AsCoM: An ecosystem-specific reference database for increased taxonomic resolution in soil microbial profiling

Christina Karmisholt Overgaard

Aalborg University

Ke Tao

Aarhus University

Sha Zhang

Aarhus University

Bent Tolstrup Christensen

Aarhus University, AU-Foulum

Zuzana Blahovska

Aarhus University

Simona Radutoiu

Aarhus University

Simon Kelly

Aarhus University

Morten Kam Dahl Dueholm (✉ md@bio.aau.dk)

Aalborg University

Research Article

Keywords: Agricultural soil, microbiota, full-length 16S rRNA genes, AutoTax, host preference

Posted Date: February 24th, 2022

DOI: <https://doi.org/10.21203/rs.3.rs-1372773/v1>

License: © ⓘ This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

1 **AsCoM: An ecosystem-specific reference database for increased taxonomic resolution in soil**
2 **microbial profiling**

3

4 **Authors:** Christina Karmisholt Overgaard¹, Ke Tao², Sha Zhang², Bent Tolstrup Christensen³,
5 Zuzana Blahovska², Simona Radutoiu², Simon Kelly^{2,*}, Morten Kam Dahl Dueholm^{1,*}

6

7 **Affiliation:**

8 ¹ Center for Microbial Communities, Department of Chemistry and Bioscience, Aalborg University,
9 Aalborg, Denmark

10 ² Department of Molecular Biology and Genetics, Aarhus University, Aarhus, Denmark

11 ³ Department of Agroecology, Aarhus University, AU-Foulum, Tjele, Denmark

12

13 *Correspondence to: Morten Kam Dahl Dueholm, Center for Microbial Communities, Department
14 of Chemistry and Bioscience, Aalborg University, Fredrik Bajers Vej 7H, 9220 Aalborg, Denmark;
15 Phone: (+45) 9940 8508; Fax: Not available; E-mail: md@bio.aau.dk or Simon Kelly, Department of
16 Molecular Biology and Genetics, Aarhus University, Gustav Wieds Vej 10, 8000 Aarhus C,
17 Denmark; Phone: Fax: Not available; E-mail: kelly@mbg.au.dk

18 **Abstract**

19 Background: Intensive agriculture systems have paved the way for a growing human population.
20 However, the abundant use of mineral fertilisers and pesticides may negatively impact nutrient cycles
21 and biodiversity. One potential alternative is to harness beneficial relationships between plants and
22 plant-associated rhizobacteria to increase nutrient-use efficiency and provide pathogen resistance.
23 Plant-associated microbiota profiling can be achieved using high-throughput 16S rRNA gene
24 amplicon sequencing. Interrogation of this data is limited by confident taxonomic classifications at
25 high taxonomic resolution (genus- or species-level), which is often not possible with the commonly
26 applied universal reference databases. The development of high-throughput full-length 16S rRNA
27 gene sequencing combined with automated taxonomy assignment (AutoTax) can be used to create
28 amplicon sequence variant resolved ecosystems-specific reference databases that are superior to the
29 traditional universal reference databases.

30 Results: Synthetic long-read sequencing of 987,353 full-length 16S rRNA genes was used to create
31 a reference database (AsCoM) with AutoTax for bacteria and archaea based on Askov and Cologne
32 soils. We evaluated the performance of AsCoM using short-read amplicon data and found that it
33 resulted in increased genus- and species-level classification compared to commonly used universal
34 reference databases. AsCoM was utilised to evaluate the ecosystem-specific primer bias and
35 taxonomic resolution of amplicon primers targeting the V5-V7 region of the 16S rRNA gene as these
36 are commonly used within the plant microbiome field. Finally, we demonstrate the benefits of
37 AsCoM through the analysis of V5-V7 amplicon data to identify new plant-associated microbes for
38 two legume (*Lotus japonicus* and *Medicago truncatula*) and two cereal (*Hordeum vulgare* and *Zea*
39 *mays*) species.

40 Conclusion: We presented AsCoM, the first comprehensive ecosystem-specific 16S rRNA reference
41 database for Askov and Cologne soil. AsCoM provides unprecedented taxonomic resolution in 16S
42 rRNA amplicon profiling of soil-grown plant microbiota. It also represents a valuable tool for
43 ecosystem-specific evaluation of amplicon primers and development of species-level hybridization
44 probes for *in situ* studies.

45

46 **Keywords:**

47 Agricultural soil, microbiota, full-length 16S rRNA genes, AutoTax, host preference.

48 **Background**

49 A growing world population necessitates continued improvements in agricultural output to ensure
50 food security. Current agricultural practices may involve excessive applications of mineral fertilisers
51 and pesticides, which need to be reduced and supplemented with more sustainable solutions [1]. One
52 proposed solution is to harness the capabilities of natural soil microbes [2–4]. In order to develop the
53 potential benefits of microbes, we require a deeper understanding of the interactions occurring
54 between plants and microbes at a community level within the complex soil environment.

55 Initial investigations of plant microbiomes largely relied on the cultivation of microbes, limiting these
56 investigations to the easily culturable fraction of the microbial communities [5]. However, advances
57 in DNA sequencing technologies and bioinformatic tools have allowed large-scale amplicon-based
58 microbiome studies to become commonplace, opening up for more comprehensive detection of the
59 microbes present. In terms of plant-microbe interactions, studies have utilised amplicon sequencing
60 to investigate the microbial diversity of soils retrieved from various ecosystems such as tallgrass
61 prairie, tundra, tropical rainforest, and agroecosystems [6–9]. Through profiling of root-associated
62 communities, key insights have been gained into the factors that govern assembly of the root
63 microbiome including soil type, root exudates, plant genotype, plant developmental state and season
64 [10–16]. The 16S rRNA V5-V7 primer pair [17–19] is often used for bacterial profiling in plant
65 studies, because amplicons originating from plant chloroplast and mitochondria can be minimised via
66 size selection and primer bias. Without the exclusion of host-derived amplicons, the fraction of
67 microbial amplicons is greatly diminished [20], resulting in a skewed picture of the microbial
68 community.

69 One of the major limiting factors in analysis of the data generated through amplicon based
70 microbiome studies is taxonomic assignment. In order to understand the biological roles of soil
71 microbes, we need to be able to identify them correctly and comprehensively. However, most bacteria
72 are only classified at the family-level or above with commonly applied reference databases [21]. This
73 is problematic as many important physiological traits only are conserved at the genus-level or below
74 [22].

75 In order to overcome this, ecosystem-specific databases can be created using a combination of high-
76 throughput full-length 16S rRNA gene sequencing and automated taxonomy assignment (AutoTax)
77 [21]. Such databases improve taxonomic classification of both short- and long-read amplicon

78 sequence variants (ASVs), providing increased taxonomic classification at the genus- and species-
79 level [21].

80 Here we introduce AsCoM, an ecosystem-specific database containing 18,174 ASV-resolved full-
81 length 16S rRNA genes (FL-ASVs) from bacteria and archaea from Askov and Cologne soils. The
82 Askov soils have been used for plant and soil studies for more than 125 years [23]. Cologne soil has
83 been extensively used in numerous plant-microbe interaction studies [15, 16, 24–26]. We use the
84 ecosystem-specific AsCoM database to (i) improve the classification of V5-V7 amplicons from
85 Askov soil, rhizosphere and endosphere at genus- and species-level, (ii) uncover ecosystem-specific
86 primer bias associated with the V5-V7 primer set, and (iii) investigate the host preference of bacterial
87 taxa in associations with two legume (*Lotus japonicus* and *Medicago truncatula*) and two cereal
88 (*Hordeum vulgare* and *Zea mays*) species.

89

90 **Methods:**

91 ***Soil and plant materials***

92 The Askov soil was obtained from the Askov Experimental Station situated in Southern Denmark
93 (GPS coordinates: 55.466 N, 9.117 E). The soil was sampled in 2018 from the 0-20 cm soil depth in
94 three replicate plots (plot no. 421, 443, and 474) in the B4-field. This field is part of the Askov Long-
95 Term Experiment established in 1894 and the plots sampled for this study have been kept without
96 manure and fertiliser application since then. The soil is a light sandy loam with 11% clay, 13% silt
97 and 76% sand and classifies as Alfisol (Typic Hapludalf, USDA Soil Taxonomy). The soil grows a
98 four-course rotation of winter wheat, silage maize, spring barley and a grass-clover mixture used for
99 cutting. Addition of lime every 4-5 year keeps soil pH in the range 5.5 to 6.5 [23]. Cologne soil was
100 obtained from the Max Planck Institute for Plant Breeding Research in Cologne, Germany (GPS
101 coordinates: 50.958 N, 6.856 E) and has not been in agricultural use for over 15 years [27, 28].

102 Seeds of *Lotus japonicus* Gifu (*Lj*) [29] and *Medicago truncatula* A17 (*Mt*) [30] were available from
103 our stocks at Aarhus University. *Hordeum vulgare* cv. Golden promise (*Hv*) and *Zea maise* cv. W22
104 (*Zm*) were received from the Crop Science Centre, University of Cambridge.

105 ***Plant setup and harvesting***

106 Seeds were surface sterilised in a 1:20 bleach solution for 5-15 min, washed five times in sterile H₂O
107 and germinated on wet filter paper for 2-5 d. Seedlings were aseptically transferred to sterile pots that
108 had been filled with ~350 g Askov soil (*Lj* and *Mt* 10 plants/pot; *Hv* and *Zm* 4 plants/pot). Three pots
109 were set up for each plant species. Plants were grown in a growth chamber under controlled
110 conditions: 16/8 h day/night, 75% humidity, 22°C / 18°C. Watering throughout the experiment was
111 with sterile H₂O only.

112 After 3 weeks of growth plants were removed from pots and roots separated from shoots. Root
113 material was vortexed for 30 sec in 50 mL Falcon tubes containing 35 mL sterile H₂O, and roots were
114 transferred to a new Falcon tube. The initial Falcon tube was centrifuged (4,000 x g, 15 min), and
115 most supernatant was removed. The soil pellet was resuspended in the remaining liquid (ca. 2 mL)
116 using a cut P1000 tip. 300 µl of this suspension was transferred to a Lysis matrix E tube (MP
117 Biomedicals) forming the rhizosphere fraction. The root material was washed three times in sterile
118 H₂O for 30 sec, once in detergent (1x TE + 0.1% Triton X-100) for 2 min, once in 80% ethanol for
119 30 sec, once in 3% bleach for 30 sec and finally five times in sterile H₂O. The surface sterilised root
120 material was then transferred to Lysis matrix E tubes forming the endosphere fraction. For *Lj* and *Mt*,
121 nodules were removed from sterilised root material and collected in separate Lysis matrix E tubes
122 forming the nodule fraction. Details about all samples can be found in **Data S1**.

123

124 ***General molecular methods and DNA extraction***

125 DNA was extracted using the FastDNA™ Spin Kit for Soil (MP Bio) with the homogenisation
126 performed using a Precellys tissue lyser (Bertin Instruments) for 2x30 sec at 6000 rpm or using the
127 RNeasy® PowerSoil® Total RNA Kit, RNeasy® with the PowerSoil® DNA Elution Kit (Qiagen). Quality
128 was determined using a genomic DNA ScreenTape and concentrations were measured using Qubit™
129 dsDNA HS Assay Kit.

130 Concentrations of DNA were measured using a Qubit 3.0 fluorometer (Thermo Fisher Scientific) and
131 quality was determined using Agilent 2200 TapeStation (Agilent Technologies). AMPure XP beads
132 were used for DNA cleanup in accordance with manufacturer's protocol except for the washing step
133 where 80% ethanol was used. All commercial kits were used according to manufacturer's protocol
134 unless otherwise stated.

135 ***Full-length 16S rRNA gene library preparation and sequencing***

136 Full-length 16S rRNA gene sequencing was carried out as previously described [31] with the addition
137 of an extra primer pair targeting archaeal 16S rRNA genes. Oligonucleotides used can be found in
138 **Table S1**. Sequencing libraries were prepared from the following sample types: Askov and Cologne
139 bulk soil and rhizosphere of *Lj*, *Mt*, *Hv*, and *Zm* grown in Askov soil for 3 weeks and *Lj* rhizosphere
140 grown in Cologne soil for 3 weeks (**Data S1**).

141
142 *Adaptor annealing by PCR*. Adaptors containing sample barcodes, unique molecular identifiers
143 (UMI) and defined primer binding sites were added to each end of the bacterial 16S rRNA genes by
144 PCR. The reaction contained 10 µL of 10x PCR Buffer (Qiagen), 2 µL of 10 mM dNTP (Qiagen), 5
145 µL of 10 µM f16S_pcr1_fw, 5 µL of 10 µM f16S_pcr1_rv, 4 µL (bacteria) or 8 µL (archaea) of 25
146 mM MgCl₂, 0 µL (bacteria) or 20 µL (archaea) of 5x Q-solution (Qiagen), 0.5 µL of 5 U/µL Taq
147 polymerase (Qiagen), 100 ng of template DNA, and nuclease-free water to 100 µL. The reaction was
148 incubated with an initial denaturation at 94°C for 3 min followed by 2 cycles of denaturation at 94°C
149 for 30 sec, annealing at 56°C (bacteria) or 54°C (archaea) for 30 sec, and extension at 72°C for 3 min,
150 and then a final extension at 72°C for 5 min. The sample was purified using 0.6x AMPure XP beads
151 and eluted in 21 µL nuclease-free water.

152
153 The following f16S_pcr1_fw forward primers were used: Askov rhizosphere (fw1), Askov bulk soil
154 (fw2), Cologne rhizosphere (fw3), Cologne bulk soil (fw4). The f16S_pcr1_rv1 reverse primer was
155 used for all samples.

156
157 *Library amplification*. The adaptor annealed 16S rRNA gene amplicons were amplified using PCR
158 to obtain enough product for quantification and sequencing. The reaction contained 19 µL of adaptor
159 annealed sample, 20 µL of 5X Phusion HF (New England Biolabs), 2 µL of 10 mM dNTP, 5 µL of
160 10 µM f16S_pcr2_fw, 5 µL of 10 µM f16S_pcr2_rv, 4 µL of 25 mM MgCl₂, 44 µL nuclease-free
161 water, and 1 µL 2U/µL Phusion HF DNA polymerase (NEB). The reaction was incubated with an
162 initial denaturation at 98°C for 30 sec followed by 15 (bacteria) or 20 (archaea) cycles of denaturation
163 at 98°C for 10 sec, annealing at 62°C for 30 sec, and extension at 72°C for 1 min, followed by a final
164 extension at 72°C for 5 min. DNA was eluted in 11 µL nuclease-free water. Quality was determined
165 using a D5000 ScreenTape and concentrations were measured using Qubit™ dsDNA HS Assay Kit.

166 Libraries for the different sample barcodes were pooled with an equal amount of DNA from each
167 sample.

168

169 *Clonal library amplification.* Adaptor annealed amplicon libraries were diluted to approximately
170 200,000 molecules/ μ L and amplified by PCR to obtain clonal copies of each uniquely tagged
171 amplicon molecule. The PCR reaction contained 63.5 μ L nuclease-free water, 10 μ L of 10X PCR
172 buffer (Qiagen), 2 μ L of 10 mM dNTP, 5 μ L of 10 μ M f16S_pcr2_fw , 5 μ L of 10 μ M f16S_pcr2_rv
173 , 4 μ L of 25 mM MgCl₂, 0.5 μ L of 5 U/ μ L Taq polymerase (Qiagen) and 10 μ L diluted adaptor
174 annealed amplicon product. The reaction was initiated by denaturation at 94°C for 3 min, followed
175 by 20 cycles of denaturation at 94°C for 30 sec, annealing at 62°C for 30 sec and extension at 72°C
176 for 2 min, followed by a final extension at 72°C for 5 min. The PCR product was purified using 0.6x
177 AMPure XP beads with elution into 21 μ L nuclease-free water. The product quality and concentration
178 were analysed on a D5000 screentape and with the Qubit dsDNA HS Assay Kit, respectively.

179

180 *Read-tag library preparation.* A Nextera library preparation kit (Illumina) was used to prepare a
181 paired-end read-tag sequencing library from the clonal library using a customised protocol. A
182 tagmentation reaction was prepared with 100 ng of the clonal library in 22.5 μ L nuclease-free water,
183 25 μ L tagment DNA buffer (Illumina) and 2.5 μ L tagment DNA enzyme (Illumina). The reaction
184 was incubated at 55°C for 5 min. The product was immediately diluted to 100 μ L and purified using
185 0.6x AMPure XP beads with elution into 42 μ L nuclease-free water.

186 The tagmentation products were PCR amplified using two separate PCRs (A and B). PCR A
187 selectively amplified fragments containing the 5' termini of the amplicons and PCR B selectively
188 amplified fragments containing the 3' termini. The reactions contained 20 μ L purified tagmentation
189 product, 5 μ L N504 nextera adaptor (Illumina), 5 μ L of 10 μ M f16S_readtag_fw (PCR A) or
190 f16S_readtag_rv (PCR B), 5 μ L PCR primer cocktail (Illumina), 10 μ L of 5x Phusion HF buffer
191 (NEB), 1 μ L of 10 mM dNTP, 3.5 μ L nuclease-free water, and 0.5 μ L of 2U/ μ L Phusion HF DNA
192 polymerase (NEB). The following PCR program was used: Initial elongation at 72°C for 3 min, initial
193 denaturation at 98°C for 30 sec, and 10 cycles of denaturation at 98°C for 10 sec, annealing at 60°C
194 for 30 sec and elongation at 72°C for 3 min, followed by a final extension at 72°C for 5 min. The raw
195 read-tag libraries were purified using 1.0x AMPure XP beads with elution into 21 μ L nuclease-free
196 water.

197 To ensure even sequencing coverage across the length of the 16S rRNA gene amplicons, the size
198 distribution of the read-tag libraries was optimised [32]. The libraries were size fractionated on an E-
199 Gel CloneWell gel (Thermo Fisher Scientific). A total of 500 ng GeneRuler 1 kb DNA ladder
200 (Thermo Fisher Scientific) was used as a length reference. The gel was run until the 500 bp marker
201 was 1 mm from the elution well, after which 20 µL elution aliquots were sampled and replaced by
202 nuclease-free water every 15 seconds, up to a total of 32 aliquots. Every two aliquots were pooled,
203 yielding 16 pooled aliquots per sample. These were then analysed on an Agilent 2200 TapeStation
204 using the High Sensitivity D1000 Screentape. Fractions with a mean fragment length of 500-1250 bp
205 were used for the pooling. The effective sequencing concentration for fractions from 500-950 bp was
206 determined based on the tapestation data and the empirical formula $C_{seq} = \text{Peak molarity [pmol/l]} * (-$
207 $0.0124 * (\text{peak size [bp]} - 215 \text{ bp}) + 10.332)$ [32]. These fractions were pooled with equal effective
208 sequencing concentration (C_{seq}). For fractions between 950-1250 bp the entire aliquot was used for
209 pooling (40-50 µL). The pooled aliquots were then purified using 1.0x AMPure XP beads with elution
210 into 11 µL nuclease-free water. The quality and concentration of the coverage optimised read-tag
211 libraries were analysed on D1000 screen tapes and with the Qubit dsDNA HS Assay Kit, respectively.

212 *Linked-tag library preparation.* Clonal libraries were end-repaired in a reaction containing 20 ng
213 clonal library, 2.5 µL of 10x NEBNext End Repair Reaction Buffer (New England Biolabs), 1.25 µL
214 NEBNext End Repair Enzyme Mix (New England Biolabs), and nuclease-free water to 25 µL. The
215 reaction was incubated at 20°C for 30 min. The end-repair reaction was purified using 1.0x AMPure
216 XP beads and eluted into 10 µL nuclease-free water.

217 The end-repaired sample was circularised in an intra molecular blunt end ligation reaction containing
218 150 µL nuclease-free water, 20 µL of 50% (w/w) PEG 4000 solution (Thermo Fisher Scientific), 20
219 µL of 10x T4 DNA ligase buffer (NEB), 8 µL of T4 DNA ligase (NEB), and 2 µL of end-repaired
220 clonal library. The reaction was incubated at 16°C for 60 min. The circularised products were purified
221 using 1.0 x AMPure XP beads and eluted in 10 µL nuclease-free water.

222 The junction sequence, which contains both UMI tags, were amplified by PCR in a reaction
223 containing 8 µL of circularized clonal library, 5 µL of 10x PCR buffer (Qiagen), 1 µL of 10 mM
224 dNTP mix, 2.5 µL of 10 µM f16S_linktag_fw, 2.5 µL of f16S_linktag_rv, 2 µL (bacteria) or 4 µL
225 (archaea) 25 mM MgCl₂, 0 µL (bacteria) or 10 µL (archaea) 5x Q-solution (Qiagen), 0.25 µL 5u/µL
226 Taq polymerase (Qiagen), and nuclease-free water to 50 µL. The PCR reaction was initiated by
227 denaturation at 94°C for 3 min, followed by 20 cycles of denaturation at 94°C for 20 sec, annealing

228 at 56°C (bacteria) or 54°C (archaea) for 20 sec, and extension at 72°C for 20 sec, followed by a final
229 extension at 72°C for 3 min. The PCR product was purified using 1.0x AMPure XP beads and elution
230 into 12 µL nuclease-free water. The quality and concentration of the linked-tag libraries were
231 analysed on D1000 screentape and with the Qubit dsDNA HS Assay Kit, respectively.

232
233 *Library pooling.* The coverage normalised read-tag library A and B were diluted to 0.9 ng/µL. The
234 linked-tag library was diluted to 0.2 ng/µL. The libraries were pooled by combining 4.6 µL read-tag
235 library A, 4.6 µL read-tag library B, and 0.8 µL linked-tag library.

236
237 *Sequencing.* The libraries were paired-end (1 x 240 bp and 1 x 25 bp) sequenced on a HiSeq 2500
238 instrument (Illumina) using on-board clustering and rapid run mode with a HiSeq PE Rapid Cluster
239 Kit v2 (Illumina) and HiSeq Rapid SBS Kit v2, 200 cycles (Illumina). The SBS reagents were
240 supplemented with 9.5 mL Incorporation Master Mix, 9.5 mL Cleavage Reagent Mix and 7 mL
241 Universal Scan Mix to enable sequencing of 265 cycles. The HiSeq was running HiSeq Control
242 Software v2.2.68 (Illumina) and Real Time analysis v1.18.66.3 (Illumina). The libraries were
243 prepared and loaded on the HiSeq using the standard procedures (Illumina: manual # 15035786 v01;
244 manual # 15050107 v02; manual # 15061846 v01) with the following changes. A volume of 10 µL
245 library pool was denatured by adding 10 µL of 0.1 N NaOH solution, mixing well by pipetting and
246 incubating for 5 min at 25°C. The denatured library pool was diluted by adding 980 µL of cold
247 Hybridisation Buffer (Illumina). 400 µL of the denatured and diluted library pool was mixed with 20
248 µL of denatured and diluted 10 pM PhiX control v3 library (Illumina), and stored on ice until loading.
249 Custom read2 primer mix was prepared by mixing 25 µL of 100 µM f16S_read2_fw and 25 µL of
250 100 µM f16S_read2_rv in a conical tube (15 mL) and diluting with 4,950 µL Hybridization Buffer
251 (final concentration 0.5 µM). When the paired-end reagent rack was loaded on the HiSeq, the Illumina
252 primer mix in position nr. 16 was replaced with the custom read2 primer mix prepared above. When
253 setting up the HiSeq run in the control software, the standard procedure was followed except for the
254 following steps: For the “Recipe Screen”, the following options were chosen: Index type options =
255 No Index. Read 1 cycles = 240. Read 2 cycles = 25. After sequencing, bcl2fastq v2.17.1.14 (Illumina)
256 was used to generate fastq files from bcl files using standard settings (manual # 15038058 RevB).

257 ***Assembly of full-length 16S rRNA genes and generation of the AsCoM database***

258 Raw sequence reads were binned, based on the UMIs, and *de novo* assembled into the synthetic long-
259 read rRNA gene sequences as previously described [21]. The assembled 16S rRNA gene sequences
260 were oriented based on the SILVA 138 SSURef Nr99 database using the usearch v.11.0.667 -orient
261 command and trimmed between the 27f and 1391r primer binding sites (bacteria) or SSU1ArF and
262 SSU1000ArR primer binding sites (archaea) using the trimming function in CLC genomics
263 workbench v. 20.0. Sequences without both primer binding sites were discarded. The trimmed full-
264 length 16S rRNA genes were processed with AutoTax v. 1.5.2 [21] to create the full-length 16S rRNA
265 gene amplicon sequence variants (FL-ASV) resolved AsCoM reference database and taxonomy.

266

267 ***Construction of phylogenetic trees***

268 FL-ASV aligned to the SILVA_138.1_SSURef_NR99 ARB database (AutoTax: temp/FL-
269 ASVs_SILVA_aln.fa) were loaded into ARB. Taxonomy was added to the sequences (AutoTax:
270 output/tax_complete.csv). Before importing the taxonomy was concatenated in excel. All bacterial or
271 archaeal sequences were hereafter exported from ARB as a fasta alignment using the positional
272 variability by parsimony filter ssuref:bacteria or ssuref:archaea. Phylogenetic trees were constructed
273 using FastTree v.2.1.11 [33] with the -gtr and -nt options.

274 ***16S rRNA V5-V7 amplicon sequencing***

275 A two-step PCR protocol was used to amplify the 16S rRNA V5-V7 region and index samples using
276 the 799F (5'-AACMGGATTAGATACCCKG-3') and 1192R (5'-ACGTCATCCCCACCTTCC-3')
277 primers as described previously [26]. Paired-end 350 bp sequencing was performed by IMGM
278 Laboratories GmbH (www.imgm.com) on an Illumina MiSeq platform.

279

280 ***Processing of short-read amplicon data***

281 16S rRNA gene V5-V7 forward and reverse reads were processed using usearch v.11.0.667 [34].
282 Forward and reverse reads were merged using the usearch -fastq_mergepairs command, filtered to
283 remove phiX sequences using usearch -filter_phix, and quality filtered using usearch -fastq_filter
284 with -fastq_maxee 1.0. Dereplication was performed using -fastx_uniques with -sizeout, and
285 amplicon sequence variants (ASVs) were resolved using the usearch -unoise3 command. An ASV
286 table was created by mapping the quality filtered reads to the ASVs using the usearch -otutab
287 command with the -zotus and -strand plus options. Taxonomy was assigned to ASVs using the
288 AsCoM database as well as universal reference databases and the usearch -sintax command with -

289 strand both and -sintax_cutoff 0.8 options. Mapping of ASVs to reference databases was done with
290 the -usearch_global command and the -id 0, -maxaccepts 0, -maxrejects 0, -top_hit_only, and -strand
291 plus options unless otherwise stated.

292

293 *Evaluation of ecosystem-specific primer bias*

294 The phylogenetic signal of V5-V7 amplicons was evaluated based on *in silico* ASVs extracted from
295 the aligned FL-ASVs in ARB. FL-ASVs were trimmed in accordance to the base pair positions after
296 the end of the forward primer and before the start of the reverse primer. A fasta file containing the
297 redundant set of *in silico* ASVs were classified using the AsCoM database using usearch -sintax with
298 -strand plus and -sintax_cutoff 0.8. The classifications were compared to those of the parental FL-
299 ASVs in the AsCoM database.

300

301 The ecosystem-specific primer bias associated with the V5-V7 primer bias was evaluated using the
302 bacterial FL-ASVs in the AsCoM database and the *analyze_primers.py* script from Primer Prospector
303 v. 1.0.1 [35]. The specificity of primer sets was defined based on the overall weighted scores (OWSs)
304 for the primer with the highest score as follows: perfect hit (OWS is 0), partial hit (OWS is >0 and
305 ≤1), and poor hit (OWS is >1). The scores are based on the following criteria: The last five bases are
306 considered the 3', non 3' mismatch penalty = 0.40 per mismatch, 3' mismatch penalty = 1.00 per
307 mismatch, last base mismatch penalty = 3.00, non 3' gap penalty = 1.00 per gap, and 3' gap penalty =
308 3.00 per gap.

309 *Availability of data and code*

310 All sequencing data have been submitted to the Sequence Read Archive under the project ID
311 PRJNA787301. Details about individual datasets can be found in **Data S1**. Data were analyzed with
312 R v.4.0.5 [36] through RStudio IDE [37], with the tidyverse v.1.3.1 [38], vegan v. 2.5.7 [39], Ampvis2
313 v.2.7.9 [40], patchwork v. 1.1.1. [41], and ggtree v. 3.1.1.991 [42] packages. All R scripts used for
314 data analysis and visualisation as well as the AsCoM reference database in SINTAX and QIIME
315 format are available at Github:
316 <https://github.com/msdueholm/Publications/tree/master/Overgaard2022a>.

317 Results

318 *Establishment of AsCoM, an ecosystem-specific reference database*

319 To create an ecosystem-specific reference database applicable for both Askov and Cologne soils, we
320 applied synthetic long-read sequencing to obtain near full-length bacterial and archaeal 16S rRNA
321 genes using DNA from bulk soil and rhizosphere samples of plants grown in the two soils.
322 Approximately the same number of sequences were obtained for each soil, resulting in a total of
323 987,353 full-length 16S rRNA gene sequences after primer and quality filtering. These sequences
324 were subsequently processed using AutoTax to resolve full-length amplicon sequence variants (FL-
325 ASVs) and create a comprehensive seven-rank taxonomy for all sequences. AutoTax incorporates *de*
326 *novo* placeholder names (AsCoM_x_y) for the many uncultured environmental taxa, which are not
327 yet taxonomically classified at all seven-ranks [21]. The result of this is a comprehensive taxonomy
328 that can be interrogated at even the lowest taxonomic ranks (species-level). This is not possible when
329 using conventional reference databases to assign taxonomy due to missing taxonomic assignments,
330 particularly in the lower ranks. The final AsCoM database contained 18,042 unique FL-ASVs, each
331 with a complete seven-rank taxonomy assigned.

333 *AsCoM reveals great taxonomic novelty within the soil community*

334 The sequence-based novelty of the AsCoM reference sequences was evaluated by mapping the FL-
335 ASVs to the SILVA_138.1_SSURef_NR99 database [43] and determining the percentage of FL-
336 ASVs that have reference sequences within the identity thresholds for each taxonomic rank suggested
337 by [44] (**Figure 1** and **Table 1**). The vast majority of the FL-ASVs had homologs within SILVA at
338 the genus-level ($\geq 94.5\%$ identity) for both archaea (97%) and bacteria (94%). However, 38% of the
339 bacterial and 14% of the archaeal FL-ASVs lacked references in SILVA at the species-level ($\geq 98.7\%$
340 identity), indicating important novelty within AsCoM at this taxonomic level.

341 We then investigated the taxonomic novelty within the AsCoM database based on the percentage of
342 AutoTax assigned placeholder names (AsCoM_x_y) at the different taxonomic ranks (**Figure 1** and
343 **Table 1**). For archaea the majority of taxa were assigned *de novo* placeholder names at the genus-
344 and species-level revealing a significant lack of taxonomic information in the lower ranks within the
345 SILVA database. However, the taxonomic diversity within archaea was very low with 112 out of the
346 132 archaeal FL-ASVs were assigned to a single family, the *Nitrososphaeraceae*. For bacteria almost
347 97% of all species, 81% of genera, 61% of families and 30% of orders were assigned *de novo*

placeholder names, revealing a great diversity of undescribed taxa within this ecosystem (**Figure 1**). This was especially apparent within the phyla of *Bacteroidota*, *Chloroflexi*, *Myxococcota*, *Patescibacteria*, *Planctomycetota* and *Proteobacteria* (**Figure 1** and **Figure S1**). The assignment of placeholder taxonomies to these sequences is essential for permitting informative community profiling, because these taxa would otherwise be excluded from taxonomy-resolved studies. The benefits of this are most apparent at the lower taxonomic ranks, which is of biological importance given that species-level assignment can be required to distinguish between, e.g., beneficial and pathogenic microbes [45–47].

AsCoM provides improved references for plant microbiome samples

Precise and comprehensive taxonomic classification requires a database with high-identity reference sequences for microbes within the ecosystem. To evaluate the coverage of the AsCoM database, we mapped 16S rRNA gene V5-V7 ASV data obtained from three different compartments (bulk soil, rhizosphere, and endosphere) recovered from *Hordeum vulgare*, *Zea mays*, *Medicago truncatula*, and *Lotus japonicus* grown in Askov soil to the AsCoM database as well as commonly applied universal reference databases and calculated the percentage of ASVs with high-identity (>99%) references in the databases (**Figure 2**).

For bulk soil and rhizosphere samples the AsCoM database (18,174 seq.) has more high-identity reference sequences (>99% identity) than the much larger universal databases GreenGenes 16S v.13.5 (1,262,986 seq.) [48], GTDB release 89 (145,904 seq.) [49], RDP 16S v16 (13,212 seq.) [50], and SILVA_138.1_SSURef_Nr99 (510,508 seq.) [43] (**Figure 2**). SILVA and GreenGenes were the best of the universal databases, likely reflecting the large number of reference sequences in those databases. For samples derived from the endosphere the SILVA and GreenGenes databases contained a slightly higher proportion of high-identity reference sequences compared to AsCoM (**Figure 2**). This likely reflects a historical focus on cultivation and isolation of bacteria from the endosphere, which have led to a high number of relevant reference sequences in the universal reference databases. A better coverage of these bacteria in AsCoM can be achieved by incorporation of relevant sequences from SILVA.

AsCoM improves taxonomic classification of ASVs from soil and plant compartments

To evaluate the taxonomic classification performance of AsCoM, we classified the V5-V7 ASVs using the SINTAX-classifier and compared the classification at the genus- and species-level with

378 those of commonly used universal databases (**Figure 3**). The AsCoM database was superior at
379 taxonomic classification both at genus- and species-level (**Figure 3**).

380 A large proportion (up to 85%) of the ASVs from bulk soil and rhizosphere samples were classified
381 at the genus-level using AsCoM, representing a significant improvement over SILVA (up to 40%).
382 The 15% of ASVs that were not assigned genus-level taxonomy by AsCoM can be explained by the
383 relatively low phylogenetic signal contained within the V5-V7 amplicons [21]. As we will see later,
384 especially members of the families *Comamonadaceae* and *Oxalobacteraceae* were hard to
385 differentiate at the genus-level based on V5-V7 amplicons (**Figure 5b**). Further analysis of the
386 taxonomic assignments by AsCoM and SILVA 138.1 demonstrated that the AsCoM database
387 provided better classifications for the 50 most abundant rhizosphere ASVs across the four plant
388 species, notably providing species-level classification for the three most abundant ASVs (**Figure 4**).
389 This serves to highlight that the improved performance of AsCoM over universal databases includes
390 abundant taxa that are likely to be of biological importance. Furthermore, these three abundant ASVs
391 were assigned *de novo* placeholder species names by AutoTax, indicating that these ASVs are missing
392 a comprehensive taxonomy in SILVA and downstream analysis of these abundant isolates at lower
393 taxonomic ranks would not be possible without the use of AsCoM.

394 Classification of the 50 most abundant ASVs in the endosphere across all four plant species was
395 slightly improved at the genus-level using the SILVA database compared to AsCoM (**Figure S2**).
396 However, AsCoM was able to provide species-level resolution for seven of these ASVs, which was
397 not possible for any of the ASVs with SILVA. Most of the ASVs that could not be classified at the
398 genus-level were from a few orders with many cultivated representatives, namely *Burkholderiales*,
399 *Rhizobiales*, and *Enterobacteriales*. The majority of the poorly classified ASVs represent isolates of
400 lower abundance identified in the endosphere of barley and/or maize. This suggests that additional
401 FL-ASV sequencing, specifically of DNA recovered from endosphere samples would improve
402 coverage of the AsCoM database in a later release.

403 So far, the capacities of AsCoM were evaluated when investigating the taxonomic assignment of V5-
404 V7 amplicons. Next, we aimed to gain insight into how the AsCoM database performs when
405 comparing the taxonomic assignment of a fully sequenced culture collection (LjSPHERE) containing
406 bacterial isolates recovered from *Lotus japonicus* grown in Cologne soil [26]. We analysed the full-
407 length 16S rRNA sequences and compared classification based on this marker gene of the isolates
408 with the AsCoM versus SILVA 138.1 database. Both databases were able to classify all isolates down

409 to family level, and most isolates (~83%) at the genus level (**Table S2**). No species level assignment
410 was possible using the SILVA database, while AsCoM was able to provide species level assignment
411 to ~5% of the isolates (15 out of 277). These results indicate that the AsCoM database indeed provides
412 equal or better classification of both short-read amplicon and full-length 16S rRNA sequences.

413 ***Taxonomic resolution and primer bias related to 16S rRNA V5-V7 amplicon sequencing***

414 Despite the increasing interest and development of protocols for full-length 16S rRNA gene
415 sequencing [51–53] and metagenomic approaches [54–56], short-read amplicon analysis remains
416 popular for studying microbial communities due to the comparative low sequencing costs and high
417 accuracy. Drawbacks of the short-read approach are the lower phylogenetic signal provided by short
418 reads and the amplification efficiency bias that primer pairs can exhibit towards certain taxa [57].

419 Taking advantage of the AsCoM database, we evaluated the taxonomic resolution provided by V5-
420 V7 ASVs, specifically in regards to the microbial diversity found within Askov and Cologne soil. *In*
421 *silico* ASVs corresponding to the V5-V7 region were extracted from the FL-ASV in AsCoM. These
422 ASVs were then classified using AsCoM using the SINTAX-classifier. Genus- and species-level
423 classification of the *in silico* ASVs were compared to the taxonomy of their corresponding FL-ASVs,
424 and the fraction of correctly or incorrectly classified ASVs was calculated (**Figure 5a**). We found
425 that 93.5% of the *in silico* ASVs were correctly classified at the genus-level, 6.5% were not classified,
426 and only 0.07% wrongly classified. Species-level classification was correctly provided for 55.1% of
427 the ASVs, 44.8% could not be classified, and only 0.2% received wrong classifications. This confirms
428 the high precision of classifications obtained using the AsCoM database.

429 In plant microbiome studies the V5-V7 primer pair is a popular choice as plant-derived amplicons
430 can be excluded from the sequencing pool. However, this feature was found to lead to a lower
431 coverage for bacteria compared to primers targeting the V3-V4 and the V4 region of the 16S rRNA
432 gene when evaluated against the complete bacterial diversity in the SILVA database [20]. Because
433 the AsCoM database only includes reference sequences from agricultural soil it can be used to
434 evaluate the ecosystem-specific primer coverage and determine the bias towards amplification of
435 certain bacterial groups resulting in underestimating the abundance or absence of certain bacterial
436 groups within the bacterial community studied (**Figure 5c**). The *in silico* primer evaluation showed
437 that the V5-V7 primer pair is poorly suited for the study of many *Cyanobacteria*, *Verrucomicrobiota*,
438 *Planctomycetota*, *Chloroflexi* and *Patescibacteria* due to partial or poor primer matches. The primer

pair does however provide perfect matches for most members of the *Proteobacteria* groups, especially the well-known plant-associated *Burkholderiales* and *Rhizobiales* (**Figure 5c**).

Host preference of legumes and cereals

The AsCoM database allowed us to investigate signatures of host preference in the microbiota associated with the roots of the model legumes *Lotus japonicus*, *Medicago truncatula* and the cereal crops *Hordeum vulgare* and *Zea mays*. To allow for a direct comparison between the host plants, all were grown concurrently in Askov unfertilised soil under controlled conditions.

Following three weeks growth rhizosphere and root endosphere compartments were harvested. For legumes, nodules were separated from the root to form an additional compartment for analysis. 16S rRNA gene amplicon sequencing was performed using the V5-V7 primers to minimise host organelle amplicon contamination and taxonomic assignment was performed using the AsCoM database. Alpha-diversity analysis showed reduced bacterial diversity within the endosphere compared to rhizosphere with a more severe reduction in diversity observed for legume plants compared to cereals (**Figure S3a**). A larger variance was observed for the microbial diversity (inverse Simpsons) than for the richness (unique ASVs) between biological replicates for all soil samples. This reflects a heterogeneous microbial composition within the soil.

Beta-diversity analysis of the complete dataset revealed separation of the samples primarily due to the compartments from which they were harvested with exception of legume endosphere and nodule samples which showed some overlap (**Figure S3b**). Separation based on plant species was evident for the endosphere and nodule compartments, whereas all rhizosphere samples clustered closely with the bulk soil.

To investigate the preference of the host plants for particular ASVs within their rhizosphere and endosphere communities these compartments were further investigated separately. Both rhizosphere and endosphere communities show distinct clustering on the basis of the host plant, except for the cereal endosphere samples which clustered together (**Figure 6a**). To gain additional information about the plant-specific microbiome, we took advantage of the improved taxonomic resolution afforded by the AsCoM database and investigated the microbial community structure of the different compartments at the genus-level (**Figure 6b**). Host preference for specific rhizosphere microbes were evident, such as *Serratia* for *H. vulgare* and *Paraburkholderia* for *Z. mays*. Microbes with a preference for legumes over cereals were also identified, e.g., *Methylothermobacter*. Abundant rhizosphere

genera with weak or no host preference according to comparable abundance across all hosts and bulk soil, included AsCoM_g_8878, AsCoM_g_179 and *Nitrospira*. Within the endosphere, host preferences were stronger than observed in the rhizosphere. The legume endospheres were dominated by symbiotic microbes, *Mesorhizobium* for *L. japonicus*, and *Allorhizobium* and *Ensifer* for *M. truncatula* (**Figure 6b**). The cereals displayed more diversity within their endospheres. *Clostridium* is preferentially abundant in *H. vulgare*, while *Ralstonia*, *Serratia* and *Paraburkholderia* are most abundant in *Z. mays*. *H. vulgare* and *Z. mays* also shared some endophytic microbes including *Pseudomonas* and *Janthinobacterium*.

The clear enrichment of compatible symbiotic rhizobia within legume endospheres is consistent with previous studies [16, 58]. To gain insight into rhizobial colonisation of cereals compared to legume under the same experimental conditions, we investigated taxa annotated as belonging to the order *Rhizobiales*. Although the relative abundance of rhizobial taxa within cereals is greatly reduced compared to legumes their presence is detected, including the symbionts identified from within the legume endospheres and nodules (**Figure S4**).

Different species or strains within the same genus can have different host-preferences [59, 60]. Therefore, we also investigated host-preference at the ASV-level, which represents the highest phylogenetic resolution afforded by the amplicon data. We identified specific ASVs that were significantly enriched or reduced in the rhizosphere of the four plant species compared to bulk soil (**Figure S5** and **Data S2**). We found that for all plants, more ASVs were enriched than reduced in the rhizosphere compared to the bulk soil, and a higher number of ASVs were enriched for the legumes than for the cereal crops (1275 versus 366, respectively) (**Figure S5**). This suggests that more bacteria are recruited by the legumes compared to the cereals. A large proportion of the enriched ASVs were classified at the species- (27-35%) and genus-level (67-80%) with AsCoM, again highlighting the unique taxonomic resolution afforded by this database. The enriched ASVs were not taxonomically restricted but covered more than 100 different families. However, the majority of the ASVs were affiliated to families that includes known rhizobacteria, including *Comamonadaceae* (219 enriched ASVs across all plant species), *Oxalobacteraceae* (160 ASVs), *Roseiflexaceae* (119 ASVs), *Flavobacteriaceae* (95 ASVs), and *Rhizobiaceae* (80 ASVs) (**Data S2**). Interestingly, at the genus-level we found that most enriched ASVs were classified as the *de novo* genus AsCoM_g_8878 (113 ASVs) belonging to the family *Roseiflexaceae*. ASVs from this genus were enriched in the rhizosphere of all plant species except *H. vulgare* and should therefore be a target for further

500 investigations. Other well-represented genera included *Flavobacterium* (95 ASVs), *Allorhizobium-*
501 *Neorhizobium-Pararhizobium-Rhizobium* (48 ASVs), *Rhizobacter* (37 ASVs), and *Devosia* (36
502 ASVs).

503 Discussion

504 The development and use of ecosystems-specified 16S rRNA reference databases, such as AsCoM,
505 enables increased taxonomic resolution in amplicon sequencing studies and provides a taxonomic
506 framework for comparing data obtained using primer pairs targeting different regions of the 16S
507 rRNA gene. A major contribution to the improved classification of ASVs at the genus- and species-
508 level is the *de novo* placeholder taxonomy created in the reference database with AutoTax for
509 environmental lineages that do not have any official taxonomy at lower ranks [21]. Because the *de*
510 *novo* taxonomy is assigned based on fixed identity thresholds for each taxonomic rank, it does not
511 take different rates of evolution across the tree of life into account. It should, therefore, not be
512 considered a replacement for proper phylogenetic classification, which ideally requires phylogenomic
513 analyses [49, 61–63]. However, the placeholder taxonomy allows us to identify new ecological
514 relevant lineages that should be targeted for the recovery of high-quality metagenome-assembled
515 genome (MAGs) and further investigations [64, 65]. A good example is the *de novo* genus
516 AsCoM_g_8878 within the family *Roseiflexaceae* for which many ASVs displayed strong host-
517 enrichment.

518 A unique feature of the ecosystem-specific databases created with AutoTax is that they are essentially
519 chimera- and error-free. The attachment of unique molecular identifiers (UMIs) to each end of the
520 original 16S rRNA gene template molecule before any PCR amplification steps allows filtering of
521 true biological sequences from chimera already in the synthetic long-read assembly [21, 32]. The few
522 sequencing errors that may occur after this initial quality control are all low abundant and removed
523 when FL-ASVs are subsequently resolved [21, 66]. The extreme quality of the reference database
524 makes it ideal for lineage-specific probe design. The development of species-specific fluorescent *in*
525 *situ* hybridisation (FISH) probes provides opportunities to visualise the morphology and spatial
526 arrangement of individual species in complex samples, and can be combined with Raman-
527 microspectroscopy to elucidate their activity and metabolic traits *in situ* [67–70].

528 Another important aspect of the AsCoM database is that it allows us to determine the ecosystem-
529 specific taxonomic resolution afforded by different 16S rRNA gene-based amplicons and the
530 expected primer-bias introduced by the primers used for amplification. This insight is important for

531 making sound conclusions, and it may also form the basis for future development of improved
532 amplicon strategies.

533 **Conclusion:**

534 In this study, we applied synthetic long-read full-length 16S rRNA sequencing and AutoTax to
535 construct AsCoM, a comprehensive ecosystem-specific reference database for Askov and Cologne
536 soil. AsCoM provided increased taxonomic resolution in microbiome analyses of soil-grown plants
537 compared to the traditional universal reference database. In addition, AsCoM represents a valuable
538 reference for evaluating the ecosystem-specific bias associated with 16S rRNA gene amplicon
539 primers and developing species-level resolved hybridisation probes for *in situ* studies.

540

541 **Declarations:**

542 **Ethics approval and consent to participate**

543 Not applicable

544

545 **Consent for publication**

546 Not applicable

547

548 **Availability of data and material**

549 All sequencing data have been submitted to the Sequence Read Archive under the project ID
550 PRJNA787301. Details about individual datasets can be found in **Data S1**. All scripts used are
551 available at Github <https://github.com/msdueholm/Publications/tree/master/Overgaard2022a>.

552

553 **Competing interests**

554 The authors declare that they have no competing interests

555

556 **Funding**

557 This study was funded by Independent Research Fund Denmark (Grant no. 9041-00236B).

558

559 **Authors' contributions**

560 KT, SZ, BTC, ZB provided samples. CKO prepared sequencing libraries for full-length 16S rRNA
561 sequencing. CKO and MKDD processed the full-length 16S rRNA sequences. KT, SZ, ZB and SK
562 performed V5-7 amplicon sequencing. CKO, MKDD, and SK performed the bioinformatic analyses.
563 CKO, MKDD, SR, and SK wrote the manuscript; CKO, MKDD, SR and SK designed the study.
564 MKDD, SR, and SK supervised the study. All authors read and approved the final manuscript.

565

566 **Acknowledgements**

567 We thank the group of P. Schulze Lefert, MPI Cologne Germany, for providing the Cologne soil.

568

569 **Authors' information (optional)**

570 Not applicable

571 **References:**

- 572 1. Tilman D, Balzer C, Hill J, Befort BL. Global food demand and the sustainable intensification of
573 agriculture. *Proc Natl Acad Sci U S A*. 2011;108:20260–4.
- 574 2. Toju H, Peay KG, Yamamichi M, Narisawa K, Hiruma K, Naito K, et al. Core microbiomes for
575 sustainable agroecosystems. *Nat Plants*. 2018;4:247–57.
- 576 3. Trivedi P, Leach JE, Tringe SG, Sa T, Singh BK. Plant–microbiome interactions: from
577 community assembly to plant health. *Nat Rev Microbiol*. 2020;18:607–21.
- 578 4. Mendes R, Garbeva P, Raaijmakers JM. The rhizosphere microbiome: significance of plant
579 beneficial, plant pathogenic, and human pathogenic microorganisms. *FEMS Microbiol Rev*.
580 2013;37:634–63.
- 581 5. Chelius MK, Triplett EW. The Diversity of Archaea and Bacteria in Association with the Roots
582 of *Zea mays* L. *Microb Ecol*. 2001;41:252–63.
- 583 6. Fierer N, Ladau J, Clemente JC, Leff JW, Owens SM, Pollard KS, et al. Reconstructing the
584 Microbial Diversity and Function of Pre-Agricultural Tallgrass Prairie Soils in the United States.
585 *Science*. 2013;342:621.
- 586 7. Gittel A, Bárta J, Kohoutová I, Mikutta R, Owens S, Gilbert J, et al. Distinct microbial
587 communities associated with buried soils in the Siberian tundra. *ISME J*. 2014;8:841–53.
- 588 8. Tripathi BM, Kim M, Singh D, Lee-Cruz L, Lai-Hoe A, Ainuddin AN, et al. Tropical Soil
589 Bacterial Communities in Malaysia: pH Dominates in the Equatorial Tropics Too. *Microb Ecol*.
590 2012;64:474–84.
- 591 9. Armalytė J, Skerniškytė J, Bakienė E, Krasauskas R, Šiugždinienė R, Kareivienė V, et al.
592 Microbial Diversity and Antimicrobial Resistance Profile in Microbiota From Soils of Conventional
593 and Organic Farming Systems. *Front Microbiol*. 2019;10:892–892.
- 594 10. Huang AC, Jiang T, Liu Y-X, Bai Y-C, Reed J, Qu B, et al. A specialised metabolic network
595 selectively modulates *Arabidopsis* root microbiota. *Science*. 2019;364:eaau6389.
- 596 11. Voges MJEEE, Bai Y, Schulze-Lefert P, Sattely ES. Plant-derived coumarins shape the
597 composition of an *Arabidopsis* synthetic root microbiome. *Proc Natl Acad Sci*. 2019;116:12558.
- 598 12. Stringlis IA, Proietti S, Hickman R, Van Verk MC, Zamioudis C, Pieterse CMJ. Root
599 transcriptional dynamics induced by beneficial rhizobacteria and microbial immune elicitors reveal
600 signatures of adaptation to mutualists. *Plant J Cell Mol Biol*. 2018;93:166–80.

13. Jacobs S, Zechmann B, Molitor A, Trujillo M, Petutschnig E, Lipka V, et al. Broad-Spectrum Suppression of Innate Immunity Is Required for Colonisation of Arabidopsis Roots by the Fungus *Piriformospora indica*. *Plant Physiol.* 2011;156:726.

14. Finkel OM, Salas-González I, Castrillo G, Spaepen S, Law TF, Teixeira PJPL, et al. The effects of soil phosphorus content on plant microbiota are driven by the plant phosphate starvation response. *PLOS Biol.* 2019;17:e3000534.

15. Thiergart T, Zgadzaj R, Bozsóki Z, Garrido-Oter R, Radutoiu S, Schulze-Lefert P. *Lotus japonicus* Symbiosis Genes Impact Microbial Interactions between Symbionts and Multikingdom Commensal Communities. *mBio.* 2019;10:e01833-19.

16. Zgadzaj R, Garrido-Oter R, Jensen DB, Koprivova A, Schulze-Lefert P, Radutoiu S. Root nodule symbiosis in *Lotus japonicus* drives the establishment of distinctive rhizosphere, root, and nodule bacterial communities. *Proc Natl Acad Sci.* 2016;113:E7996.

17. Buchholz F, Antonielli L, Kostić T, Sessitsch A, Mitter B. The bacterial community in potato is recruited from soil and partly inherited across generations. *PloS One.* 2019;14:e0223691–e0223691.

18. Bonder MJ, Abeln S, Zaura E, Brandt BW. Comparing clustering and pre-processing in taxonomy analysis. *Bioinformatics.* 2012;28:2891–7.

19. Montijn R, Nocker A, Schuren F, Kort R, Richter-Heitmann T. Discrimination between live and dead cells in bacterial communities from environmental water samples analysed by 454 pyrosequencing. *Int Microbiol.* 2010;13:59–65.

20. Beckers B, Op De Beeck M, Thijs S, Truyens S, Weyens N, Boerjan W, et al. Performance of 16s rDNA Primer Pairs in the Study of Rhizosphere and Endosphere Bacterial Microbiomes in Metabarcoding Studies. *Front Microbiol.* 2016;7:650–650.

21. Dueholm MS, Andersen KS, McIlroy SJ, Kristensen JM, Yashiro E, Karst SM, et al. Generation of Comprehensive Ecosystem-Specific Reference Databases with Species-Level Resolution by High-Throughput Full-Length 16S rRNA Gene Sequencing and Automated Taxonomy Assignment (AutoTax). *mBio.* 2020;11:e01557-20.

22. Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A phylogenetic perspective. *Science.* 2015;350:aac9323.

23. Christensen BT, Thomsen IK, Eriksen J. The Askov long-term experiments: 1894-2019: a unique research platform turns 125 years. DCA report No 151. Aarhus University, Danish Centre for Food and Agriculture, Tjele: Denmark; 2019.

632 24. Bulgarelli D, Rott M, Schlaeppi K, Ver Loren van Themaat E, Ahmadinejad N, Assenza F, et al.
633 Revealing structure and assembly cues for *Arabidopsis* root-inhabiting bacterial microbiota. *Nature*.
634 2012;488:91–5.

635 25. Thiergart T, Durán P, Ellis T, Vannier N, Garrido-Oter R, Kemen E, et al. Root microbiota
636 assembly and adaptive differentiation among European *Arabidopsis* populations. *Nat Ecol Evol*.
637 2020;4:122–31.

638 26. Wippel K, Tao K, Niu Y, Zgadzaj R, Kiel N, Guan R, et al. Host preference and invasiveness of
639 commensal bacteria in the *Lotus* and *Arabidopsis* root microbiota. *Nat Microbiol*. 2021;6:1150–62.

640 27. Harbort CJ, Hashimoto M, Inoue H, Niu Y, Guan R, Rombolà AD, et al. Root-Secreted
641 Coumarins and the Microbiota Interact to Improve Iron Nutrition in *Arabidopsis*. *Cell Host*
642 *Microbe*. 2020;28:825-837.e6.

643 28. Bai Y, Müller DB, Srinivas G, Garrido-Oter R, Potthoff E, Rott M, et al. Functional overlap of
644 the *Arabidopsis* leaf and root microbiota. *Nature*. 2015;528:364–9.

645 29. Handberg K, Stougaard J. *Lotus japonicus*, an autogamous, diploid legume species for classical
646 and molecular genetics. *Plant J*. 1992;2:487–96.

647 30. Young ND, Debellé F, Oldroyd GED, Geurts R, Cannon SB, Udvardi MK, et al. The *Medicago*
648 genome provides insight into the evolution of rhizobial symbioses. *Nature*. 2011;480:520–4.

649 31. Dueholm MS, Nierychlo M, Andersen KS, Rudkjøbing V, Knutsson S, Consortium the MG, et
650 al. MiDAS 4: A global catalogue of full-length 16S rRNA gene sequences and taxonomy for studies
651 of bacterial communities in wastewater treatment plants. *bioRxiv*. 2021;2021.07.06.451231.

652 32. Karst SM, Dueholm MS, McIlroy SJ, Kirkegaard RH, Nielsen PH, Albertsen M. Retrieval of a
653 million high-quality, full-length microbial 16S and 18S rRNA gene sequences without primer bias.
654 *Nat Biotechnol*. 2018;36:190–5.

655 33. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for
656 Large Alignments. *PLOS ONE*. 2010;5:e9490.

657 34. Edgar RC. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*.
658 2010;26:2460–1.

659 35. Walters WA, Caporaso JG, Lauber CL, Berg-Lyons D, Fierer N, Knight R. PrimerProspector:
660 de novo design and taxonomic analysis of barcoded polymerase chain reaction primers.
661 *Bioinformatics*. 2011;27:1159–61.

36. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria: R Foundation for Statistical Computing, Vienna; 2008. URL <https://www.R-project.org/>

37. RStudio Team. RStudio: Integrated Development for R. Boston, MA: RStudio, PBC; 2020. URL <http://www.rstudio.com/>

38. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the tidyverse. *J Open Source Softw.* 2019;4:1686.

39. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. *vegan*: Community Ecology Package. 2019.

40. Andersen KS, Kirkegaard RH, Karst SM, Albertsen M. *ampvis2*: an R package to analyse and visualise 16S rRNA amplicon data. *bioRxiv.* 2018;:299537.

41. Pedersen TL. *patchwork*: The Composer of Plots. 2020.

42. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. *ggtree*: an r package for visualisation and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol.* 2017;8:28–36.

43. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* 2013;41:D590–6.

44. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol.* 2014;12:635–45.

45. Garrido-Sanz D, Redondo-Nieto M, Martin M, Rivilla R. Comparative genomics of the *Pseudomonas corrugata* subgroup reveals high species diversity and allows the description of *Pseudomonas ogarae* sp. nov. *Microb Genomics.* 2021;7.

46. Berendsen RL, van Verk MC, Stringlis IA, Zamioudis C, Tommassen J, Pieterse CMJ, et al. Unearthing the genomes of plant-beneficial *Pseudomonas* model strains WCS358, WCS374 and WCS417. *BMC Genomics.* 2015;16:539.

47. Elshafie HS, Camele I. An Overview of Metabolic Activity, Beneficial and Pathogenic Aspects of *Burkholderia* Spp. *Metabolites.* 2021;11:321.

48. DeSantis TZ, Hugenholtz P, Larsen N, Rojas M, Brodie EL, Keller K, et al. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol.* 2006;72:5069–72.

694 49. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete
695 domain-to-species taxonomy for Bacteria and Archaea. *Nat Biotechnol.* 2020;38:1079–86.

696 50. Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal Database Project:
697 data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* 2014;42:D633–42.

698 51. Jeong J, Yun K, Mun S, Chung W-H, Choi S-Y, Nam Y, et al. The effect of taxonomic
699 classification by full-length 16S rRNA sequencing with a synthetic long-read technology. *Sci Rep.*
700 2021;11:1727.

701 52. Matsuo Y, Komiya S, Yasumizu Y, Yasuoka Y, Mizushima K, Takagi T, et al. Full-length 16S
702 rRNA gene amplicon analysis of human gut microbiota using MinION™ nanopore sequencing
703 confers species-level resolution. *BMC Microbiol.* 2021;21:35.

704 53. Callahan BJ, Wong J, Heiner C, Oh S, Theriot CM, Gulati AS, et al. High-throughput amplicon
705 sequencing of the full-length 16S rRNA gene with single-nucleotide resolution. *Nucleic Acids Res.*
706 2019;47:e103.

707 54. Milanese A, Mende DR, Paoli L, Salazar G, Ruscheweyh H-J, Cuenca M, et al. Microbial
708 abundance, activity and population genomic profiling with mOTUs2. *Nat Commun.* 2019;10:1014.

709 55. Lu J, Salzberg SL. Ultrafast and accurate 16S rRNA microbial community analysis using
710 Kraken 2. *Microbiome.* 2020;8:124.

711 56. Ye SH, Siddle KJ, Park DJ, Sabeti PC. Benchmarking Metagenomics Tools for Taxonomic
712 Classification. *Cell.* 2019;178:779–94.

713 57. Albertsen M, Karst SM, Ziegler AS, Kirkegaard RH, Nielsen PH. Back to basics - the influence
714 of DNA extraction and primer choice on phylogenetic analysis of activated sludge communities.
715 *PloS One.* 2015;10:e0132783.

716 58. Brown SP, Grillo MA, Podowski JC, Heath KD. Soil origin and plant genotype structure
717 distinct microbiome compartments in the model legume *Medicago truncatula*. *Microbiome.*
718 2020;8:139.

719 59. Ofek M, Voronov-Goldman M, Hadar Y, Minz D. Host signature effect on plant root-associated
720 microbiomes revealed through analyses of resident vs. active communities. *Environ Microbiol.*
721 2014;16:2157–67.

722 60. Tovi N, Frenk S, Hadar Y, Minz D. Host Specificity and Spatial Distribution Preference of
723 Three *Pseudomonas* Isolates. *Front Microbiol.* 2019;9.

724 61. Zhu Q, Mai U, Pfeiffer W, Janssen S, Asnicar F, Sanders JG, et al. Phylogenomics of 10,575
725 genomes reveals evolutionary proximity between domains Bacteria and Archaea. *Nat Commun.*
726 2019;10:5477.

727 62. Parks DH, Chuvochina M, Waite DW, Rinke C, Skarshewski A, Chaumeil P-A, et al. A
728 standardised bacterial taxonomy based on genome phylogeny substantially revises the tree of life.
729 *Nat Biotechnol.* 2018;36:996.

730 63. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of
731 the tree of life. *Nat Microbiol.* 2016;1:1–6.

732 64. Nierychlo M, Singleton CM, Petriglieri F, Thomsen L, Petersen JF, Peces M, et al. Low Global
733 Diversity of *Candidatus* Microthrix, a Troublesome Filamentous Organism in Full-Scale WWTPs.
734 *Front Microbiol.* 2021;12:690251.

735 65. Petriglieri F, Singleton C, Kondrotaitė Z, Dueholm MKD, McDaniel EA, McMahon KD, et al.
736 Re-evaluation of the phylogenetic diversity and global distribution of the genus *Candidatus*
737 *Accumulibacter*. *bioRxiv.* 20121;:473458.

738 66. Edgar RC. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon
739 sequencing. *bioRxiv.* 2016;:81257.

740 67. Wagner M, Nielsen PH, Loy A, Nielsen JL, Daims H. Linking microbial community structure
741 with function: fluorescence in situ hybridization-microautoradiography and isotope arrays. *Curr*
742 *Opin Biotechnol.* 2006;17:83–91.

743 68. Huang WE, Stoecker K, Griffiths R, Newbold L, Daims H, Whiteley AS, et al. Raman-FISH:
744 combining stable-isotope Raman spectroscopy and fluorescence in situ hybridisation for the single
745 cell analysis of identity and function. *Environ Microbiol.* 2007;9:1878–89.

746 69. Singer E, Wagner M, Woyke T. Capturing the genetic makeup of the active microbiome in situ.
747 *ISME J.* 2017;11:1949–63.

748 70. Fernando EY, McIlroy SJ, Nierychlo M, Herbst F-A, Petriglieri F, Schmid MC, et al. Resolving
749 the individual contribution of key microbial populations to enhanced biological phosphorus removal
750 with Raman–FISH. *ISME J.* 2019;13:1933–46.

751 71. Bodenhausen N, Horton MW, Bergelson J. Bacterial Communities Associated with the Leaves
752 and the Roots of *Arabidopsis thaliana*. *PLOS ONE.* 2013;8:e56329.

754

755 **Figure 1: Novel sequences and *de novo* taxa observed in the AsCoM reference database.** Phylogenetic trees are based
756 on the FL-ASVs in the AsCoM database. The eight phyla with most FL-ASVs are highlighted. Sequence novelty was
757 determined by the percent identity between each FL-ASV and their closest relative in the SILVA_138.1_SSURef_Nr99
758 database and taxonomic thresholds proposed by [44] shown in Table 1. Taxonomy novelty was defined based on the
759 assignment of *de novo* taxa by AutoTax [21].

Figure 2: Database evaluation based on mapping of short-read amplicon data. V5-V7 ASVs obtained from Askov bulk soil as well as rhizosphere and endosphere samples from *Lotus japonicus*, *Medicago truncatula*, *Hordeum vulgare*, and *Zea mays* were mapped to the following reference databases and the proportion of high-identity hits ($\geq 99\%$ id) calculated: AsCoM, GreenGenes 16S 13.5 (GG) [48], GTDB release 89 [49], RDP 16S v16 [50], and SILVA SSU 138.1 Ref NR99 [43]. The ASVs were filtered based on their relative abundance (only ASVs with $\geq 0.01\%$ relative abundance were kept) before the analyses. Coverage shows how much of the accumulated read abundance these abundant ASVs accounted for (mean $\pm$ standard deviation, number of samples).

Figure 3: Database evaluation based on classification of short-read amplicon data. The V5-V7 amplicon data used in figure 2 was classified using the SINTAX-classifier with AsCoM, GreenGenes 16S 13.5 (GG) [48], GTDB release 89 [49], RDP 16S v16 [50], and SILVA SSURef 138.1 NR99 [43]. The ASVs were filtered based on their relative abundance (only ASVs with $\geq 0.01\%$ relative abundance were kept) before the analyses. Coverage shows how much of the accumulated read abundance these abundant ASVs accounted for (mean $\pm$ standard deviation, number of samples).

775

776 **Figure 4: Classification of the 50 most abundant ASVs in the Askov rhizosphere with four plant types.** ASVs were
777 classified with the SINTAX-classifier using AsCoM and the SILVA SSURef NR 138.1 database and best classification
778 (lowest rank) is presented. Three biological replicates are shown for each plant type. *De novo* placeholder taxa provided
779 by the AsCoM database is highlighted in red. ANPR: *Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium*; BCP:
780 *Burkholderia-Caballeronia-Paraburkholderia*.

781

782

783

784

785

786

787

788

789

790

Figure 5: *In silico* evaluation of taxonomic resolution and primer mismatch of the V5-V7 primer. a and b) *In silico* ASVs corresponding to the region amplified with the V5-V7 primer pair [5, 71] were extracted from the all bacterial FL-ASVs (a) or selected families (b) in AsCoM and classified using the AsCoM database. Classifications were evaluated as correctly classified, wrongly classified, or not classified using the taxonomy of FL-ASV in the AsCoM database as the ground truth. c) *In silico*, ecosystem-specific evaluation of primer specificity of the V5-V7 primer pair using Primer Prospector and bacterial FL-ASVs from the AsCoM database. The gray branches show perfect hits, the yellow branches show partial hits, and the red branches show poor hits (see materials and methods for definitions). The phyla most influenced from primer bias as well as *Burkholderiales* and *Rhizobiales* are highlighted in different colors in the taxonomy tree.

Figure 6. Effect of plant species on the rhizosphere and endosphere microbiota. a) Beta-diversity analysis using Bray-Curtis dissimilarities separates samples on the basis of host species for both the rhizosphere and endosphere compartments. b) Percent relative abundance of the 20 most abundant genera within the rhizosphere and endosphere compartments. Three biological replicates are shown for each host species. The relative abundance for bulk soil is included for reference.

797

798

799

800

Figure S1: Taxonomic novelty within different phyla in the AsCoM database. The number of taxa assigned *de novo* name by AutoTax within the taxonomic rank of class, order, family, and genus for each phylum. *De novo* phyla are highlighted in red.

801

802

803

804

805

806

Figure S2: Classification of the 50 most abundant ASVs in the Askov endosphere with four plant types. ASVs were classified with the SINTAX-classifier using AsCoM and the SILVA SSURef NR 138.1 database and best classification (lowest rank) is presented. Three biological replicates are shown for each plant type. *De novo* placeholder taxa provided by the AsCoM database is highlighted in red. ANPR: *Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium*; BCP: *Burkholderia-Caballeronia-Paraburkholderia*.

807

808 **Figure S3. Diversity metrics.** a) Alpha-diversity analysis using the ASV richness and the inverse Simpson's index of
809 bulk soil, rhizosphere (RZ), endosphere (ES) and nodule (Nod) compartments harvested from *Lotus japonicus* (*Lj*),
810 *Medicago truncatula* (*Mt*), *Hordeum vulgare* (*Hv*) and *Zea mays* (*Zm*). Each dot represents the average of three technical
811 replicates. b) Beta-diversity analysis using Bray-Curtis dissimilarities of the same microbiome samples.

812

813 **Figure S4. Host preference for rhizobia.** Relative abundance of all genera in the Order Rhizobiales across the
814 rhizosphere (RZ), endosphere (ES) and nodule (Nod) compartments harvested from *Lotus japonicus* (Lj), *Medicago*
815 *truncatula* (Mt), *Hordeum vulgare* (Hv) and *Zea mays* (Zm). Bulk soil is included for reference. The means of three
816 biological replicates are shown. *De novo* placeholder taxa provided by the AsCoM database are highlighted in red. ANPR:
817 *Allorhizobium-Neorhizobium-Pararhizobium-Rhizobium*.

Figure S5. MA-plot of ASVs enriched in the rhizosphere of four plant species compared to bulk soil. DEseq2 analysis was used to determine which ASVs were enriched in the rhizosphere. ASVs having a significant ($p_{adj} < 0.01$) positive log2 fold change are enriched in the rhizosphere compared to bulk soil (red) whereas ASVs having a significant ($p_{adj} < 0.01$) negative log2 fold change are reduced (blue). The adjusted p-values, log2 fold change, and taxonomy for individual ASVs can be found in Data S2.

825 **Tables:**

826 **Table 1: Novel sequences and *de novo* taxa observed in the AsCoM reference database.** Sequence novelty was determined based
827 on the percent identity between each FL-ASV and their closest relative in the SILVA_138.1_SSURef_Nr99 database and taxonomic
828 thresholds proposed by [44]. Taxonomic novelty was defined based on the number of *de novo* taxa assigned by AutoTax at each
829 taxonomic rank.

	Sequence novelty (bacteria/archaea)		Taxonomy novelty* (bacteria/archaea)	
	Sequences	Percentage (%)	<i>De novo</i> Taxa	Percentage (%)
New phylum (<75.0%)	10 / 0	0.06 / 0	3 / 0	8.33 / 0
New class (<78.5%)	16 / 0	0.09 / 0	15 / 0	15.15 / 0
New order (<82.0%)	26 / 0	0.15 / 0	91 / 1	30.33 / 12.50
New family (<86.5%)	116 / 0	0.65 / 0	420 / 8	61.49 / 80.00
New genus (<94.5%)	1138 / 4	6.36 / 3.03	1997 / 16	80.95 / 84.21
New species (<98.7%)	6719 / 18	37.55 / 13.64	7604 / 41	96.64 / 100

830 **De novo* species also include known species that cannot be resolved based on full-length 16S rRNA genes.

831

832 **Table S1: Oligonucleotides used for full-length 16S rRNA gene library preparation.** Unique molecular tags, and
 833 sample barcodes are marked with blue and red, respectively. Primers specific for bacteria or archaea are marked with
 834 “_b_” or “_a_” in their names, respectively.

Name	Sequence	Protocol step
f16S_b_pcr1_fw1	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN TGCCTCT TAGAGTTTGATCMTGGCTCAG	Barcoding and unique molecular identifier (UMI) tagging
f16S_b_pcr1_fw2	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN TCCTCTA CAGAGTTTGATCMTGGCTCAG	
f16S_b_pcr1_fw3	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN TCATGAGC AGAGTTTGATCMTGGCTCAG	
f16S_b_pcr1_fw4	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN CCTGAGAT AGAGTTTGATCMTGGCTCAG	
f16S_b_pcr1_rv1	AGCGCGGCAAAGATGAAGATNNNNNNNNNNNNNN TGAACCTT GACGGGCGGTGWGTRCA	
f16S_a_pcr1_fw1	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN TGCCTCTT TCCGGTTGATCCYGCBRG	
f16S_a_pcr1_fw2	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN TCCTCTA CTCCGGTTGATCCYGCBRG	
f16S_a_pcr1_fw3	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN TCATGAGC TCCGGTTGATCCYGCBRG	
f16S_a_pcr1_fw4	CTCCACCCAGACTCATCCATNNNNNNNNNNNNNN CCTGAGAT TCCGGTTGATCCYGCBRG	
f16S_a_pcr1_rv1	AGCGCGGCAAAGATGAAGATNNNNNNNNNNNNNN TGAACCTT GGCCATGCAMYWCCTCTC	
f16S_pcr2_fw	CTCCACCCAGACTCATCCAT	Library and clonal amplification
f16S_pcr2_rv	AGCGCGGCAAAGATGAAGAT	
f16S_readtag_fw	CAAGCAGAAGACGGCATAACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTCTCCACCCAGACTCATCCAT	Read-tag library preparation
f16S_readtag_rv	CAAGCAGAAGACGGCATAACGAGATGTGACTGGAGTTCAGACGTGTGCTCTTCCGATCTAGCGCGGCAAAGATGAAGAT	
f16S_b_linktag_fw	CAAGCAGAAGACGGCATAACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCTCTGAGCCAKGATCAAACCTCT	Linked-tag library preparation
f16S_b_linktag_rv	AATGATACGCGACCAACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCTTGACWACCGCCCGTC	
f16S_a_linktag_fw	CAAGCAGAAGACGGCATAACGAGATCGGTCTCGGCATTCTGCTGAACCGCTCTTCCGATCTCYVGCRRGGATCAACCGGA	
f16S_a_linktag_rv	AATGATACGCGACCAACCGAGATCTACACTCTTCCCTACACGACGCTCTTCCGATCTGAGAGGWRKTGCATGGCC	
f16S_read2_fw	GCTCTTCCGATCTCTCCACCCAGACTCATCCAT	Illumina MiSeq/HiSeq sequencing
f16S_read2_rv	GCTCTTCCGATCTAGCGCGCAAAGATGAAGAT	

836 **Table S2: Classification of bacterial isolates from the rhizosphere of Lotus japonicus grown in Cologne soil.** The
837 number and percentage of isolates classified at different taxonomic ranks is shown for AsCoM and
838 SILVA_138.1_SSURef_NR99.

839

Taxonomic rank	AsCoM	SILVA 138.1
Kingdom	277 (100%)	277 (100%)
Phylum	277 (100%)	277 (100%)
Class	277 (100%)	277 (100%)
Order	277 (100%)	277 (100%)
Family	277 (100%)	277 (100%)
Genus	230 (83.0%)	231 (83.4%)
Species	15 (5.4%)	0

840