the soil samples. In the 2nd and 3rd experiments, 92.5 – 99.5% and 98.6 – 99.6%
281 of the sequences from the control samples, and 2.9 – 17.3% and 5.9 – 9.4% of the sequences
282 from the soil samples, respectively, were covered by the potentially contaminant SVs.

283 The estimated bacterial abundance in a gram of soil based on the qPCR results was in general
284 higher in the samples from the 2nd experiment but followed the same pattern as in the samples

285 from the 1st experiment with no differences between the 1 mg, 5 mg, and 250 mg samples but
286 significantly higher values in the 25 mg samples (Figure S3). The small soil homogenate samples
287 did not show the degree of heterogeneity in prokaryotic community structure we observed
288 among the small soil samples of the 1st experiment. Abundant SVs (≥ 0.1% relative abundance in
289 at least one sample) in the 25 mg soil homogenate samples were all detectable in at least five of
290 the eight replicates. Out of the 354 abundant SVs in the 5 mg soil homogenate samples, one
291 was present in only three of the samples but the others were detectable in at least six. The 1 mg
292 soil homogenate samples harbored 446 abundant SVs. None of them was unique to a single
293 sample and 442 were present in five or more of the eight samples. The Aitchison distances of
294 community structure were much higher among the 5 mg, and especially among the 1 mg soil
295 samples of the 1st experiment compared to the distances between the 250 mg soil samples
296 (Figure 5). In contrast, the distances between the 1 mg, 5 mg, or 25 mg soil homogenate
297 samples of the 2nd experiment were very similar to the distances among the 250 mg soil
298 samples indicating no difference in the heterogeneity of prokaryotic community structure
299 between these sample groups.

300 *Bacterial and archaeal co-occurrence patterns in 250 mg-soil samples and aggregates*

301 Networks of prokaryotic co-occurrence were constructed using the 272 SVs that reached ≥0.2%
302 relative abundance in at least one of the samples of the 3rd experiment. No network was
303 obtained from the 250 mg samples unless the removal of unstable edges and the correction of
304 the p-values for multiple testing were skipped. The resulting network has 78 edges between 35
305 nodes (Figure 6A). Thus, this spatial resolution revealed only a small number of putative
306 associations many of which are likely false discoveries. In contrast, a network of 137 edges and

307 67 nodes (with the removal of unstable edges and control of the false discovery rate) was
308 obtained from the individual soil aggregates (Figure 6B). A total of 54 of the nodes are part of a
309 connected component in which there are three nodes with high betweenness centrality: SV15
310 (*Verrucomicrobia*, candidatus *Udaeobacter*), SV31 (*Actinobacteria*), and SV36 (*Acidobacteria*
311 subgroup 6). These SVs potentially serve a keystone function by connecting two clusters in the
312 network. One of the clusters contains several SVs of *Thaumarchaeota*, *Verrucomicrobia*, and
313 *Actinobacteria*. The other is dominated by *Acidobacteria* subgroup 6. The hub of the latter
314 cluster is SV58 (*Acidobacteria* subgroup 6) that has the highest degree in the network being
315 connected to 19 nodes. SV399 (*Chloroflexi*) and SV123 (*Acidobacteria* subgroup 6) are linked
316 with negative associations to several members of this cluster.

317 Discussion

318 The large disparity between the scale in which the soil microbiota is usually studied with
319 molecular methods (0.25 to 1 g of soil) and the distance over which microbial interactions occur,
320 impede the detection of interacting partners [43]. In order to gain information on the soil
321 microbial diversity at an increased spatial resolution that considers soil structure, in this study
322 we reduced the amount of soil used for DNA extraction from 250 to 1 mg and also extracted
323 individual soil aggregates. Bacterial and archaeal DNA was recovered with similar efficiency
324 from the 250 mg and the 5 mg and 1 mg samples as shown by the qPCR results. This was not
325 true for fungal DNA. Either the DNA-extraction kit was not efficient in isolating fungal DNA from
326 samples below 25 mg, or fungi may preferentially colonize larger soil aggregates. Our results
327 show that the DNA extraction kit was most efficient in recovering bacterial and archaeal DNA

328 from 25 to 125 mg soil, although the variation in the yield of 16S rRNA gene copies was large
329 among these samples. Since this increased variation was apparent among the 25 mg soil
330 homogenate samples of the 2nd experiment as well, it must be a property of the DNA extraction
331 kit and not an indication of uneven distribution of bacterial cells at the scale of 25 mg samples.

332 As a consequence of sampling small amounts of soil, the DNA extracts had low template
333 concentrations for the subsequent PCR analyses. Thereby we had to anticipate a high risk of
334 contamination affecting the results [44]. In fact, quantifiable amounts of *Bacteria* and *Fungi*, but
335 not *Archaea*, were detected in the control samples without soil. However, they reached no
336 more than 4% of the number of bacterial rRNA genes copies in the smallest soil samples, and
337 thus, the influence of contamination on our results is negligible. The bacterial community found
338 in the control samples was clearly distinct from the soil communities suggesting that the
339 contamination originated from the reagents of the DNA extraction and sequencing library
340 preparation rather than cross-contamination between samples [45, 46]. Another concern of
341 working with small samples is that molecular methods applied to such low amounts of template
342 may perform inconsistently leading to artificial variation in the results. The 2nd experiment
343 showed that the DNA extraction, PCR, and sequencing did not artificially generate more
344 variation in the results from small samples than the variation present among the 250 mg
345 samples. Therefore, the heterogeneity in prokaryotic community composition and structure
346 among the 1 mg and 5 mg soil samples from the 1st experiment was not caused by stochastic
347 effects or PCR bias.

348 Samples of 25 mg up to 250 mg of soil were close to identical in prokaryotic community
349 composition, thus they provide a good representation of the overall prokaryotic diversity of our

350 soil. In contrast, the 1 mg and 5 mg samples and single aggregates were heterogeneous in
351 community structure. We found that while small soil samples could recover some SVs not
352 necessarily detected with the conventionally used 250 mg samples; these SVs were typically low
353 in abundance. Very few exceeded the relative abundance threshold we applied to control the
354 sparsity of the data in our analysis of community structure. Therefore, the large heterogeneity
355 of prokaryotic community structure we observed among the small samples was not because
356 they would have enabled the detection of more SVs. Instead, it appears that they contained
357 different subsets of the total community present in the 250 mg samples. Interestingly, the small
358 samples didn't have more variation in the abundance of *Bacteria, Archaea*, and *Fungi* than the
359 250 mg samples. Microbial cells in soil have a patchy distribution at the scale of a few
360 micrometers [47] but, for the soil of this study apparently not at the scale of macroaggregates
361 or 1 to 5 mg samples.

362 Network analyses based on microbial co-occurrence have been applied to soil samples as large
363 as 10 grams [48]. The scientific value of utilizing much smaller samples is shown by our result
364 that 1 and 5 mg samples contained subsets of the total microbial diversity. The smaller spatial
365 scale increases the likelihood that the observed co-occurrences actually indicate interactions
366 [49]. From 37 soil aggregates, we obtained a complex network of bacterial and archaeal co-
367 occurrence that contained two clusters, one with several *Thaumarchaeota*, *Verrucomicrobia*
368 and *Actinobacteria* SVs, the other mainly with *Acidobacteria* subgroup 6 SVs. Three SVs, which
369 could represent keystone taxa, were found to connect these clusters. Keystone taxa are
370 considered to be the best predictors for whole-community compositional changes [50]. If these

371 putative keystone SVs are abundant in an aggregate, we can expect that members of both
372 clusters are present there.

373 While the spatial scale that we reached in this study is not yet fine enough to reveal most
374 microbial interactions as they may occur in microaggregates [51], it can support the
375 development of hypotheses and experiments to understand the patterns and processes shaping
376 the soil microbiota and modelling its behavior [52-54]. Developing DNA extraction protocols
377 from even small soils samples, as typical for microaggregates, should be a way forward to unveil
378 interactions within soil microbial communities, thereby further enhancing our understanding of
379 how soil microbial communities are organized at structural and functional levels.

380 Acknowledgements

381 We thank Ines Merbach, Department Biozönoseforschung, Helmholtz-Zentrum für
382 Umweltforschung – UFZ, Bad Lauchstädt, Germany, for providing the soil of this study. We
383 thank Naomie Oßwald and Jerome Rischke who supported us at early stages of this work as
384 bachelor students, and Karin Trescher for excellent technical assistance. The study was
385 financially supported by the German Research Foundation (DFG; Grant number TE 383/10-1) in
386 context of the Priority Programme "Rhizosphere spatiotemporal organisation (SPP2089).

387 Conflict of interest

388 The authors declare no conflict of interest.

389 **References**

390 1. Green JL, Bohannan BJM, Whitaker RJ. Microbial biogeography: from taxonomy to traits.
391 Science. 2008. 320;5879:1039-43.

392 2. Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, et al. Challenges in microbial
393 ecology: building predictive understanding of community function and dynamics. ISME J.
394 2016. 10;11:2557-68.

395 3. Vos M, Wolf AB, Jennings SJ, Kowalchuk GA. Micro-scale determinants of bacterial diversity in
396 soil. FEMS Microbiol Rev. 2013. 37;6:936-54.

397 4. Thakur MP, Phillips HRP, Brose U, De Vries FT, Lavelle P, Loreau M, et al. Towards an
398 integrative understanding of soil biodiversity. Biol Rev. 2020. 95;2:350-64.

399 5. Six J, Elliott ET, Paustian K. Soil macroaggregate turnover and microaggregate formation: a
400 mechanism for C sequestration under no-tillage agriculture. Soil Biol Biochem. 2000.
401 32;14:2099-103.

402 6. Ranjard L, Poly F, Combrisson J, Richaume A, Gourbiere F, Thioulouse J, et al. Heterogeneous
403 cell density and genetic structure of bacterial pools associated with various soil
404 microenvironments as determined by enumeration and DNA fingerprinting approach (RISA).
405 Microbial Ecol. 2000. 39;4:263-72.

406 7. Smith P, Cotrufo ME, Rumpel C, Paustian K, Kuikman PJ, Elliott JA, et al. Biogeochemical cycles
407 and biodiversity as key drivers of ecosystem services provided by soils. Soil. 2015. 1;2:665-85.

408 8. Wilpiszeski RL, Aufrecht JA, Retterer ST, Sullivan MB, Graham DE, Pierce EM, et al. Soil
409 aggregate microbial communities: towards understanding microbiome interactions at
410 biologically relevant scales. Appl Environ Microbiol. 2019. 85;14:UNSP e00324-19

411 9. Rillig MC, Müller LAH, Lehmann A. Soil aggregates as massively concurrent evolutionary
412 incubators. ISME J. 2017. 11;9:1943-8.

413 10. Young JM, Rawlence NJ, Weyrich LS, Cooper A. Limitations and recommendations for
414 successful DNA extraction from forensic soil samples: A review. Sci Justice. 2014. 54;3:238-
415 44.

416 11. Fierer N, Jackson RB. The diversity and biogeography of soil bacterial communities. Proc Nat
417 Acad Sci USA. 2006. 103;3:626-31.

418 12. Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AKM, Kent AD, et al. Pyrosequencing
419 enumerates and contrasts soil microbial diversity. ISME J. 2007. 1;4:283-90.

420 13. Griffiths RI, Thomson BC, James P, Bell T, Bailey M, Whiteley AS. The bacterial biogeography
421 of British soils. Environ Microbiol. 2011. 13;6:1642-54.

422 14. Karimi B, Terrat S, Dequiedt S, Saby NPA, Horriguel W, Lelievre M, et al. Biogeography of soil
423 bacteria and archaea across France. Sci Adv. 2018. 4;7:aat1808.

424 15. Szoboszlay M, Dohrmann AB, Poeplau C, Don A, Tebbe CC. Impact of land-use change and
425 soil organic carbon quality on microbial diversity in soils across Europe. FEMS Microbiol
426 Ecol. 2017. 93;12:fix146.

427 16. Hemkemeyer M, Dohrmann AB, Christensen BT, Tebbe CC. Bacterial preferences for specific
428 soil particle size fractions revealed by community analyses. Front Microbiol. 2018. 9:00149.

429 17. Hemkemeyer M, Christensen BT, Tebbe CC, Hartmann M. Taxon-specific fungal preference
430 for distinct soil particle size fractions. Eur J Soil Biol. 2019. 94:103103.

431 18. Constancias F, Prevost-Boure NC, Terrat S, Aussems S, Nowak V, Guillemin JP, et al.
432 Microscale evidence for a high decrease of soil bacterial density and diversity by cropping.
433 Agron Sustain Dev. 2014. 34;4:831-40.

434 19. Davinic M, Fultz LM, Acosta-Martinez V, Calderon FJ, Cox SB, Dowd SE, et al. Pyrosequencing
435 and mid-infrared spectroscopy reveal distinct aggregate stratification of soil bacterial
436 communities and organic matter composition. Soil Biol Biochem. 2012. 46:63-72.

437 20. Fox A, Ikoyi I, Torres-Sallan G, Lanigan G, Schmalenberger A, Wakelin S, et al. The influence
438 of aggregate size fraction and horizon position on microbial community composition. Appl
439 Soil Ecol 2018. 127;19-29.

440 21. Ivanova EA, Kutovaya OV, Tkhakakhova AK, Chernov TI, Pershina EV, Markina LG, et al. The
441 structure of microbial community in aggregates of a typical chernozem aggregates under
442 contrasting variants of its agricultural use. Eurasian Soil Sci. 2015. 48;11:1242-56.

443 22. Bach EM, Williams RJ, Hargreaves SK, Yang F, Hofmockel KS. Greatest soil microbial diversity
444 found in micro-habitats. Soil Biol Biochem. 2018. 118:217-26.

445 23. Ranjard L, Nazaret S, Gourbiere F, Thioulouse J, Linet P, Richaume A. A soil microscale study
446 to reveal the heterogeneity of Hg(II) impact on indigenous bacteria by quantification of
447 adapted phenotypes and analysis of community DNA fingerprints. FEMS Microbiol Ecol.
448 2000. 31;2:107-15.

449 24. Bailey VL, Fansler SJ, Stegen JC, McCue LA. Linking microbial community structure to beta-
450 glucosidic function in soil aggregates. ISME J. 2013. 7;10:2044-53.

451 25. Kravchenko AN, Negassa WC, Guber AK, Hildebrandt B, Marsh TL, Rivers ML. Intra-aggregate
452 pore structure influences phylogenetic composition of bacterial community in
453 macroaggregates. Soil Sci Soc Am J. 2014. 78;6:1924-39.

454 26. Bailey VL, McCue LA, Fansler SJ, Boyanov MI, DeCarlo F, Kemner KM, et al. Micrometer-scale
455 physical structure and microbial composition of soil macroaggregates. Soil Biol Biochem.
456 2013. 65;60-8.

457 27. Direito SOL, Zaura E, Little M, Ehrenfreund P, Roling WFM. Systematic evaluation of bias in
458 microbial community profiles induced by whole genome amplification. Environ Microbiol.
459 2014. 16;3:643-57.

460 28. Wang Y, Gao ZM, Xu Y, Li GY, He LS, Qian PY. An evaluation of multiple annealing and
461 looping based genome amplification using a synthetic bacterial community. Acta Oceanol
462 Sin. 2016. 35;2:131-6.

463 29. Kim H, Nunan N, Dechesne A, Juarez S, Grundmann G. The spatial distribution of exoenzyme
464 activities across the soil micro-landscape, as measured in micro- and macro-aggregates, and
465 ecosystem processes. Soil Biol Biochem. 2015. 91;258-67.

466 30. Miller DN, Bryant JE, Madsen EL, Ghiorse WC. Evaluation and optimization of DNA extraction
467 and purification procedures for soil and sediment samples. Appl Environ Microbiol. 1999.
468 65;11:4715-24.

469 31. Hemkemeyer M, Christensen BT, Martens R, Tebbe CC. Soil particle size fractions harbour
470 distinct microbial communities and differ in potential for microbial mineralisation of organic
471 pollutants. Soil Biol Biochem. 2015. 90;255-65.

472 32. Kozich JJ, Westcott SL, Baxter NT, Highlander SK, Schloss PD. Development of a dual-index
473 sequencing strategy and curation pipeline for analyzing amplicon sequence data on the
474 MiSeq Illumina sequencing platform. Appl Environ Microbiol. 2013. 79;17:5112-20.

475 33. Walters W, Hyde ER, Berg-Lyons D, Ackermann G, Humphrey G, Parada A, et al. Improved
476 bacterial 16S rRNA gene (V4 and V4-5) and fungal internal transcribed spacer marker gene
477 primers for microbial community surveys. mSystems. 2016. 1;1:e00009-15.

478 34. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-
479 resolution sample inference from Illumina amplicon data. Nat Methods. 2016. 13;7:581-3.

480 35. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig WG, Peplies J, et al. SILVA: a comprehensive
481 online resource for quality checked and aligned ribosomal RNA sequence data compatible
482 with ARB. Nucleic Acids Res. 2007. 35;21:7188-96.

483 36. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan: community
484 ecology package. R package version 2.5-2. 2018. https://CRAN.R-
485 project.org/package=vegan.

486 37. Palarea-Albaladejo J, Martin-Fernandez JA. zCompositions - R Package for multivariate
487 imputation of left-censored data under a compositional approach. Chemometrics Intell Lab
488 Syst. 2015. 143;85-96.

489 38. Gloor GB, Wu JR, Pawlowsky-Glahn V, Egozcue JJ. It's all relative: analyzing microbiome data
490 as compositions. Ann Epidemiol. 2016. 26;5:322-9.

491 39. Gloor GB, Macklaim JM, Pawlowsky-Glahn V, Egozcue JJ. Microbiome datasets are
492 compositional: and this is not optional. Front Microbiol. 2017. 8;6:02224.

493 40. Faust K, Raes J. CoNet app: inference of biological association networks using Cytoscape
494 [version 2; peer review: 2 approved]. 2016. 5:1519. f1000research.9050.2.

495 41. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial Co-
496 occurrence relationships in the human microbiome. PLoS Comput Biol. 2012. 8;7:
497 e1002606.

498 42. Assenov Y, Ramirez F, Schelhorn SE, Lengauer T, Albrecht M. Computing topological
499 parameters of biological networks. Bioinformatics. 2008. 24;2:282-4.

500 43. Nunan N. The microbial habitat in soil: Scale, heterogeneity and functional consequences. J
501 Plant Nutr Soil Sci. 2017. 180;4:425-9.

502 44. Weiss S, Amir A, Hyde ER, Metcalf JL, Song SJ, Knight R. Tracking down the sources of
503 experimental contamination in microbiome studies. Genome Biol. 2014. 15;12:564.

504 45. Glassing A, Dowd SE, Galandiuk S, Davis B, Chiodini RJ. Inherent bacterial DNA
505 contamination of extraction and sequencing reagents may affect interpretation of
506 microbiota in low bacterial biomass samples. Gut Pathog. 2016. 8:24.

507 46. Salter SJ, Cox MJ, Turek EM, Calus ST, Cookson WO, Moffatt MF, et al. Reagent and
508 laboratory contamination can critically impact sequence-based microbiome analyses. BMC
509 Biol. 2014. 12:87.

510 47. Nunan N, Wu KJ, Young IM, Crawford JW, Ritz K. Spatial distribution of bacterial
511 communities and their relationships with the micro-architecture of soil. FEMS Microbiol
512 Ecol. 2003. 44;2:203-15.

513 48. Khan MAW, Bohannan BJM, Nüsslein K, Tiedje JM, Tringe SG, Parlade E, et al. Deforestation
514 impacts network co-occurrence patterns of microbial communities in Amazon soils. FEMS
515 Microbiol Ecol. 2019. 95;2:fiy230.

516 49. Cordero OX, Datta MS. Microbial interactions and community assembly at microscales. Curr
517 Opin Microbiol. 2016. 31:227-34.

518 50. Herren CM, McMahon KD. Keystone taxa predict compositional change in microbial
519 communities. Environ Microbiol. 2018. 20;6:2207-17.

520 51. Raynaud X, Nunan N. Spatial ecology of bacteria at the microscale in soil. PLOS ONE. 2014.
521 9;1:10.1371.

522 52. Faust K, Raes J. Microbial interactions: from networks to models. Nat Rev Microbiol. 2012.
523 10;8:538-50.

524 53. Rillig MC, Ryo M, Lehmann A, Aguilar-Trigueros CA, Buchert S, Wulf A, et al. The role of
525 multiple global change factors in driving soil functions and microbial biodiversity. Science.
526 2019. 366;6467:886-90.

527 54. Tecon R, Or D. Biophysical processes supporting the diversity of microbial life in soil. FEMS
528 Microbiol Rev. 2017. 41;5:599-623.

529 **Figure legends**

530 **Figure 1** Estimates of (A) bacterial, (B) archaeal, and (C) fungal abundance in a gram of soil
531 based on qPCR from the samples from the 1st experiment (gene copy numbers per g of soil wet
532 weight). One of the 250 mg soil samples was an outlier in the fungal ITS qPCR results and is not
533 included in the plot. Sample groups not labelled with the same letter were significantly different
534 in Tukey's HSD tests.

535 **Figure 2** Number of (A) sequences and (B) sequence variants (SVs) in the samples after the
536 removal of potentially contaminant SVs. Letters indicate significant differences between sample
537 groups of the 1st experiment according to Tukey's HSD tests. * indicates significant difference
538 based on Welch's t-test between the aggregate and the 250 mg soil samples of the 3rd
539 experiment.

540 **Figure 3** Principle component analyses (PCA) plots from the 16S rRNA gene sequencing data
541 from the 1st (A) and 3rd (B) experiments.

542 **Figure 4** SVs arranged according to in how many of the 1 mg, 5 mg, 25 mg, 125 mg, or 250 mg
543 samples from the 1st experiment they were detected. Only SVs that reached at least 0.1%
544 relative abundance in any of the samples are included. Each node represents one SV colored
545 based on its phylum-level classification and sized according to its average relative abundance
546 across all samples excluding those in which it was not detected.

547 **Figure 5** Aitchison distances in the bacterial and archaeal community structure (16SrRNA gene
548 amplicons) within sample groups from the 1st and 2nd experiments.

549 **Figure 6** Co-occurrence networks from the (A) 250 mg samples and (B) the single aggregates
550 from the 3rd experiment. The frames mark the two clusters discussed in the text. It should be
551 noted that in (A) unstable edges were *not* removed and the Benjamini-Hochberg correction for
552 multiple comparisons was not applied.

553 **Table 1** Number of samples and soil weights within each sample category

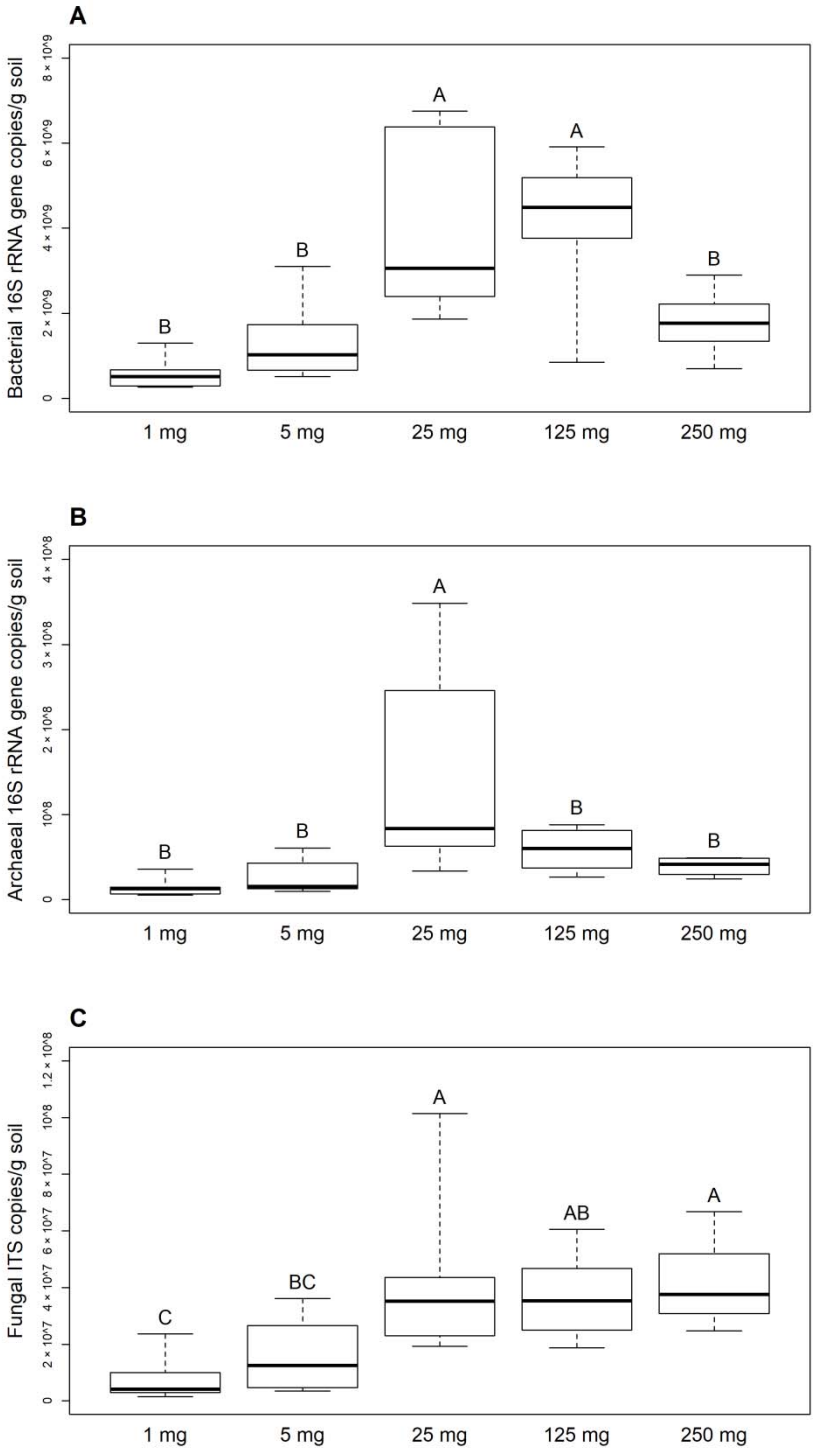
554

Soil weight class or sample type	Sample weight [mg] ± SD	Number of soil samples
1st Experiment		
250-mg	251 ± 1	8
125-mg	125 ± 1	8
25-mg	25.1 ± 0.4	8
5-mg	4.9 ± 0.2	8
1-mg	1.1 ± 0.2	8
Control, no soil	-	6
2nd Experiment		
250-mg	252 ± 3	8
25-mg soil homogenate	-	8
5-mg soil homogenate	-	8
1-mg soil homogenate	-	8
Control, no soil	-	6
3rd Experiment		
250-mg	251 ± 0	35
Soil aggregate	5.3 ± 1.6	37
Control	-	9

555

556

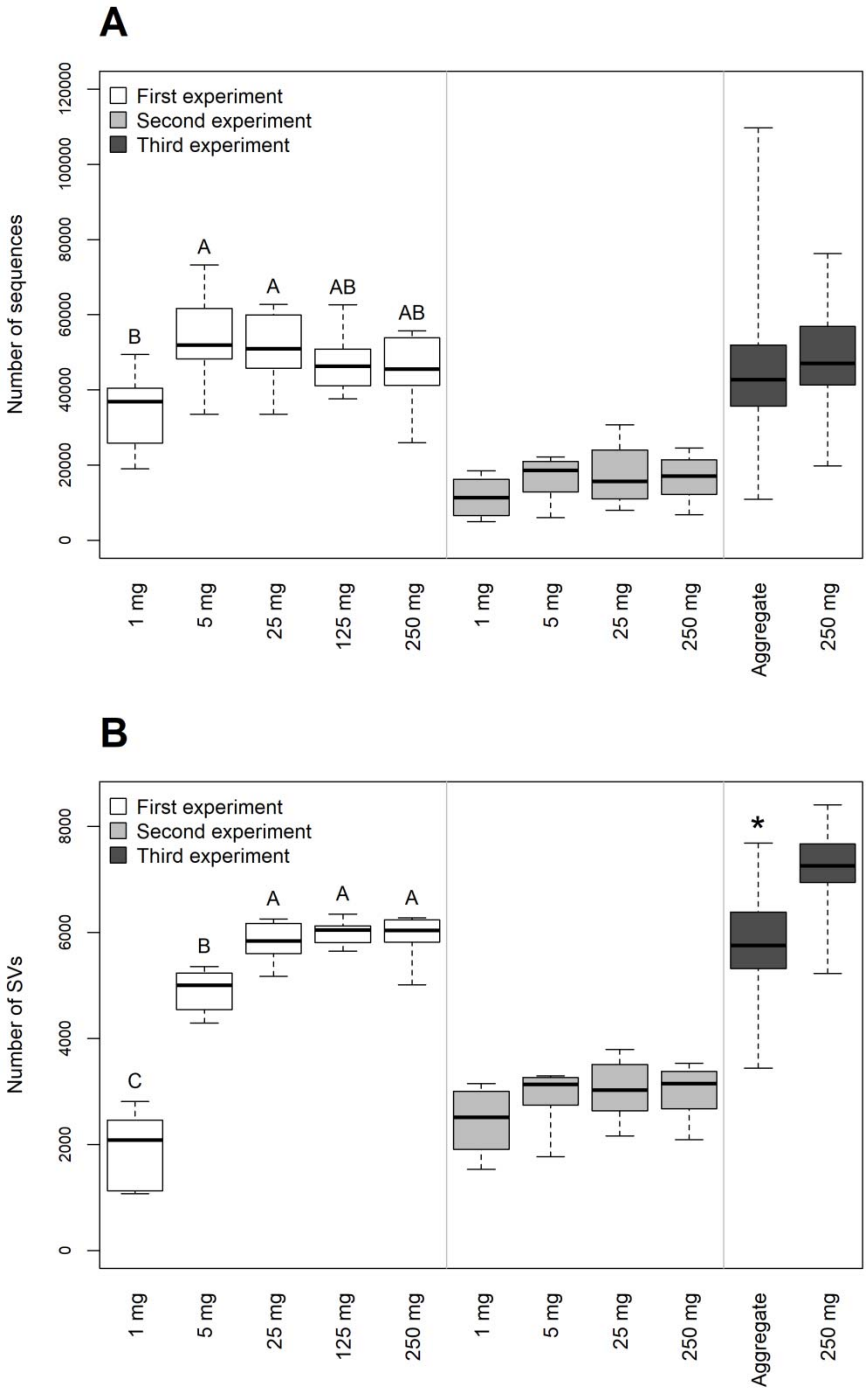
557 Figure 1



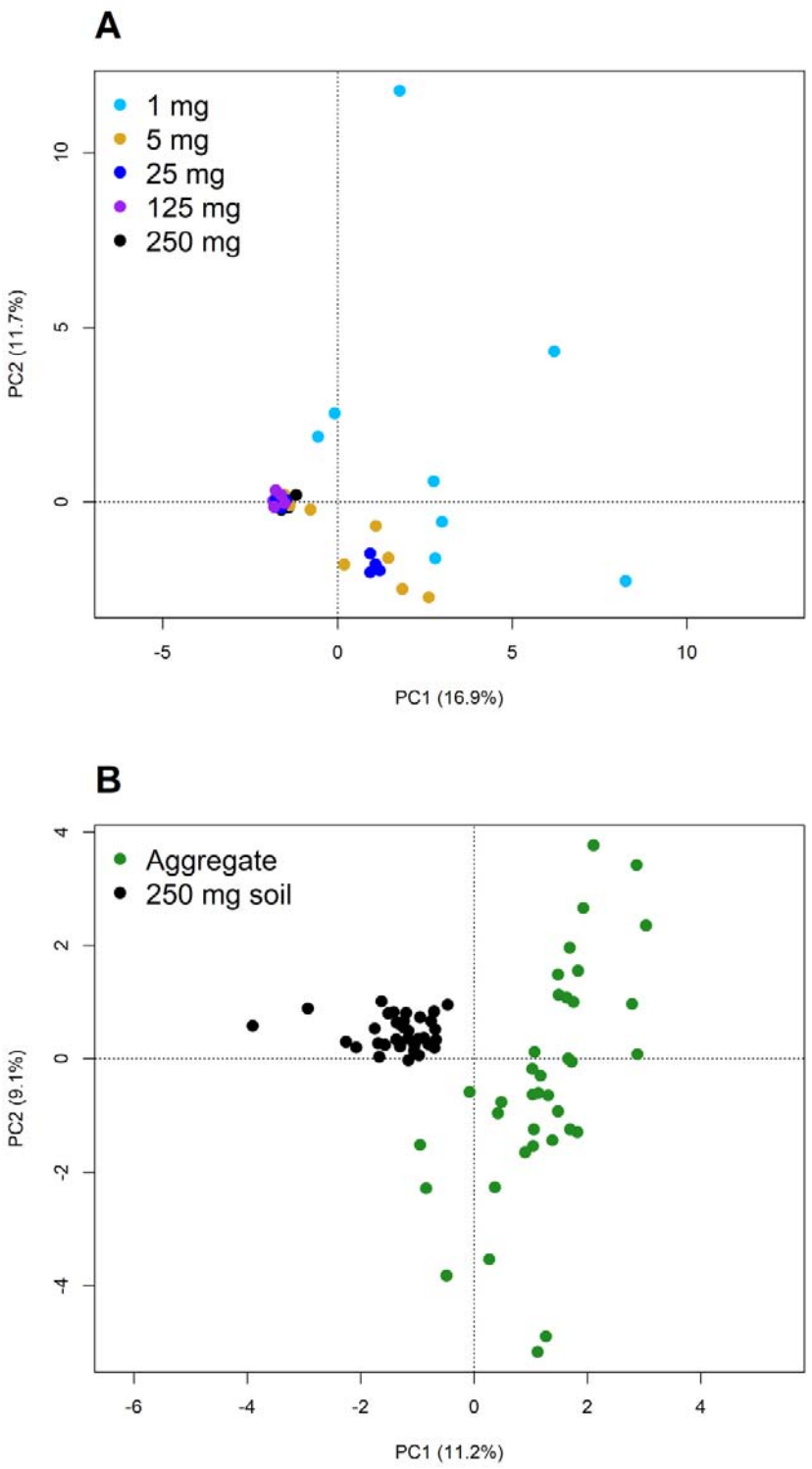
558

559

Figure 2


A
First experiment
Second experiment
Third experiment
Number of sequences
0
20000
40000
60000
80000
100000
120000
B
A
A
AB
AB
1 mg
5 mg
25 mg
125 mg
250 mg
1 mg
5 mg
25 mg
250 mg
Aggregate
250 mg
B
First experiment
Second experiment
Third experiment
Number of SVs
0
2000
4000
6000
8000
C
B
A
A
A
*
1 mg
5 mg
25 mg
125 mg
250 mg
1 mg
5 mg
25 mg
250 mg
Aggregate
250 mg

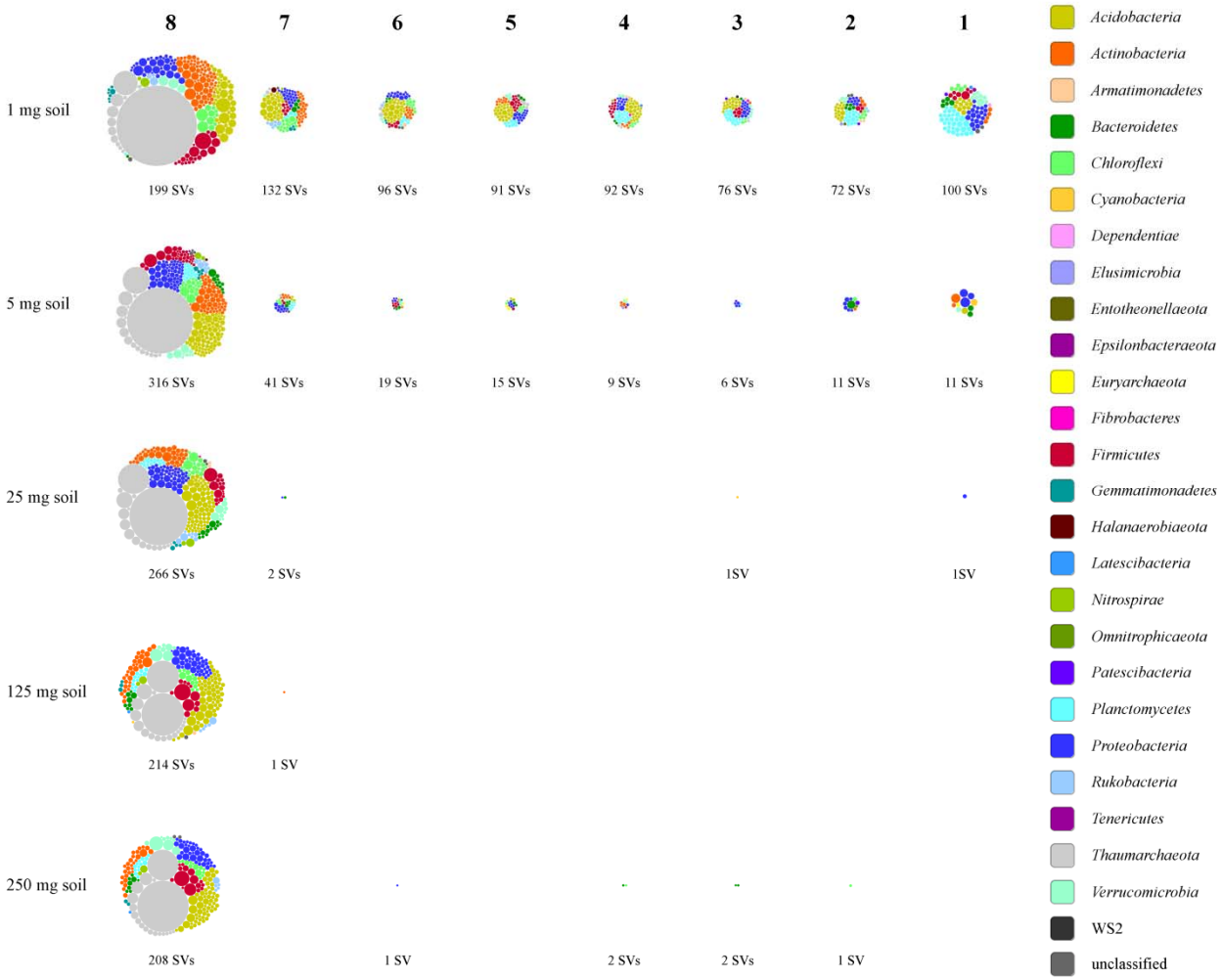
563 Figure 3



564

565

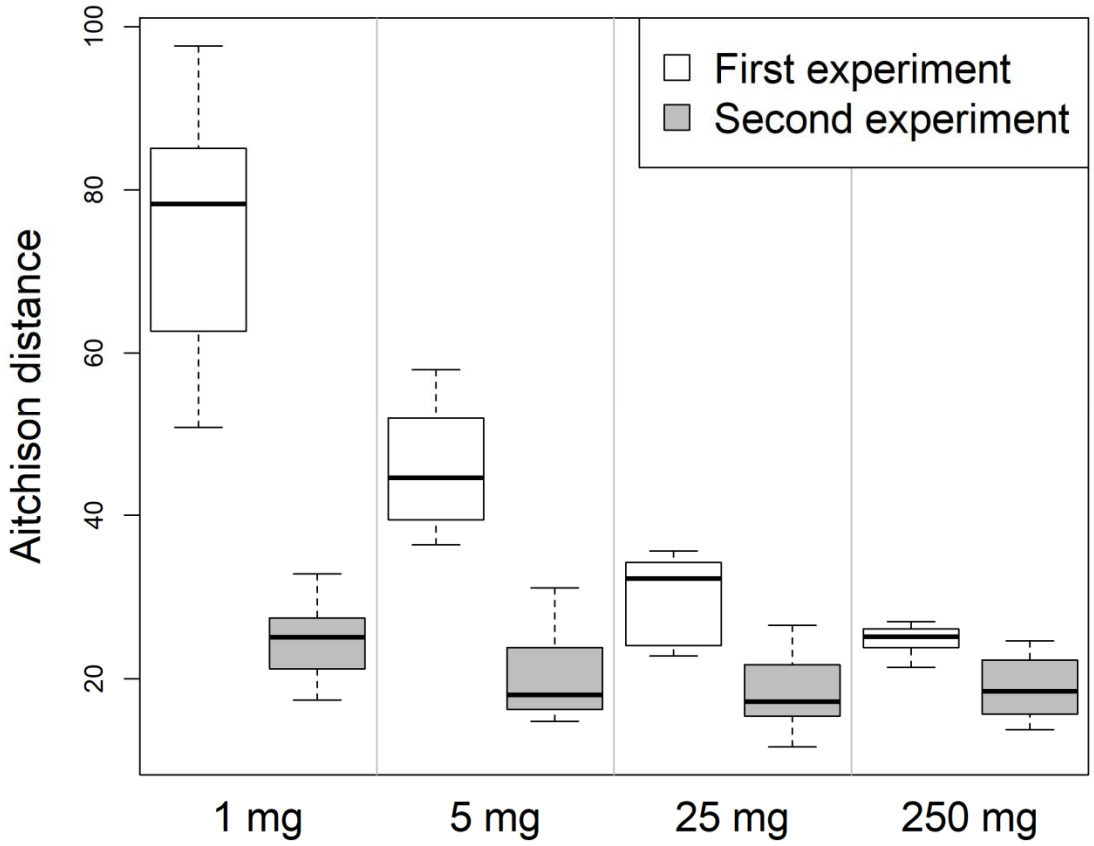
566 Figure 4



567

568

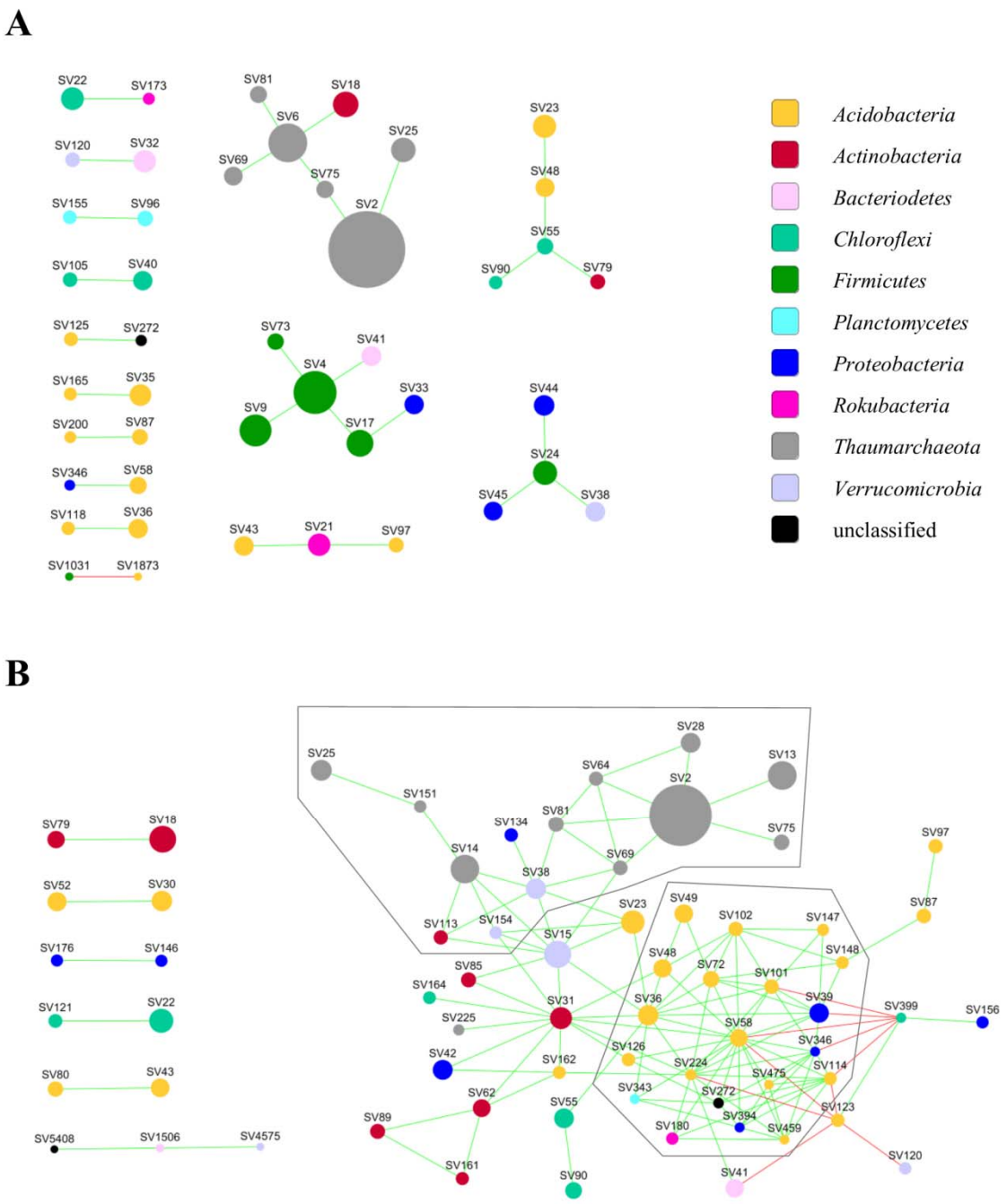
569 Figure 5



570

571

572 Figure 6



573