

**Root-zone associated core microbiome**

1 **COREMIC: a web-tool to search for a root-zone associated CORE MICrobiome**

2 Richard R. Rodrigues<sup>a,b,\*</sup>, Nyle C. Rodgers<sup>c</sup>, Xiaowei Wu<sup>d</sup>, and Mark A. Williams<sup>a,e</sup>

3 <sup>a</sup>Interdisciplinary Ph.D. Program in Genetics, Bioinformatics, and Computational Biology, Virginia Tech, Blacksburg 24061, Virginia,  
4 United States of America.

5 <sup>b</sup>Department of Pharmaceutical Sciences, Oregon State University, Corvallis 97331, Oregon, United States of America.

6 <sup>c</sup>Department of Electrical and Computer Engineering, Virginia Tech, Blacksburg 24061, Virginia, United States of America.

7 <sup>d</sup>Department of Statistics, Virginia Tech, Blacksburg 24061, Virginia, United States of America.

8 <sup>e</sup>Department of Horticulture, Virginia Tech, Blacksburg 24061, Virginia, United States of America.

9

10 Richard Rodrigues ([richrr@vt.edu](mailto:richrr@vt.edu))

11 Nyle Rodgers ([nyle@vt.edu](mailto:nyle@vt.edu))

12 Xiaowei Wu ([xwwu@vt.edu](mailto:xwwu@vt.edu))

13 Mark Williams ([markwill@vt.edu](mailto:markwill@vt.edu))

14 **Contact:** Richard Rodrigues ([richrr@vt.edu](mailto:richrr@vt.edu)); 409 Pharmacy Bldg., Oregon State University, Corvallis OR 97331

15 \*To whom correspondence should be addressed.

16

17 **Abstract**

18 Microbial diversity on earth is extraordinary, and soils alone harbor thousands of species per gram of soil. Understanding  
19 how this diversity is sorted and selected into habitat niches is a major focus of ecology and biotechnology, but remains  
20 only vaguely understood. A systems-biology approach was used to mine information from databases to show how it can  
21 be used to answer questions related to the core microbiome of habitat-microbe relationships. By making use of the bur-  
22 geoning growth of information from databases, our tool “COREMIC” meets a great need in the search for understanding  
23 niche partitioning and habitat-function relationships. The work is unique, furthermore, because it provides a user-friendly  
24 statistically robust web-tool (<http://coremic2.appspot.com>), developed using Google App Engine, to help in the process  
25 of database mining to identify the “core microbiome” associated with a given habitat. A case study is presented using  
26 data from 31 switchgrass rhizosphere community habitats across a diverse set of soil and sampling environments. The  
27 methodology utilizes an outgroup of 28 non-switchgrass (other grasses and forbs) to identify a core switchgrass  
28 microbiome. Even across a diverse set of soils (5 environments), and conservative statistical criteria (presence in more  
29 than 90% samples and FDR  $q$ -val < 0.05% for Fisher’s exact test) a core set of bacteria associated with switchgrass was

**R. Rodrigues et al.**

---

30 observed. These included, among others, closely related taxa from *Lysobacter spp.*, *Mesorhizobium spp.*, and  
31 *Chitinophagaceae*. These bacteria have been shown to have functions related to the production of bacterial and fungal  
32 antibiotics and plant growth promotion. COREMIC can be used as a hypothesis generating or confirmatory tool that  
33 shows great potential for identifying taxa that may be important to the functioning of a habitat (e.g. host plant). The case  
34 study, in conclusion, shows that COREMIC can identify key habitat-specific microbes across diverse samples, using cur-  
35 rently available databases and a unique freely available software.

36

37 **Keywords:** microbiome; root-zone; rhizosphere; web-tool; software; app; meta-analysis; database; data mining

38

### 39 **1. Introduction**

40 Microbial diversity on earth is extraordinary, and soils alone harbor thousands of species per gram (Hughes et al., 2001).  
41 Understanding how this diversity is sorted and selected into habitat niches is a major focus of ecology and biotechnology,  
42 but remains only vaguely understood. The advent of next-generation sequencing technologies now allow for the potential  
43 to make great leaps in the study of microbe-habitat relationships of highly diverse microbial communities and environ-  
44 ments. The identity and functions of this overwhelming multitude of microbes are in the beginning stages of being de-  
45 scribed, and are already providing insights into microbial impacts on plant and animal health (Berg, 2009; Evans and  
46 Schwarz, 2011; Clemente et al., 2012). Making use of the overwhelming amount of information on microbial taxa and  
47 habitats has enormous potential for use to further understand microbial-habitat relationships. Thus, the advent of new  
48 methods and approaches to utilize this data and describe microbiomes will benefit microbial ecology and biotechnology.

49 Though variations exist, a core microbiome can be defined, conceptually, using Venn diagrams, where over-lapping  
50 circles and non-overlapping areas of circles represent shared and non-shared members of a habitat, respectively (Shade  
51 and Handelsman, 2012). Typically, microbiomes identified in this manner are not statistically evaluated, or by nature,  
52 seek to answer specific hypothesis that are specific to an experiment. For example, studies often identify microbes asso-  
53 ciated with different plant growth stages, species, cultivars, and locations but rarely, if at all, mine databases or perform  
54 meta-analysis to statistically identify microbiomes across studies and experimental conditions (Chaudhary et al., 2012;  
55 Liang et al., 2012; Mao et al., 2013; Mao et al., 2014; Hargreaves et al., 2015; Rodrigues et al., 2015; Jesus et al., 2016;  
56 Rodrigues et al., 2017). Describing differences due to treatment or habitat conditions are informative in their own right,  
57 however, extending this framework to include an easy to use, and statistically robust tool to help in the mining of data  
58 from underutilized and burgeoning databases (e.g. the National Center for Biotechnology Information (NCBI), Riboso-

### **Root-zone associated core microbiome**

59 mal Database Project) can help transform the ecological study of microbes in their natural environment. Using the vast  
60 and growing databases of organism and habitat metadata will allow for both the testing and development of hypotheses  
61 associated with habitat-microbe relationships that were not formerly possible.

62 To address the challenges described above, we developed COREMIC - a novel, easy to use, and freely available web  
63 tool to identify the “core microbiome”, of any well-defined habitat (e.g. plant root-zone) or niche (Shade and  
64 Handelsman, 2012). This straightforward approach is a novel and powerful way to complement existing analysis (e.g.  
65 indicator species analysis (ISA) (Dufrene and Legendre, 1997)) by allowing for the use of data that is now overflowing  
66 among freely available databases. It seeks to determine the core set of microbes (core microbiome) that are explicitly  
67 associated with a host system or habitat. The ability to identify core microbiomes at this scale has great potential to de-  
68 scribe host-microbe interactions and habitat preferences of microbes.

69 A meta-analysis based case study was performed, combining diverse sequencing datasets derived from NCBI, to test  
70 for the occurrence of a core microbiome in the rhizosphere (root-zone) of switchgrass. Switchgrass is a US-native, peren-  
71 nial grass studied by many researchers, and thus has a growing database to mine for genetic information. Its widespread  
72 study is likely a result of its bioenergy potential, and the capacity of the grass to grow on marginal lands not dedicated to  
73 crops. Studies have identified different bacteria found in the root-zones of switchgrass (Jesus et al., 2010; Mao et al.,  
74 2011; Chaudhary et al., 2012; Liang et al., 2012; Mao et al., 2013; Bahulikar et al., 2014; Mao et al., 2014; Werling et al.,  
75 2014; Hargreaves et al., 2015; Jesus et al., 2016; Rodrigues et al., 2017), however, there has been no integrative study of  
76 different datasets identifying the core microbiome in switchgrass rhizospheres. It is thus proposed to identify host-habitat  
77 relationships as a proof of concept for a core microbiome. In this paper we utilize a plant host to define a habitat, but the-  
78 oretically any habitat and associated organisms could make use of COREMIC and its approach to identify a core  
79 microbiome.

80

## **81 2. Material and methods**

### *82 2.1. Datasets used in the study*

83 A diverse set of data composed of 61 samples from two different published datasets and collected from multiple locations  
84 (Jesus et al., 2016; Rodrigues et al., 2017) were used for this study. Data were obtained from the NCBI and selected  
85 based on the availability of the raw (16S rRNA) sequence data of root-zone bacteria from switchgrass and that for an out-  
86 group of reference (native and/or other grasses) plants.

**R. Rodrigues et al.**

---

87 The dataset “Jesus 2016”(Jesus et al., 2016), PRJEB6704, compared the rhizosphere soil microbial communities asso-  
88 ciated with restored prairie with three grass crops, namely corn, switchgrass, and mixed prairie grasses. The grasses were  
89 grown in fields of Michigan and Wisconsin and were harvested after two and ten years. The V6-V8 region of the 16S  
90 rRNA gene was amplified and sequenced using the Roche 454 pyrosequencing. In our study, we used a total of 43 sam-  
91 ples (3 each from corn, switchgrass, mixed grasses (2 yrs. only), and restored prairie grasses grown in Wisconsin and  
92 Michigan, and sampled after 2 and 10 years. Switchgrass grown in Michigan, composed of 4 samples, were collected  
93 following 10 years of plant growth.

94 The dataset “Rodrigues 2017”(Rodrigues et al., 2017), PRJNA320123, compared the root-zone soil microbial commu-  
95 nities associated with switchgrass cultivars: “Alamo” and “Dacotah”. The switchgrass were grown in the greenhouse us-  
96 ing soil derived from plots growing Switchgrass (>7 years) near Blacksburg, VA. Switchgrass rhizosphere bacteria were  
97 sampled at three different growth stages. The V3-V4 region of the 16S rRNA gene was amplified and sequenced using  
98 Illumina MiSeq sequencing. In our study, we used a total of 18 switchgrass samples for Alamo (A) and Dacotah (D) from  
99 stages V2 and E3 (4 AV2, 4 DV2, 5 AE3, 5 DE3 = 18).

100 Overall, these datasets served as a diverse resource (relevant differences are summarized in Figure 1) to compare the  
101 root-zone bacteria and identify core-bacteria associated with switchgrass.

102

### 103 *2.2. Sequence data analysis and picking of Operational Taxonomic Units (OTU)*

104 For the Rodrigues 2017 dataset, the OTU table was obtained from previously performed analysis (Rodrigues et al., 2017).  
105 For the Jesus 2016 dataset, quality score (25) and read lengths (150) thresholds were enforced using cutadapt (1.8.1)  
106 (Martin, 2011) and an open reference OTU picking (enable\_rev\_strand\_match True) was performed in QIIME v1.8.0  
107 (Caporaso et al., 2010), as previously described (Rodrigues et al., 2015; Rodrigues et al., 2017), to allow comparison with  
108 the other dataset. Briefly, uclust (Edgar, 2010) was used to cluster reads into OTUs (97% sequence similarity) and assign  
109 taxonomy against the Greengenes reference database version 13.8 (DeSantis et al., 2006; McDonald et al., 2012). Two  
110 samples from the Jesus 2016 dataset were removed from downstream analysis due to very few sequences assigned to  
111 OTUs.

112

### 113 *2.3. Combining two datasets*

114 Within each OTU table, sequences assigned to identical OTUs in a sample were summed to retain unique taxa. The  
115 common (678) OTUs from the two datasets were selected, converted to biom format and used for further analyses (Figure

### **Root-zone associated core microbiome**

116 1). The data table was filtered and rarefied using a sequence threshold of 1150, and the beta diversity was calculated using  
117 Bray-Curtis (Beals, 1984) distance and visualized using Principal Coordinate Analysis (Gower, 2005). Multivariate  
118 data analysis methods of MRPP (Mielke, 1984), Permanova (Anderson, 2001) and ANOSIM (Clarke, 1993) were used to  
119 identify whether the plant type (switchgrass versus non-switchgrass) were associated with different bacterial communi-  
120 ties.

121

#### **2.4. Core microbiome analysis**

123 To find the set of core OTUs, the samples in the combined OTU table (original data) were first divided into the interest  
124 group samples (switchgrass) and out-group samples. The abundance values for each OTU in each sample are then con-  
125 verted to binary (present/absent) values based on whether they are zero or nonzero. For each OTU a one-tailed Fisher's  
126 Exact Test was used to calculate a  $p$ -value testing whether an OTU was present in a significantly higher portion in the  
127 interest in-group (Switchgrass) compared to the out-group samples (numerous other grass species).

128 These  $p$ -values were corrected for multiple-testing using Benjamini Hochberg. The OTUs with a  $q$ -value  $< 0.05$  were  
129 then selected to only the OTUs that are present in at least 90% of the interest group samples. Uninformative OTUs (e.g.,  
130 k\_Bacteria;p\_c\_o\_f\_g\_s\_) were filtered out and the remaining OTUs were candidates for the core microbiome.

131

#### **2.5. Implementation of COREMIC**

133 COREMIC and the datasets are available at <http://coremic2.appspot.com>. Its code is available on github  
134 (<https://github.com/richrr/coremicro>). The web-tool was developed in Python 2.7, and is hosted on Google App Engine.  
135 Other requirements include GoogleAppEnginePipeline 1.9.22.1, pyqi 0.3.1, requests 2.10.0, requests-toolbelt 0.6.2,  
136 mailjet-rest 1.2.2, biom-format 1.1.2, ete3 3.0.0 (for tree generation—see below for details), webapp2 2.5.2, numpy 1.6.1,  
137 matplotlib 1.2.0, jinja2 2.6, ssl 2.7. COREMIC is accessible via any internet connected browser and emails the results to  
138 the user. The processing times with the default settings after uploading the data are provided in Table S1.

139 A custom python script generates a phylogenetic tree using the taxonomic labels for each OTU displaying the relation-  
140 ship between the core OTUs obtained from the group of interest and the out-group. This tree is generated using the ete3  
141 3.0.0 library.

142

### **3. Results**

143

*R. Rodrigues et al.*

144 After quality filtering, a total of 319,821 reads were obtained from the Jesus 2016 dataset (mean 461.45 and std. dev.  
145 69.34). Two samples with very few (48 and 75) counts were removed; each of the remaining samples had more than 1150  
146 sequences assigned to OTUs. The number of OTUs in the Jesus 2016 and Rodrigues 2017 datasets was 771 and 1118,  
147 respectively. The combined dataset had 678 OTUs, 31 switchgrass and 28 non-switchgrass (other grasses) samples.

148 The bacterial communities in switchgrass and grasses from the combined dataset were significantly different  
149 (Permanova, MRPP, and ANOSIM  $p$ -values  $< 0.01$ ) and as can be observed using the PCoA plot using the Bray-Curtis  
150 dissimilarity metric (Figure 2). These differences were apparent despite significant difference across datasets  
151 (Permanova, MRPP, and ANOSIM  $p$ -values  $< 0.01$ ); which could be the result, for example, of the heterogeneity of the  
152 data set related to climate, soil type-condition, growth conditions, and plant age. In this regard, at the phylum level, Mann  
153 Whitney test identified Bacteroidetes and Verrucomicrobia had significantly greater ( $p$ -value  $< 0.05$ ) relative abundance  
154 in switchgrass, whereas, Gemmatimonadetes were more abundant in other grasses (Figure S1).

155 We used a very conservative criterion of  $>90\%$  threshold i.e., an OTU has to be present in at least 90% of switchgrass  
156 samples and observed five OTUs with FDR  $q$ -values  $< 0.05$  (Table 1). The relative abundance and a phylogenetic tree  
157 exhibiting their relationship with the core-OTUs from the non-switchgrass samples is shown in Figure S2 and Figure S3,  
158 respectively. Despite the enormous variability across the many different sampling locations, there is support for the oc-  
159 currence of a core microbiome in the root-zone of switchgrass.

160

161 **Table 1: Bacterial OTUs associated with switchgrass.**

OTU	present(%)
p_Proteobacteria;c_Gammaproteobacteria;o_Xanthomonadales;f_Xanthomonadaceae;g_Lysobacter;s_	100
p_Planctomycetes;c_Planctomycetia;o_B97;f_;g_;s_	96.8
p_Bacteroidetes;c_[Saprospirae];o_[Saprospirales];f_Chitinophagaceae	96.8
p_Proteobacteria;c_Alphaproteobacteria;o_Rhizobiales;f_Phyllobacteriaceae;g_Mesorhizobium;s_	90.3
p_Proteobacteria;c_Gammaproteobacteria;o_Legionellales;f_;g_;s_	90.3

162 The core bacterial OTUs those were significantly ( $q$ -value  $< 0.05$ ) associated with switchgrass, calculated using pres-  
163 ence/absence data and present in  $>90\%$  switchgrass samples.

## Root-zone associated core microbiome

164

### 165 **4. Discussion**

166 The case study showed how COREMIC can identify key habitat-specific microbes across diverse samples, using current-  
167 ly available databases and a unique freely available software. The core set of bacteria associated with switchgrass includ-  
168 ed, among others, closely related taxa from *Lysobacter spp.*, *Mesorhizobium spp.*, and *Chitinophagaceae*. The functional  
169 relevance of these bacteria related to switchgrass is unknown, but it is notable that these bacteria have been shown to  
170 produce bacterial and fungal antibiotics and promote the growth of plants (Kaneko et al., 2000; Kilic-Ekici and Yuen,  
171 2004; Weir et al., 2004; Islam et al., 2005; Jochum et al., 2006; Ji et al., 2008; Park et al., 2008; Nandasena et al., 2009;  
172 Yin, 2010; Bailey et al., 2013; Degefu et al., 2013; Guerrouj et al., 2013; Madhaiyan et al., 2015). The analyses from the  
173 highly diverse data sets thus provided information that helps to greatly narrow down possibilities and thus set the stage  
174 for testing, using controlled studies, how the core microbiota potentially support or antagonize the function of a native  
175 grass. This novel toolkit is simple to use and supports use by a broad range of biological scientists, and is particularly  
176 relevant to those with expertise in their field but with limited bioinformatics background. Overall, in a dataset derived  
177 from a complex and diverse set of habitats and ecosystems, this tool was shown to pinpoint microbiota of the microbiome  
178 that might have important functional implications within their habitat or host.

179

#### 180 *4.1. Methodological considerations in the use of COREMIC*

181 COREMIC performs a complementary analysis different from that of existing methods by using presence/absence data.  
182 For two groups (A and B) it checks whether (pre-determined percentage of) samples from group A have a non-zero value  
183 for the OTU. This allows scientists to operate without making assumptions about the PCR-based OTU relative abundanc-  
184 es. This is considered a potential advantage of the method because it is unknown whether relative abundance of sequence  
185 data is representative of true relative differences between communities. Further research, in this regard, will be aimed  
186 towards investigating other measures of OTU “presence”, namely the extent of exclusivity, consistency, or abundance of  
187 the group that is eventually determined to be a core microbiome.

188 Sampling plots used in this study were located across a range of diverse environments to help create a backdrop of het-  
189 erogeneity. While this diversity of habitat conditions ignores the potential for microbe-environment interactions that  
190 might be important for the plant-microbial relationship, it has the advantage of being a conservative approach with high  
191 veracity for defining a core microbiome regardless of habitat heterogeneity. The locations from which samples were  
192 grown (Michigan, Wisconsin, Virginia) were treated as independent to help isolate the overall habitat effect of

193 switchgrass (Werling et al., 2014; Jesus et al., 2016). When the effects of habitat are thought to be habitat specific, re-  
194 searchers can take this into account during the design and analysis using COREMIC.

195 It is notable that the representation of an outgroup (multiple non-switchgrass species) is an important criteria and  
196 choice made by researchers, and is an approach that has both advantages and caveats. By definition, a habitat is defined  
197 by its differences from that of other habitats, and therefore the use of the outgroup is an important choice. A counter-  
198 argument for the current dataset might argue for exclusion of breeding lines of a cultivated grass (maize) as being unre-  
199 presentative of the grass outgroup. In our case, it was thought, *a priori*, that a diverse set of grasses would provide the best  
200 comparison; and no compelling argument was found that supported the exclusion of maize from the analysis. An implicit  
201 assumption was also made that the taxonomy of plant species (root-zone habitats) play an important role in determining  
202 root-zone microbial communities, an approach supported by extensive findings that different grass species associate with  
203 different microbial communities (Kuske et al., 2002; Kennedy et al., 2004; Berendsen et al., 2012; Chaudhary et al.,  
204 2012; Turner et al., 2013). So although there is a need for careful consideration of the experimental questions of interest  
205 when using COREMIC, this is a common, if not ubiquitous foundation of all experimentation and hypothesis testing. The  
206 results provide a statistically valid approach using freely available software to describe and define a core microbiome of  
207 switchgrass.

208 The choice of the outgroup, furthermore, for determining a core microbiome is amenable to choice using deductive rea-  
209 soning but ultimately limited by available data. This issue almost certainly limits inclusion of many functionally im-  
210 portant rhizosphere microbes that could affect the growth of switchgrass. In this study, the proof of concept utilized a  
211 conservative approach to highlight the methodology across a diversity of geographies, soil types, and plant ages. The  
212 COREMIC tool as well as the multiple methods for defining a core microbiome (e.g., QIIME (Caporaso et al., 2010),  
213 ISA (Dufrene and Legendre, 1997)) will always be defined by the expertise, and the nature of the hypotheses defined and  
214 defended by individual researchers.

215

#### 216 *4.2. Core Microbes*

217 The individual datasets described in this study had previously focused on identifying abundant microbes and differences  
218 due to experimental conditions. The current meta-analysis goes a step further to find common microbiota that are associ-  
219 ated with switchgrass across the diverse experimental conditions. The members of the *Lysobacter* genus, an identified  
220 core microbe of switchgrass, are known to live in soil and have been shown to be ecologically important due to their abil-  
221 ity to produce exo-enzymes and antibiotics (Reichenbach, 2006). Their antimicrobial activities against bacteria, fungi,



### **Root-zone associated core microbiome**

222 unicellular algae, and nematodes have been described (Islam et al., 2005; Jochum et al., 2006; Park et al., 2008; Yin,  
223 2010). Strains of this genus, for example, have been used for control of diseases caused by bacteria in rice (Ji et al., 2008)  
224 and tall fescue (Kilic-Ekici and Yuen, 2004). Reports of their function thus support the idea that they may play an im-  
225 portant role in switchgrass growth and survival. The core microbiome results thus support further research into the role  
226 played by this bacterium in the switchgrass rhizosphere.

227 Similarly, members of the *Mesorhizobium* genus are well-known diazotrophs (Kaneko et al., 2000) and previously  
228 shown to be symbiotically associated with switchgrass (DeAngelis et al., 2010; Bahulikar et al., 2014) and legumes (Weir  
229 et al., 2004; Nandasena et al., 2009; Degefu et al., 2013; Guerrouj et al., 2013). Another identified core microbiome taxa,  
230 soil-dwelling members of the *Chitinophagaceae* family are known to have  $\beta$ -glucosidase (Bailey et al., 2013) and  
231 Aminocyclopropane-1-carboxylate (ACC) deaminase activities and ability to produce indole-3-acetic acid (IAA)  
232 (Madhaiyan et al., 2015). These molecules and enzymes are well known for their effects on plant growth (Zhao, 2010;  
233 Van de Poel and Van Der Straeten, 2014). The capacity to degrade cellulose might provide additional and readily availa-  
234 ble options to aid survival of these bacteria near switchgrass root zones during times of environmental stress. ACC  
235 deaminase and IAA production, in contrast, are potent plant growth modulators (Glick, 2014) that could play a role in  
236 plant productivity and survival, especially under conditions of plant physiological stress. Though these examples above  
237 would need further study, they provide consistent examples describing how a core microorganism could play a role in  
238 determining plant function and growth. The power of the approach stems from the ability to identify the core microbes  
239 associated with a plant (or other habitat), and that can, with veracity, narrow down potentially important core microbes  
240 from otherwise hyperdiverse samples.

241 From a technological standpoint, it is important to put the current approach into context with research before the  
242 metagenomics era. The search and identification of antagonistic plant growth promoting microbes has previously been  
243 tedious and labor intensive. Screenings of hundreds of microbes were used to cultivate and identify candidate microbes  
244 that might support (or deter) plant growth. In the case of beneficial microbes, even when identified under greenhouse  
245 conditions, the beneficial effects rarely translated into plant supportive growth under field growth conditions (Babalola,  
246 2010; Hayat et al., 2010). With the aid of hindsight and new knowledge suggesting the importance of the soil habitat and  
247 root-soil interactions in the development of growth promoting plant-microbial relationships, the approach used in this  
248 study reverses the focus (from top-down to bottom-up) to search for microbes that appear to already be naturally well-  
249 adapted to the root-soil habitats of interest (Trabelsi and Mhamdi, 2013; Souza et al., 2015). This process streamlines the  
250 search for suitable microbes from a daunting pool of thousands of bacterial taxa. Bacteria and fungi with well-known

*R. Rodrigues et al.*

---

251 partnerships with members of the core microbiome, it would be expected, to be more readily adaptable to their native  
252 environment. Indeed, the concept of adaptability to an environment has been shown to be true for many types of microbes  
253 across the environmental spectrum, and has given rise to the concept of the niche (Lennon et al., 2012). The COREMIC  
254 tool provides an alternative and logical approach to help mine available datasets, in the search for core microbiomes as-  
255 sociated with habitats that are ecologically and agriculturally important.

256

### 257 *4.3. Conclusions*

258 The COREMIC tool, by helping to mine multiple datasets fills a major gap in the search for the core microbiome associ-  
259 ated with a host or habitat. It allows for the development of a working hypothesis in the search for microbes well suited  
260 for a habitat or host-microbe interaction. It can also be used to confirm laboratory studies that have identified target mi-  
261 crobes that might be important symbionts or thought to be associated with a specific habitat. In the case of plants, but not  
262 limited to them, the COREMIC approach can identify microbial targets that might be useful for plant growth promotion.  
263 An example of this would be the identification of diazotrophic bacteria that aid the growth of bioenergy grasses and help  
264 to serve the development of sustainable agricultural systems. This combined with the ongoing efforts of plant breeding  
265 and genetic modification would help to catalyze microbe-driven crop yield improvement while practicing environmental  
266 stewardship through reduced fertilizer use. Here we show the applicability of COREMIC in rhizosphere-associated mi-  
267 crobes, but the overall concepts are translational across disciplines with interests in host-microbe and microbe-habitat  
268 relationships. The applicability of COREMIC for the identification of core genes and microbes has excellent potential to  
269 help understand the roles of microorganisms in complex and diverse microbial communities.

270

### 271 **Declarations**

### 272 **Ethics approval and consent to participate**

273 Not applicable.

274

### 275 **Consent for publication**

276 Not applicable.

277

### 278 **Availability of data and materials**

### **Root-zone associated core microbiome**

279 The datasets and results supporting the conclusions of this article are included within the article and supplementary files.  
280 COREMIC and the datasets are available at <http://coremic2.appspot.com>. An archived version of its code is available on  
281 github (<https://github.com/richrr/coremicro>). COREMIC and its code are freely available under the GPL license.

282

### **283 Competing interests**

284 The authors declare that they have no competing interests.

285

### **286 Authors' contributions**

287 Conceived and designed the experiments: RRR MAW. Implemented software tools: RRR NCR. Performed the experi-  
288 ments: RRR NCR. Analyzed the data: RRR NCR XW MAW. Wrote the paper: RRR NCR XW MAW. All authors read  
289 and approved the final manuscript.

290

### **291 Acknowledgements**

292 The authors thank Dr. Roderick Jensen and James Wrenn for their suggestions in improving the manuscript. We thank  
293 Dr. James McClure for his help in the web-tool development. We also acknowledge BIOM, Google App Engine and their  
294 developers. The authors acknowledge Virginia Polytechnic Institute and State University's Open Access Subvention  
295 Fund. We thank Virginia Tech's Genetics, Bioinformatics, and Computational Biology, Department of Horticulture for  
296 providing personnel funding. Research was also partially funded by grants from USDA-NIFA (2011-03815).

297

### **298 References**

299 Anderson, M.J., 2001. A new method for non-parametric multivariate analysis of variance. *Austral Ecology* 26, 32-46.  
300 Babalola, O.O., 2010. Beneficial bacteria of agricultural importance. *Biotechnology Letters* 32, 1559-1570.  
301 Bahulikar, R.A., Torres-Jerez, I., Worley, E., Craven, K., Udvardi, M.K., 2014. Diversity of nitrogen-fixing bacteria  
302 associated with switchgrass in the native tallgrass prairie of northern Oklahoma. *Applied and Environmental*  
303 *Microbiology* 80, 5636-5643.  
304 Bailey, V.L., Fansler, S.J., Stegen, J.C., McCue, L.A., 2013. Linking microbial community structure to beta-glucosidic  
305 function in soil aggregates. *The ISME journal* 7, 2044-2053.  
306 Beals, E.W., 1984. Bray-Curtis Ordination: An Effective Strategy for Analysis of Multivariate Ecological Data.  
307 *Advances in Ecological Research* 14, 1-55.

**R. Rodrigues et al.**

---

- 308 Berendsen, R.L., Pieterse, C.M., Bakker, P.A., 2012. The rhizosphere microbiome and plant health. Trends in Plant  
309 Science 17, 478-486.
- 310 Berg, G., 2009. Plant-microbe interactions promoting plant growth and health: perspectives for controlled use of  
311 microorganisms in agriculture. Applied Microbiology and Biotechnology 84, 11-18.
- 312 Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., Fierer, N., Pena, A.G.,  
313 Goodrich, J.K., Gordon, J.I., Huttley, G.A., Kelley, S.T., Knights, D., Koenig, J.E., Ley, R.E., Lozupone, C.A.,  
314 McDonald, D., Muegge, B.D., Pirrung, M., Reeder, J., Sevinsky, J.R., Turnbaugh, P.J., Walters, W.A., Widmann, J.,  
315 Yatsunenko, T., Zaneveld, J., Knight, R., 2010. QIIME allows analysis of high-throughput community sequencing data.  
316 Nature Methods 7, 335-336.
- 317 Chaudhary, D., Saxena, J., Lorenz, N., Dick, L., Dick, R., 2012. Microbial Profiles of Rhizosphere and Bulk Soil  
318 Microbial Communities of Biofuel Crops Switchgrass (*Panicum virgatum* L.) and *Jatropha* (*Jatropha curcas* L.). Applied  
319 and Environmental Soil Science 2012, 1-6.
- 320 Clarke, K.R., 1993. Non-parametric multivariate analyses of changes in community structure. Australian Journal of  
321 Ecology 18, 117-143.
- 322 Clemente, J.C., Ursell, L.K., Parfrey, L.W., Knight, R., 2012. The impact of the gut microbiota on human health: an  
323 integrative view. Cell 148, 1258-1270.
- 324 DeAngelis, K.M., Gladden, J.M., Allgaier, M., D'haeseleer, P., Fortney, J.L., Reddy, A., Hugenholtz, P., Singer, S.W.,  
325 Gheynst, J.S.V., Silver, W.L., Simmons, B.A., Hazen, T.C., 2010. Strategies for Enhancing the Effectiveness of  
326 Metagenomic-based Enzyme Discovery in Lignocellulolytic Microbial Communities. BioEnergy Research 3, 146-158.
- 327 Degefu, T., Wolde-Meskel, E., Liu, B., Cleenwerck, I., Willems, A., Frostegard, A., 2013. *Mesorhizobium shonense* sp.  
328 nov., *Mesorhizobium hawassense* sp. nov. and *Mesorhizobium abyssinicae* sp. nov., isolated from root nodules of  
329 different agroforestry legume trees. International Journal of Systematic and Evolutionary Microbiology 63, 1746-1753.
- 330 DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., Huber, T., Dalevi, D., Hu, P., Andersen,  
331 G.L., 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Applied and  
332 Environmental Microbiology 72, 5069-5072.
- 333 Dufrene, M., Legendre, P., 1997. Species Assemblages and Indicator Species: The Need for a Flexible Asymmetrical  
334 Approach. Ecological Monographs 67, 345-366.
- 335 Edgar, R.C., 2010. Search and clustering orders of magnitude faster than BLAST. Bioinformatics 26, 2460-2461.

**Root-zone associated core microbiome**

- 336 Evans, J.D., Schwarz, R.S., 2011. Bees brought to their knees: microbes affecting honey bee health. Trends in  
337 Microbiology 19, 614-620.
- 338 Glick, B.R., 2014. Bacteria with ACC deaminase can promote plant growth and help to feed the world. Microbiological  
339 Research 169, 30-39.
- 340 Gower, J.C., 2005. Principal Coordinates Analysis, Encyclopedia of Biostatistics, 2 ed. John Wiley and Sons, Ltd, The  
341 Open University, Milton Keynes, UK.
- 342 Guerrouj, K., Perez-Valera, E., Chahboune, R., Abdelmoumen, H., Bedmar, E.J., El Idrissi, M.M., 2013. Identification of  
343 the rhizobial symbiont of Astragalus glombiformis in Eastern Morocco as Mesorhizobium camelthorni. Antonie Van  
344 Leeuwenhoek 104, 187-198.
- 345 Hargreaves, S.K., Williams, R.J., Hofmockel, K.S., 2015. Environmental Filtering of Microbial Communities in  
346 Agricultural Soil Shifts with Crop Growth. PLoS One 10, e0134345.
- 347 Hayat, R., Ali, S., Amara, U., Khalid, R., Ahmed, I., 2010. Soil beneficial bacteria and their role in plant growth  
348 promotion: a review. Annals of Microbiology 60, 579-598.
- 349 Hughes, J.B., Hellmann, J.J., Ricketts, T.H., Bohannan, B.J.M., 2001. Counting the Uncountable: Statistical Approaches  
350 to Estimating Microbial Diversity. Applied and Environmental Microbiology 67, 4399-4406.
- 351 Islam, M.T., Hashidoko, Y., Deora, A., Ito, T., Tahara, S., 2005. Suppression of damping-off disease in host plants by the  
352 rhizoplane bacterium Lysobacter sp. strain SB-K88 is linked to plant colonization and antibiosis against soilborne  
353 Peronosporomycetes. Applied and Environmental Microbiology 71, 3786-3796.
- 354 Jesus, Susilawati, E., Smith, S., Wang, Q., Chai, B., Farris, R., Rodrigues, J., Thelen, K., Tiedje, J., 2010. Bacterial  
355 Communities in the Rhizosphere of Biofuel Crops Grown on Marginal Lands as Evaluated by 16S rRNA Gene  
356 Pyrosequences. BioEnergy Research 3, 20-27.
- 357 Jesus, E.d.C., Liang, C., Quensen, J.F., Susilawati, E., Jackson, R.D., Balser, T.C., Tiedje, J.M., 2016. Influence of corn,  
358 switchgrass, and prairie cropping systems on soil microbial communities in the upper Midwest of the United States.  
359 Global Change Biology Bioenergy 8, 481-494.
- 360 Ji, G.-H., Wei, L.-F., He, Y.-Q., Wu, Y.-P., Bai, X.-H., 2008. Biological control of rice bacterial blight by Lysobacter  
361 antibioticus strain 13-1. Biological Control 45, 288-296.
- 362 Jochum, C.C., Osborne, L.E., Yuen, G.Y., 2006. Fusarium head blight biological control with Lysobacter enzymogenes  
363 strain C3. Biological Control 39, 336-344.

**R. Rodrigues et al.**

---

- 364 Kaneko, T., Nakamura, Y., Sato, S., Asamizu, E., Kato, T., Sasamoto, S., Watanabe, A., Idesawa, K., Ishikawa, A.,  
365 Kawashima, K., Kimura, T., Kishida, Y., Kiyokawa, C., Kohara, M., Matsumoto, M., Matsuno, A., Mochizuki, Y.,  
366 Nakayama, S., Nakazaki, N., Shimpo, S., Sugimoto, M., Takeuchi, C., Yamada, M., Tabata, S., 2000. Complete genome  
367 structure of the nitrogen-fixing symbiotic bacterium *Mesorhizobium loti*. *DNA Research* 7, 331-338.
- 368 Kennedy, N., Brodie, E., Connolly, J., Clipson, N., 2004. Impact of lime, nitrogen and plant species on bacterial  
369 community structure in grassland microcosms. *Environmental Microbiology* 6, 1070-1080.
- 370 Kilic-Ekici, O., Yuen, G.Y., 2004. Comparison of strains of *Lysobacter enzymogenes* and PGPR for induction of  
371 resistance against *Bipolaris sorokiniana* in tall fescue. *Biological Control* 30, 446-455.
- 372 Kuske, C.R., Ticknor, L.O., Miller, M.E., Dunbar, J.M., Davis, J.A., Barns, S.M., Belnap, J., 2002. Comparison of soil  
373 bacterial communities in rhizospheres of three plant species and the interspaces in an arid grassland. *Applied and*  
374 *Environmental Microbiology* 68, 1854-1863.
- 375 Lennon, J.T., Aanderud, Z.T., Lehmkuhl, B.K., Schoolmaster, D.R., 2012. Mapping the niche space of soil  
376 microorganisms using taxonomy and traits. *Ecology* 93, 1867-1879.
- 377 Liang, C., Jesus, E., Duncan, D., Jackson, R., Tiedje, J., Balsler, T., 2012. Soil microbial communities under model  
378 biofuel cropping systems in southern Wisconsin, USA: Impact of crop species and soil properties. *Applied Soil Ecology*  
379 54, 24-31.
- 380 Madhaiyan, M., Poonguzhali, S., Senthilkumar, M., Pragatheswari, D., Lee, J.S., Lee, K.C., 2015. *Arachidicoccus*  
381 *rhizosphaerae* gen. nov., sp. nov., a plant-growth-promoting bacterium in the family Chitinophagaceae isolated from  
382 rhizosphere soil. *International Journal of Systematic and Evolutionary Microbiology* 65, 578-586.
- 383 Mao, Y., Li, X., Smyth, E., Yannarell, A., Mackie, R., 2014. Enrichment of specific bacterial and eukaryotic microbes in  
384 the rhizosphere of switchgrass (*Panicum virgatum* L.) through root exudates. *Environmental Microbiology Reports* 6, 13.
- 385 Mao, Y., Yannarell, A., Davis, S., Mackie, R., 2013. Impact of different bioenergy crops on N-cycling bacterial and  
386 archaeal communities in soil. *Environmental Microbiology* 15, 928-942.
- 387 Mao, Y., Yannarell, A., Mackie, R., 2011. Changes in N-Transforming Archaea and Bacteria in Soil during the  
388 Establishment of Bioenergy Crops. *PLoS One* 6, e24750.
- 389 Martin, M., 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. *EMBnet.journal* 17, 10.
- 390 McDonald, D., Price, M.N., Goodrich, J., Nawrocki, E.P., DeSantis, T.Z., Probst, A., Andersen, G.L., Knight, R.,  
391 Hugenholtz, P., 2012. An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of  
392 bacteria and archaea. *The ISME journal* 6, 610-618.

**Root-zone associated core microbiome**

- 393 Mielke, P.W., 1984. Meteorological applications of permutation techniques based on distance functions., In: Krishnaiah,  
394 P.R., Sen, P.K. (Eds.), Handbook of statistics: Nonparametric methods, Amsterdam: North-Holland, pp. 813-830.
- 395 Nandasena, K.G., O'Hara, G.W., Tiwari, R.P., Willems, A., Howieson, J.G., 2009. Mesorhizobium australicum sp. nov.  
396 and Mesorhizobium opportunistum sp. nov., isolated from Biserrula pelecinus L. in Australia. International Journal of  
397 Systematic and Evolutionary Microbiology 59, 2140-2147.
- 398 Park, J.H., Kim, R., Aslam, Z., Jeon, C.O., Chung, Y.R., 2008. Lysobacter capsici sp. nov., with antimicrobial activity,  
399 isolated from the rhizosphere of pepper, and emended description of the genus Lysobacter. International Journal of  
400 Systematic and Evolutionary Microbiology 58, 387-392.
- 401 Reichenbach, H., 2006. The Genus Lysobacter, The Prokaryotes. Springer New York, pp. 939-957.
- 402 Rodrigues, R.R., Moon, J., Zhao, B., Williams, M.A., 2017. Microbial communities and diazotrophic activity differ in the  
403 root-zone of Alamo and Dacotah switchgrass feedstocks. GCB Bioenergy 9, 1057-1070.
- 404 Rodrigues, R.R., Pineda, R.P., Barney, J.N., Nilsen, E.T., Barrett, J.E., Williams, M.A., 2015. Plant Invasions Associated  
405 with Change in Root-Zone Microbial Community Structure and Diversity. PLoS One 10, e0141424.
- 406 Shade, A., Handelsman, J., 2012. Beyond the Venn diagram: the hunt for a core microbiome. Environmental  
407 Microbiology 14, 4-12.
- 408 Souza, R., Ambrosini, A., Passaglia, L.M., 2015. Plant growth-promoting bacteria as inoculants in agricultural soils.  
409 Genetics and Molecular Biology 38, 401-419.
- 410 Trabelsi, D., Mhamdi, R., 2013. Microbial inoculants and their impact on soil microbial communities: a review. BioMed  
411 Research International 2013, 863240.
- 412 Turner, T., Ramakrishnan, K., Walshaw, J., Heavens, D., Alston, M., Swarbreck, D., Osbourn, A., Grant, A., Poole, P.,  
413 2013. Comparative metatranscriptomics reveals kingdom level changes in the rhizosphere microbiome of plants. The  
414 ISME journal 7, 2248-2258.
- 415 Van de Poel, B., Van Der Straeten, D., 2014. 1-aminocyclopropane-1-carboxylic acid (ACC) in plants: more than just the  
416 precursor of ethylene! Frontiers in Plant Science 5, 640.
- 417 Weir, B.S., Turner, S.J., Silvester, W.B., Park, D.C., Young, J.M., 2004. Unexpectedly diverse Mesorhizobium strains  
418 and Rhizobium leguminosarum nodulate native legume genera of New Zealand, while introduced legume weeds are  
419 nodulated by Bradyrhizobium species. Applied and Environmental Microbiology 70, 5980-5987.
- 420 Werling, B.P., Dickson, T.L., Isaacs, R., Gaines, H., Gratton, C., Gross, K.L., Liere, H., Malmstrom, C.M., Meehan,  
421 T.D., Ruan, L., Robertson, B.A., Robertson, G.P., Schmidt, T.M., Schrottenboer, A.C., Teal, T.K., Wilson, J.K., Landis,

*R. Rodrigues et al.*

---

422 D.A., 2014. Perennial grasslands enhance biodiversity and multiple ecosystem services in bioenergy landscapes.  
423 Proceedings of the National Academy of Sciences 111, 1652-1657.

424 Yin, H., 2010. Detection Methods for the Genus *Lysobacter* and the Species *Lysobacter enzymogenes*, Biological  
425 Sciences. University of Nebraska, Lincoln.

426 Zhao, Y., 2010. Auxin biosynthesis and its role in plant development. Annual Review of Plant Biology 61, 49-64.

427

428

429

430 **Figure 1: The COREMIC approach.** The workflow indicating the Jesus 2016 and Rodrigues 2017 datasets and differ-  
431 ences between them, and the methodology used to identify core microbiome. Switchgrass and other grasses are indicated  
432 by “Swg” and “Non-Swg,” respectively.

433

434 **Figure 2: Beta-diversity of the combined dataset.** PCoA plot showing Bray-Curtis dissimilarities for bacterial commu-  
435 nities at the OTU level in switchgrass (blue colored) and other grasses (red colored).

436

437 **Figure S1: Taxonomic summary of the relative abundance of bacterial phyla in the combined dataset.** The taxa and  
438 the labels are arranged as per total relative abundance across all samples, with the most abundant phyla at the bottom and  
439 the least abundant phyla at the top of the y-axis. Mann Whitney test was used to identify phyla with significantly different  
440 (p value < 0.05) relative abundance.

441

442 **Figure S2: Abundance of core microbiome of switchgrass.** The bar plot compares the relative abundance of  
443 switchgrass (red colored) core OTUs (90% threshold and  $q$ -value < 0.05) and non-switchgrass (yellow colored) samples.

444

445 **Figure S3: Core microbiome of switchgrass.** Phylogenetic tree showing relationships between core OTUs (90% thresh-  
446 old and  $q$ -value < 0.05) identified from switchgrass (blue colored) and non-switchgrass samples.

447

448

449 **Table S1: Processing times for COREMIC.**



**Root-zone associated core microbiome**

<b>Rows = 678*numb</b>	<b>Cols = 59*numb</b>	<b>Trial 1</b>	<b>Trial 2</b>	<b>Trial 3</b>	<b>Trial 4</b>	<b>Trial 5</b>	<b>Trial 6</b>	<b>Mean</b>	<b>Std. Er- ror</b>
1	1	13.102	12.017	12.015	12.314	11.924	11.603	12.163	0.210
2	1	28.426	26.511	27.832	28.623	25.742	30.245	27.896	0.655
10	1	37.913	84.115	41.965	70.986	43.540	46.456	54.163	7.671
1	2	12.924	13.924	12.914	14.639	16.016	17.961	14.730	0.802
1	10	30.127	41.331	24.405	32.020	34.582	48.253	35.120	3.467
2	2	29.118	29.512	29.586	34.621	36.447	35.057	32.390	1.359

450 The run times (in seconds) for different sized inputs with a 678 OTUs (rows) and 59 samples (columns) dataset using

451 default settings for COREMIC.

452

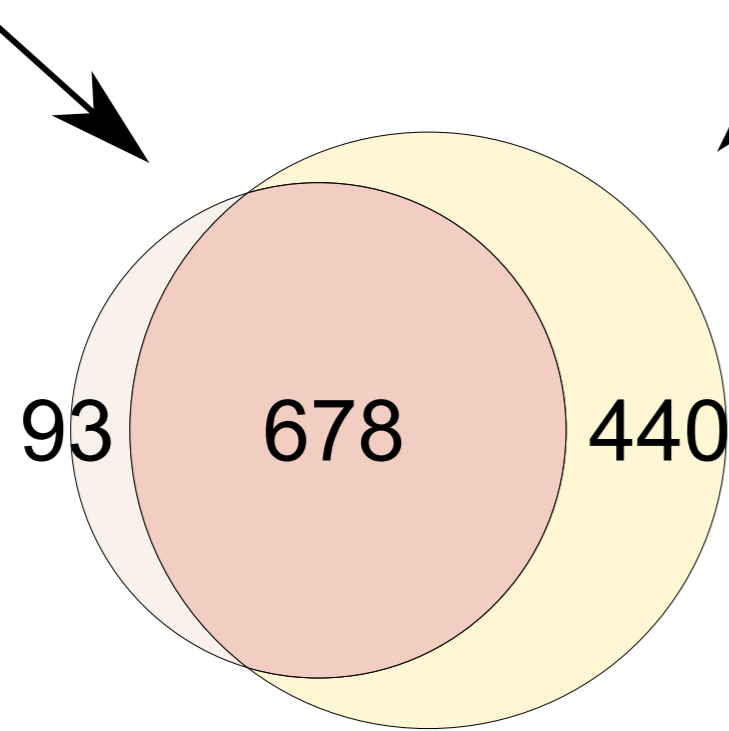
453

# Jesus

# Rodrigues

13 Swg, 28 Non-Swg  
771 OTUs

18 Swg  
1118 OTUs



678 common OTUs | absolute/relative abundance

### Original data (treated as binary)

	S-1	S-2	S-3	S-n	N-1	N-2	N-3	N-n
OTUx	1	1	1	1	0	0	1	0
OTUy	1	1	1	1	0	0	1	1
OTUz	1	1	1	1	1	0	1	1
OTUn	0	0	1	0	1	1	1	0

Fisher's exact test



### Core microbiome

OTUx OTUy

OTU is significant if  $q$ -value < 5%

### Jesus 2016

### Rodrigues 2017

Amplicon regions

V6-V8

V3-V4

Sequencing platform

Pyroseq

Illumina

Reads

Single

Paired

Lengths

~500 bp

~250 bp

Location

Wisconsin, Michigan

Virginia

Age

2 yrs, 10 yrs

1.5 months, 3.5 months

Site

Field

Greenhouse

Plants

Corn, Mixed grasses,  
Switchgrass, Praire grasses

Switchgrass

