CaptureSeq: Capture-based enrichment of *cpn60* gene fragments empowers pan-Domain profiling of microbial communities without universal PCR

Matthew G. Links [1,2], Tim J. Dumonceaux [3,4], Luke McCarthy [5], Sean M. Hemmingsen [5], Edward Topp [6], Alexia Comte [3], Jennifer R. Town [3*]

[1]Department of Animal and Poultry Science, University of Saskatchewan, Saskatoon, SK, Canada

[2]Department of Computer Science, University of Saskatchewan, Saskatoon, SK, Canada

[3]Agriculture and Agri-Food Canada, Saskatoon Research and Development Centre, Saskatoon, SK, Canada

[4]Department of Veterinary Microbiology, University of Saskatchewan, Saskatoon, SK, Canada

[5]National Research Council of Canada, Saskatoon, SK, Canada

[6]Agriculture and Agri-Food Canada, London Research and Development Centre, London, ON, Canada

*author for correspondence: jennifer.town@agr.gc.ca

Running title: Quantitative microbiome study using hybridization

DNA sequencing data associated with this work has been deposited at NCBI under BioProject PRJNA406970 and SRA deposits SRX3181274-SRX3181276 and SRX3187583-SRX3187601.

ABSTRACT

Molecular profiling of complex microbial communities has become the basis for examining the relationship between the microbiome composition, structure and metabolic functions of those communities. Microbial community structure can be partially assessed with universal PCR targeting taxonomic or functional gene markers. Increasingly, shotgun metagenomic DNA sequencing is providing more quantitative insight into microbiomes. Unfortunately both amplicon-based and shotgun sequencing approaches have significant shortcomings that limit the ability to study microbiome dynamics. We present a novel, amplicon-free, hybridization-based method (CaptureSeq) for profiling complex microbial communities using probes based on the chaperonin-60 gene. This new method generates a quantitative, pan-Domain community profile with significantly less expenditure and sequencing effort than a shotgun metagenomic sequencing approach. Molecular microbial profiles were compared for antibiotic-amended soil samples using CaptureSeq, shotgun metagenomics, and amplicon-based techniques. The CaptureSeq method generated a microbial profile that provided a much greater depth and sensitivity than shotgun metagenomic sequencing while simultaneously mitigating the bias effects associated with amplicon-based methods. The resulting community profile provided quantitatively reliable information about all three Domains of life (Bacteria, Archaea, and Eukarya). The applications of CaptureSeq are globally impactful and will facilitate highly accurate studies of host-microbiome interactions for environmental, crop, animal and human health.

INTRODUCTION

Life on Earth is classified into hierarchical taxonomic lineages that describe all living systems as having descended from a common ancestor along three evolutionary lines. Using ribosomal RNA-encoding gene sequences, Woese and Fox [1] delineated these Domains, which are now known as Bacteria, Archaea, and Eukarya [2]. Most complex microbial communities exist as assemblages replete with representatives from each of these Domains, the total genomic complement of which is called a microbiome. Understanding microbial community dynamics requires tools to examine the composition of these complex ecosystems. Advancements in DNA sequencing technology have created new opportunities to simplify the profiling of microbial communities from a diverse range of environments. As new insights are gained into the diversity of microbiomes in soil, water, plant and animal-associated ecosystems, we are collectively realizing the powerful effects that microbiome composition and structure can have on how these communities function [3]. To characterize the multifaceted relationships between microorganisms and their environment, it is critical to obtain a comprehensive microbial community profile that most accurately reflects its original composition and quantitative structure.

Microbiologists have increasingly embraced culture-independent methods of identification in recent decades [4]. By far the most commonly employed culture independent method is PCR-based amplification of informative gene sequences. In adapting the use of PCR for amplifying a conserved region of 16S rRNA, Weller and Ward provided the first example of microbial profiling [5]. More recently, Paul Hebert's proposed DNA barcoding criteria for Eukarya have established standards for what comprises a robust target for phylogenetic profiling [6]. Alternative universal gene markers for 16S [7], *cpn60* [8], *rpo*B [9], *mcr*A [10] and ITS [11] have been used for profiling microorganisms from bacterial, archaeal and eukaryotic Domains, however no single

amplification is able to profile microbes from all three Domains simultaneously. In order to obtain phylogenetic information for microorganisms across all three Domains of life, separate target amplification and processing protocols are required[12], increasing the cost and analytical complexity of accurately assessing dynamic changes in the community across Domains. Moreover, stochastic effects of primer interaction with a complex template, along with the difficulty in designing primers and amplification conditions that will equally target all members of a community[13], result in an unavoidable bias in community representation both in terms of presence/absence and relative abundances[13-16].

In recent years metagenomic approaches in which whole nucleic acid recovered from a sample is fragmented and sequenced using shotgun methods have become increasingly popular. This approach has a significant advantage over barcode-specific methods in that shotgun-sequencing data can overcome issues of bias and representation that are inherent in amplicon sequencing approaches, and provides the additional advantage of describing the metabolic potential of the microbial community [17-19]. Sequencing of all DNA present in an environmental sample can therefore be considered somewhat of a “gold standard” for taxonomic profiling. However, this approach is not without its own limitations. For example, it can be a wasteful enterprise in terms of the phylogenetic information recovered per sequencing cost. Shotgun sequencing is also not easily able to connect the functional potential observed in the sequencing data with the exact microbe within which that functionality resides. Additionally, DNA acquired from a community of microorganisms is inherently unbalanced; there are not equal numbers of each taxon, nor do all taxa have genomes that are of equal sizes. Thus shotgun sequencing can provide a view of microbial community composition that is biased by genome size and microbial abundances. Overcoming this bias requires significant amounts of sequencing; therefore, chasing

the rarity of the least abundant microbes by shotgun metagenomics sequencing carries a high financial cost [14,15,20,21]. The abundances of microbes within characterized complex microbial communities range over many orders of magnitude. While shotgun sequencing efforts provide a reasonable estimate of abundance there is a significant loss in dynamic range when compared to PCR-based profiling.

The chaperonin 60 gene [8] (type I chaperonin) and its Archaeal homologue thermosome complex [22] (type II chaperonin) have been previously recognized as highly discriminating targets across all three Domains of life [23], meet standard International Barcode of Life criteria [24] and enable *de novo* assembly of operational taxonomic units (OTU) [25]. While "universal" PCR primers are available [8,26], they are not expected to capture the pan-Domain diversity of a complex microbial community through amplification. Moreover, *cpn60* amplification provides OTU abundances that do not always correlate to the true abundance of the microorganism in the sample [27]. If these limitations can be overcome, there is significant opportunity to dramatically improve research assessing host-microbiome interactions in plant, human and animal settings.

Recent advances in hybridization-based DNA capture combined with high throughput sequencing (CaptureSeq), which have proven to be remarkably powerful means of enriching samples for DNA sequences of interest [28-30], led us to consider the possibility of exploiting the unique features of *cpn60* to provide a pan-Domain microbial community profile without the use of universal PCR amplifications. A custom array of biotinylated RNA capture baits was designed based on the entire taxonomic composition of the chaperonin database cpnDB (www.cpndb.ca) [8] and evaluated as a tool for enriching total genomic DNA simultaneously for type I and type II chaperonin target sequences. Samples were selected that encompassed taxonomic diversity across all three Domains of life. Soil samples comprised primarily of Bacteria, manure samples

with increased Archaeal diversity and a terrestrial pond sample with a larger number of Eukarya were used to compare the CaptureSeq method to standard shotgun metagenomic and amplicon-based approaches. The results indicate that CaptureSeq provides the taxonomic reach associated with shotgun metagenomic sequencing combined with the sampling depth of amplicon-based sequencing, giving an essentially complete, balanced, quantitatively accurate view of complex microbial ecosystems with reduced sequencing effort.

RESULTS

CaptureSeq generates Pan-Domain microbial community profiles

Microbial profiles were generated by CaptureSeq using samples from very different environmental ecosystems including soil, manure and a non-aerated terrestrial pond using CaptureSeq. These profiles provided a taxonomic overview of Bacteria, Archaea and Eukarya simultaneously, and identified sequencing reads from 9,361 (soil), 9,306 (manure), and 6,568 (pond) distinct taxonomic clusters (Supplemental Dataset S1). Additionally, the CaptureSeq profile facilitated inter-Domain comparisons of read abundances between taxonomic groups, since the abundances could be expressed in relation to the total pan-Domain community as opposed to reflecting only the proportions within a single Domain (Figure 1).

The soil sample microbiomes were composed primarily of Bacteria, with Proteobacteria and Actinobacteria comprising 60% and 25% of the pan-Domain community respectively. Members of the phyla Acidobacteria and Gemmatinomonadetes represented an additional 5% each of the microbiome. Total archaeal reads only accounted for 0.03-0.08% of the soil pan-Domain community, however there were still 165 archaeal taxonomic clusters identified in the soil. Eukarya represented 0.18-0.21% of the soil microbiome, with Fungi and Metazoa the most

abundant taxonomic groups. While the manure samples also contained a diverse array of Bacteria, they only represented 77-80% of the microbiome, compared to >99% for all of the soil samples. CaptureSeq libraries from the manure samples contained 19-22% archaeal reads, of which the vast majority were methanogens from the Phylum Euryarchaeota. The terrestrial pond contained a much greater proportion and diversity of Eukaryotes, representing 6.7% of the sequencing reads and 361 taxonomic clusters (Supplemental Dataset S1). *De novo* assembly of eukaryotic sequencing reads from the terrestrial pond sample generated 11 OTU most closely related to members of the Phylum Chlorophyta (green algae). Additionally, the assembly of OTU most similar to *Aenopholes* sp. (mosquitoes), and three members of the Phylum Alveolata (protists), suggests that CaptureSeq was able to retrieve *cpn60* DNA from higher level Eukarya. Compared to reference sequences in cpnDB, these *de novo* assembled OTU had nucleotide identities ranging from 59-84%, suggesting that the current probe array design and hybridization conditions were sufficiently permissive to allow capture of novel *cpn60* sequences (true unknowns).

CaptureSeq provides a similar microbial community profile to shotgun metagenomic sequencing

The complex taxonomic diversity found in soil provided an opportunity to determine if CaptureSeq yields a microbial community profile that accurately reflects the composition of the community and facilitates insights into the response of the communities to perturbation. Therefore, replicate plots amended with antibiotics were compared to control (unamended) soil samples using CaptureSeq, shotgun metagenomics, or *cpn60*-based amplicon sequencing techniques. In this setting, the ability of CaptureSeq to achieve in-depth sampling that is a more

accurate reflection of the community composition is critical to elucidate the effects of antimicrobial exposure on microbial ecosystem dynamics.

Both CaptureSeq and metagenomic techniques generated type I chaperonin sequences from all three Domains unencumbered by amplification and primer design biases. However, the number of chaperonin containing sequences represented only 0.08% of the total reads from the shotgun metagenomic library compared to an average of 16.7% (± 0.8%) for CaptureSeq and 94.8% (± 0.6%) for amplicon libraries (Supplemental Table S1). For a complex community such as soil, a greater sampling depth is required in order to make meaningful conclusions regarding microbial community composition and structure. Using a metagenomic approach requires orders of magnitude more sequencing effort to achieve a high level of community coverage and is not financially feasible for a large number of samples (Figure 2).

Examination of OTU abundance patterns revealed that the CaptureSeq and shotgun metagenomic profiled samples displayed patterns of microbial abundances that were more similar to one another and distinct from the pattern shown by the amplicon datasets (Figure 3). Moreover, of the three methods analyzed, only CaptureSeq showed a hierarchical clustering pattern that showed a difference between the antibiotic-treated and untreated soil samples (Figure 3). Similarly, when intra-technique beta diversity was assessed, only the CaptureSeq data provided measures that showed a separation of the soil samples by antibiotic treatment (Supplemental Figure S1). These results highlight the importance of profiling method on the ability to gain meaningful insights into microbiome structure and function.

Comparing alpha diversity metrics of the soil communities between the three profiling techniques suggested that both richness (Chao1) and diversity (Shannon H') were higher when

profiled using shotgun metagenomic compared to amplicon sequencing (Supplemental Figure S2). The CaptureSeq method provided alpha diversity metrics that were between those of the shotgun metagenomic shotgun method and amplicon sequencing (Supplemental Figure S2). Additionally, the alpha diversity metrics of the CaptureSeq method showed the least variability among the biological replicates of each treatment, even when libraries were down-sampled to very low levels (Supplemental Figure S2). Samples examined by *cpn60* amplification and sequencing displayed the highest inter-sample variability compared to CaptureSeq and metagenomic sequencing.

CaptureSeq permits de novo assembly of OTU from taxonomic clusters

To determine if *de novo* assembly of OTU representing individual organisms was reliable using CaptureSeq, we selected one target microorganism from each Domain for quantification using OTU-specific qPCR. For Bacteria, we quantified *Microbacterium* sp. C448, which was cultured from these soil samples and has previously been shown to degrade and metabolize the sulfonamide antibiotic added to the field plots[31]. While the presence of this target in the soil samples was confirmed using culture methods, it was under-represented in the amplicon and shotgun metagenomic libraries when compared to the CaptureSeq profiles. Only the CaptureSeq library provided a sufficient number of target sequencing reads for *de novo* assembly, generating a 1,066 bp OTU that was >99% identical to the *cpn60* sequence obtained from the genome of this organism [32]. We also assembled OTU targets from the Domains Eukarya (type I-*Phythophthora infestans*) and Archaea (type II-*Methanoculleus* sp.). Reads that mapped to the reference chaperonin sequences for these organisms were assembled *de novo* into OTU and were then quantified in each soil sample using ddPCR. Quantification of *Microbacterium* sp. C448 showed that the bacterium was present at a low level in all soil samples of between 10^3 and 10^4

196 gene copies per gram of soil, and that the levels were significantly higher in the antibiotic-treated
197 soil samples (Table 1). The archaeal OTU was quantified at levels between 495 and 527 gene
198 copies per gram of soil. The OTU corresponding to *P. infestans* was present at levels below the
199 limit of detection of ddPCR for these samples, yet was detectable by CaptureSeq (Table 1).
200 These results confirm the potential of the CaptureSeq method to almost completely sample
201 complex microbial communities with a limit of detection beyond the dynamic range of even very
202 sensitive quantification methods like ddPCR.

203 ***CaptureSeq provides a quantitatively accurate view of bacterial abundance***

204 Using a synthetic community of 20 microorganisms spiked into carrier DNA from a seed
205 wash facilitated a quantitative examination microbial community profiles using CaptureSeq.
206 Quantification of *cpn60* DNA from the synthetic community before and after hybridization using
207 qPCR revealed an enrichment of 3-4 orders of magnitude for *cpn60*-containing DNA fragments
208 compared to 16S rRNA-encoding genes (Supplemental Figure S3). For the 5 microorganisms
209 that were quantified, the ~10-fold reduction in gene copy number observed between the high,
210 medium, and low spike levels was consistent with the starting composition of the synthetic
211 community samples (Supplemental Figure S3). Furthermore, the number of *cpn60* gene copies
212 for the microorganisms added to the seed wash DNA extract was highly reproducible within each
213 spike level across the 1000-fold difference analyzed (Supplemental Figure S3). Across the
214 different spiking levels, there was a linear correlation between qPCR-determined input gene
215 copies and the number of sequencing reads observed for each of the five targets using the
216 CaptureSeq method, providing Pearson correlation coefficients (r^2) ranging from 0.995-1.000.
217 This compared to a range of 0.532-0.878 for libraries profiled by amplicon sequencing, with

more apparent distortion at the higher spike levels when targets were the most abundant (Figure 4).

While all 20 bacteria from the synthetic community were identified using both amplicon and CaptureSeq profiling techniques, only the CaptureSeq method generated profiles that accurately reflected the relative amounts of DNA spiked into the seed wash background (Figure 5 and Supplemental Table S2). In the CaptureSeq libraries, the number of mapped sequencing reads for each member of synthetic community was within one order of magnitude from the mean for each spike level. In the amplicon libraries however, the *cpn60* sequences of *Bifidobacterium infantis* and *Bifidobacterium bifidum*, which feature a high G/C content, were over 10- and 100-fold lower than the mean for both the High and Medium spiked samples (Supplemental Figure S4). This improved representation of high G/C Actinobacteria by CaptureSeq was also apparent in the microbial community profiles generated for the soil samples. Compared to the CaptureSeq libraries, the *cpn60* sequences of the 25 most under-represented taxonomic clusters in the amplicon libraries had very high G/C content (64-71%) and included several members of the genera *Nocardioides*, *Marmoricola* and *Pseudonocardia* (Supplemental Table S3).

De novo assembly of the mapped sequencing reads for each microorganism from the synthetic panel for both amplicon and CaptureSeq libraries generated OTU that were >99% identical to the known *cpn60* sequences.

DISCUSSION

Targeted capture of *cpn60* gene fragments resulted in an approximately 200-fold enrichment of the soil samples for the taxonomic marker of interest, from under 0.1% of reads in

the shotgun metagenomic sequencing to over 15% of reads in the CaptureSeq datasets. This level of enrichment enabled very deep sampling of the soil microbial communities (similar to that attained using PCR-based enrichment) with far less sequencing data (i.e. a significant cost savings). This is of particular importance when the organisms of interest are very low in abundance, such as *Microbacterium* sp. C448 in this study. OTU were observed in the CaptureSeq datasets that were present at extremely low levels in the soil genomic DNA, near or below the detection limit for ddPCR. Based on the assay setup and dilution factors we used, the theoretical ddPCR detection limit was 3570 copies/g soil, assuming detection of 10 copies per assay[33]. Although increased sequencing effort can result in more complete coverage of complex microbial communities using shotgun metagenomic sequencing[15,21], application of this method to investigate the taxonomic composition of a sample is not an efficient use of budgetary resources. In addition, CaptureSeq provided a balanced view of the relative abundances of microorganisms within the community. PCR-associated representational bias, which presents a skewed representation of microbial taxon abundance [34], is a well-known phenomenon[35-37], and is likely the result of using end-point PCR product to generate the sequencing library as the exponential accumulation of amplicon serves to compress the dynamic range of relative DNA abundance in the end product of the reaction. CaptureSeq also resulted in an improvement of the representation of high G/C content microorganisms compared to amplification. Difficulty in amplification of high G/C content targets is a phenomenon that has been previously observed using both 16S and *cpn60* taxonomic markers from mixed communities [26,38]. *De novo* assembly of taxonomic clusters from the CaptureSeq datasets into OTU for which probes were not explicitly designed, such as *Microbacterium* sp. C448, also suggests that off-target *cpn60* sequence capture can expand the breadth of OTU observed in the dataset beyond the sequences represented in the

263 probe array and can include sequences that have not been previously observed. While
264 CaptureSeq may be biased by the probe sequences employed, it is clearly capable of detecting
265 novel microbes, expanding the breadth of microorganisms that are included in the microbial
266 community profile beyond microbes that have been previously identified.

267 The overall patterns of OTU abundances in each of the three methods showed that the
268 amplicon-based method provided a pattern that was distinct from the patterns observed for both
269 CaptureSeq and shotgun metagenomic sequencing, which were more similar to one another.
270 While the three methods all provided discernably different overall community profiles, the
271 difference observed in the relative abundances of microorganisms was likely the result of
272 different biases inherent in each of the methods. The over-representation in the amplicon datasets
273 of several of the microorganisms that were very rare in the metagenomic and CaptureSeq
274 libraries was likely the result of amplification effects on the relative abundances of
275 microorganisms [16,39]. PCR amplification also introduced a higher experimental error in various
276 alpha diversity parameters (Chao1, Shannon, Simpson) among the biological replicates analyzed
277 compared to CaptureSeq and shotgun metagenomic sequencing. This observation is consistent
278 with previous studies using 16S rRNA amplicon profiles of soil communities [16,40]. Among the
279 three methods, CaptureSeq displayed the lowest inter-sample variation for these diversity
280 parameters. CaptureSeq therefore has the potential to improve insight into microbial community
281 dynamics by reducing experimental variability, and thereby improving reproducibility, compared
282 to both amplicon-based and shotgun metagenomic sequencing. The consistency in alpha
283 diversity calculations is likely a reflection of the reduced biases inherent in the CaptureSeq
284 protocol and facilitates making meaningful conclusions about community richness and diversity.

The *cpn60* taxonomic marker enables *de novo* assembly of OTU [23,25] providing greater discrimination between closely related microorganisms and facilitating OTU-specific assay design. The *cpn60*-based CaptureSeq approach generates assembled chaperonin sequences that may also include regions flanking the sequence amplified by the universal primers, as observed with the OTU over 1 kb in length generated for *Microbacterium* sp. C448 and *Methanoculleus marisnigri* in this study. This additional sequencing information can provide further taxonomic discrimination of many prokaryotes, especially if the assembled region includes the *cpn10* co-chaperonin that is adjacent to *cpn60* in many bacterial genomes [41]. The OTU that were *de novo* assembled provided suitable targets for ddPCR, facilitating the enumeration of targeted microorganisms from each Domain, which had initially been identified by sequencing and assembly. Such an approach can be used to identify biological interactions between/among microorganisms that can explain their relative abundance patterns [23].

Both CaptureSeq and shotgun metagenomic sequencing provided the means to identify OTU from all Domains simultaneously, facilitating the characterization of inter-Domain relationships among microorganisms. The ability to calculate the abundances of organisms as a proportion of the entire pan-Domain community facilitates the identification of inter-Domain relationships and syntrophies. This is of particular importance in many settings (e.g. manure or gut health) in identifying the syntrophic relationships between volatile fatty acid producing Bacteria and methanogenic Archaea [42]. In soil, the complex relationship between saprophytic Fungi and Bacteria is critical to examining the role of the microbiome in nutrient cycling [43]. Similarly in the terrestrial pond, the bacterial and eukaryotic components of the microbial ecosystems can be directly compared numerically, which may allow insights into inter-Domain relationships that impact elemental cycles or other ecosystem services. This advantage is not

offered using amplification of universal targets, although it does provide the benefit of very deep coverage of complex microbial communities. Shotgun metagenomic genome sequencing does not provide the community coverage of either the amplicon-based or CaptureSeq methods at a similar sequencing effort, suggesting that complex microbiomes will likely require additional phylogenetic data to make any informed examination of microbial diversity metrics. CaptureSeq enabled deep coverage of complex microbial communities, although the community representation is naturally biased by the hybridization probes used. However, we observed off-target hybridization, as evidenced by the appearance of *cpn60* OTU in the CaptureSeq datasets. Optimizing the hybridization parameters may result in further improvements to the enrichment of taxonomic markers in complex templates, increasing the efficiency of this approach to microbial community profiling. Shotgun metagenomics can reasonably be considered the least biased means of determining the taxonomic composition of an environmental sample, and may be a suitable choice when sufficient sequencing resources are available. However the abiding popularity of amplicon-based profiling is at least partially a result of the high degree of enrichment of taxonomically informative sequence reads that it generates. CaptureSeq provides an alternative that avoids the amplification biases associated with PCR while retaining the sequencing efficiency of amplicon-based profiling.

Molecular microbial community profiling is one of the foundational steps in exploring microbiome structure-function relationships in an experimental system [44-46]. To generate and evaluate scientific hypotheses it is critical to generate a microbiome profile that reflects the natural state a closely as possible with sufficient sensitivity to evaluate both abundant and rare microorganisms. The *cpn60*-based method described herein permits taxonomically broad and deep microbial community profiling of complex microbiomes. Thus CaptureSeq has the potential

to impact life sciences research wherever microbes are thought to be important, including human health and nutrition [47], agriculture [48], biotechnology [49], and environmental sciences [50]. Several methodologies are available for microbial community profiling, including 16S and ITS amplification and sequencing, as well as profiling using 16S rRNA-based capture probes [30]. While all microbial community profiling techniques have inherent limitations and biases, compared to shotgun metagenomic and universal target amplification, CaptureSeq is a suitable alternative that provides quantitative, pan-Domain analysis of complex communities.

MATERIALS AND METHODS

Soil sample preparation

Soil samples were obtained from a long-term study initiated in 1999 evaluating the effect of annual antibiotic exposure on soil microbial communities, described in Cleary *et al.* [51]. Soil samples evaluated in the present study were obtained in 2013 following 15 sequential annual applications of a mixture of sulfamethazine, chlortetracycline and tylosin, each added at 10 mg kg^{-1} soil. Soil was sampled 30 days after the spring application of antibiotics. The plots were planted with soybeans (*Glycine max*, v. Harosoy) immediately after incorporation of the antibiotics. One triplicate group of plots had experienced no antibiotic treatment, and the other triplicate set had received yearly antibiotic treatments since 1999 as described [51]. Genomic DNA was extracted from 3.5 g of each soil sample using the PowerMax Soil DNA isolation kit (Mo-Bio Laboratories, Carlsbad, CA) with a 5 mL elution volume. DNA extracts were quantified using a Qubit fluorimeter (Thermo Fisher Scientific, Waltham, MA, USA) and stored at -80°C until processing and analysis.

Terrestrial pond sample preparation

353 A water sample was obtained from a pond located on a Saskatchewan farm (51.99°N, -
354 106.46°W) on May 13, 2016. Biological material was recovered from 2L of water by
355 centrifugation at 20,000 *g* for 20 minutes. Total DNA was extracted using a PowerWater DNA
356 extraction kit (Mo-Bio Laboratories, Carlsbad, CA) and quantified as described above.

357 ***Seed wash carrier DNA preparation***

358 Genomic DNA to act as carrier DNA for spiking 10-fold decreasing amounts of a
359 synthetic community was generated by washing wheat seeds as previously described [23], and
360 known to lack all of the microorganisms comprising the synthetic community panel [23].

361 ***Synthetic community sample preparation***

362 Amplicons corresponding to the *cpn60* UT of 20 bacteria associated with the human
363 vaginal tract [25] were cloned into the pGEM-T Easy plasmid (Promega, WI, USA) and purified
364 using the Qiagen Miniprep kit (Qiagen, CA, USA). The synthetic community was formed by
365 combining equimolar concentrations of plasmids containing the *cpn60* UT for all 20
366 microorganisms [25]. Dilutions of this mixture (corresponding to 0.4, 0.04, and 0.004 ng plasmid
367 DNA, or approximately 10^8, 10^7, and 10^6 copies of each plasmid) were spiked into a background
368 of 10 ng/µl of wheat seed carrier DNA. Spiked genomic DNA samples prepared in this way were
369 sequenced using *cpn60* universal target amplification and CaptureSeq as described below.

370 The efficacy of the CaptureSeq hybridization was assessed prior to sequencing using
371 quantitative PCR (qPCR) targeting plasmids added to the seed wash background. qPCR primers
372 and amplification conditions were as described previously [52]. Total bacteria were enumerated
373 using qPCR targeting the 16S ribsosomal RNA-encoding gene as described previously [53].

374 ***Amplicon-based sequencing***

375 The *cpn60* UT was amplified from synthetic community-spiked DNA or soil genomic
376 DNA samples using 40 cycles of PCR with the type I chaperonin universal primer cocktail
377 containing a 1:3 ratio of H279/H280:H1612/H1613 [26] and cycling conditions of 1x 95°C, 5 min;
378 40x 95°C 30sec, 42-60°C 30sec, 72°C 30sec; 1x 72°C 2min. Replicate reactions from each
379 amplification temperature for each sample were pooled and gel purified using the Blue Pippin
380 Prep system (Sage Science, MA, USA) with a 2% agarose cassette, and concentrated using
381 Amicon 30K 0.5 ml spin columns (EMD Millipore, MA, USA). Amplicon from all samples was
382 prepared for sequencing using the NEBNext Illumina library preparation kit (New England
383 Biolabs, location), and sequenced with 400 forward cycles of v2 Miseq chemistry.

384 ***CaptureSeq array design***

385 Capture probes were designed based on all type I and type II chaperone sequences in the
386 public domain (i.e. CpnDB; www.cpndb.ca)[8]. 15,733 probes were designed to be complementary
387 to the type I and type II chaperone sequences. Design of probes was based on identifying 120bp
388 sequences from the reference database using a 60bp incrementing step. Thus the resulting probes
389 should share a 50% overlap with the next probe in a tiling-like fashion. The custom oligos were
390 bound to magnetic beads in equimolar concentration as a custom Mybaits array by Mycoarray
391 (Ann Arbor, MI, USA).

392 ***Shotgun metagenomic sequencing and CaptureSeq preparation***

393 Genomic DNA from each of the soil samples was diluted to 2.5 ng/µl and split into two
394 aliquots of 100 µl each for shearing using a water bath sonicator as described [54]. Shotgun
395 metagenomic genomic sequencing libraries were prepared directly from one aliquot of each

sheared genomic DNA sample using the NEBNext Illumina library preparation kit according to the manufacturer's directions (New England Biolabs, MA, USA). Samples were then sequenced with 2x250 bp cycles of v2 Miseq chemistry (Illumina, CA, USA).

To generate the CaptureSeq libraries, the second aliquots of sheared genomic DNA samples were subjected to end repair and index addition using NEBNext as above, then hybridized to the capture probe array as described [54]. The chaperonin-enriched products were then sequenced with 2x250 bp cycles of v2 Miseq chemistry (Illumina, CA, USA).

Sequencing analysis

To compare the number of output sequencing reads for the different spiking levels, sequencing reads from the synthetic community-spiked samples were down-sampled to the smallest library size for each profiling technique (30,091 for amplicon and 506,247 for CaptureSeq) and mapped to a reference set of *cpn60* UT sequences for the 20 microorganisms in the panel by local paired alignment using bowtie2 (v. 2.2.3) [55].

A reference database of all publically available chaperonin sequences was generated by selecting a list of seven chaperonin protein sequences representing each taxonomic group: fungi, bacteria, archaea, plant mitochondria, plant chloroplast, and animal mitochondria. These probes were used as queries for a BLAST search of GenBank using the default parameters to blastp. Matching protein sequences were manually vetted to generate a list of 30,141 protein identifiers. These protein identifiers were then used to retrieve the corresponding 30,120 nucleotide sequences available in GenBank according to the procedure described in Supplemental Information. The accession numbers of those nucleotide sequences are provided in Supplemental Dataset S2. The breadth of taxa that were retrieved by this method was similar to the taxonomic

418 breadth represented in the 16S and ITS reference datasets (Supplemental Dataset S3).
419 Sequencing reads from all soil samples were grouped into taxonomic clusters by paired local
420 alignment to this reference set of chaperonin genes using bowtie2. The sequencing libraries were
421 down-sampled to the size of the smallest shotgun metagenomic library (2,777 mapped paired
422 reads), and the relative abundances of each of the resulting taxonomic clusters was used as the
423 basis for assessing the alpha and beta diversity metrics of the three profiling methods for
424 equivalent sampling effort.

425 ***De novo OTU assembly and quantification***

426 Read pairs from target taxonomic clusters were assembled *de novo* into *cpn60* OTU using
427 Trinity (v. 2.4.0) with a kmer of 31. OTU-specific primer and hydrolysis probe sets were
428 designed using Primer3 [56] or Beacon Designer (v.7) (Premier Biosoft, Palo Alto, CA, USA) as
429 described previously [57]. Annealing temperatures were optimized for each reaction using gradient
430 PCR with ddPCR Supermix for Probes (Bio-Rad, Mississauga, ON, Canada) using 900 nM each
431 primer and 250 nM of hydrolysis probe in a 20 µl reaction volume. Primer/probe sequences and
432 optimized amplification conditions are shown in Supplemental Table S1. Template DNA was
433 digested prior to amplification using *Eco*RI at 37°C for 60 minutes. A final volume of 2-5 µl was
434 used as template for droplet digital PCR (ddPCR). Emulsions were formed using a QX100
435 droplet generator (Bio-Rad, Hercules, CA, USA), and amplifications were carried out using a
436 C1000 Touch thermocyler (Bio-Rad). Reactions were analyzed using a QX100 droplet reader
437 (Bio-Rad) and quantified using QuantaSoft (v.1.6.6) (Bio-Rad). Results were converted to copy
438 number/g soil extracted by accounting for sample preparation and dilution. For the prepared
439 CaptureSeq libraries, results were converted to copy number/µl by considering dilution factors.

440 ***Alpha diversity analysis***

To compare the richness and diversity metrics between the three profiling techniques, mapped sequencing reads were down-sampled from 250-2,750 reads to simulate a uniform sampling effort across profiling techniques. Metrics were averaged across 100 bootstrapped datasets using the multiple_rarefactions.py and alpha_diversity.py scripts from QIIME (v. 1.8.0) [58].

In the cases where the total effect of sequencing effort was required for comparisons across estimates of community coverage read thresholds were transformed to reflect total sequencing effort for each sample.

Beta diversity analysis

To compare the community similarity between different sequencing methods, mapped sequencing reads were down-sampled to the size of the smallest metagenomic library sample (2,777 mapped reads). For intra-technique comparisons, mapped sequencing reads were down-sampled to the smallest library size within each profiling method; 2,777 for metagenomic, 127,642 for CaptureSeq, and 27,388 reads for amplicon libraries. Principal Coordinate Analysis of inter- and intra-technique Bray-Curtis distance was calculated using the vegan package (v. 2.4.2) in R (v. 3.2.4).

REFERENCES

1 Woese, C. R. & Fox, G. E. Phylogenetic structure of the prokaryotic domain: The primary kingdoms. *Proc. Natl. Acad. Sci. U.S.A.* **74**, 5088-5090, doi:10.1073/pnas.74.11.5088 (1977).

2 Woese, C. R., Kandler, O. & Wheelis, M. L. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. *Proc. Natl. Acad. Sci. U.S.A.* **87**, 4576-4579, doi:10.1073/pnas.87.12.4576 (1990).

3 Tikhonovich, I. A. & Provorov, N. A. Microbiology is the basis of sustainable agriculture: An opinion. *Ann. Appl. Biol.* **159**, 155-168, doi:10.1111/j.1744-7348.2011.00489.x (2011).

4 J T Staley, a. & Konopka, A. Measurement of *in situ* activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* **39**, 321-346, doi:10.1146/annurev.mi.39.100185.001541 (1985).

5 Weller, R. & Ward, D. M. Selective recovery of 16S rRNA sequences from natural microbial communities in the form of cDNA. *Appl. Environ. Microbiol.* **55**, 1818-1822 (1989).

6 Hebert, P. D. N., Cywinska, A., Ball, S. L. & deWaard, J. R. Biological identifications through DNA barcodes. *Proc R Soc Lond [Biol]* **270**, 313-321, doi:10.1098/rspb.2002.2218 (2003).

7 Singer, E. *et al.* High-resolution phylogenetic microbial community profiling. *ISME J* **10**, 2020-2032, doi:10.1038/ismej.2015.249 (2016).

8 Hill, J. E., Penny, S. L., Crowell, K. G., Goh, S. H. & Hemmingsen, S. M. cpnDB: A chaperonin sequence database. *Genome Res.* **14**, 1669-1675 (2004).

9 Adékambi, T., Drancourt, M. & Raoult, D. The *rpoB* gene as a tool for clinical microbiologists. *Trends Microbiol.* **17**, 37-45 (2009).

10 Barret, M. *et al.* Identification of *Methanoculleus* spp. as active methanogens during anoxic incubations of swine manure storage tank samples. *Appl. Environ. Microbiol.* **79**, 424-433 (2013).

11 Schoch, C. L. *et al.* Nuclear ribosomal internal transcribed spacer (ITS) region as a universal DNA barcode marker for Fungi. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 6241-6246, doi:10.1073/pnas.1117018109 (2012).

12 Barret, M. *et al.* Emergence shapes the structure of the seed-microbiota. *Appl. Environ. Microbiol.* **81**, 1257-1266 (2015).

13 Walker, A. W. *et al.* 16S rRNA gene-based profiling of the human infant gut microbiota is strongly influenced by sample processing and PCR primer choice. *Microbiome* **3**, 26, doi:10.1186/s40168-015-0087-4 (2015).

14 Guo, J., Cole, J. R., Zhang, Q., Brown, C. T. & Tiedje, J. M. Microbial community analysis with ribosomal gene fragments from shotgun metagenomes. *Appl. Environ. Microbiol.* **82**, 157-166, doi:10.1128/aem.02772-15 (2016).

15 Lynch, M. D. J. & Neufeld, J. D. Ecology and exploration of the rare biosphere. *Nat Rev Micro* **13**, 217-229, doi:10.1038/nrmicro3400 (2015).

16 Poretsky, R., Rodriguez-R, L. M., Luo, C., Tsementzi, D. & Konstantinidis, K. T. Strengths and limitations of 16S rRNA gene amplicon sequencing in revealing temporal microbial community dynamics. *PLOS ONE* **9**, e93827, doi:10.1371/journal.pone.0093827 (2014).

17 Hess, M. *et al.* Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. *Science* **331**, 463-467 (2011).

18 Raymond, F. *et al.* The initial state of the human gut microbiome determines its reshaping by antibiotics. *ISME J* **10**, 707-720, doi:10.1038/ismej.2015.148 (2016).

19 Handley, K. M. *et al.* Biostimulation induces syntrophic interactions that impact C, S and N cycling in a sediment microbial community. *ISME J.* **7**, 800-816, doi:10.1038/ismej.2012.148 (2013).

20 Fierer, N. *et al.* Cross-biome metagenomic analyses of soil microbial communities and their functional attributes. *Proc. Natl. Acad. Sci. U.S.A.* **109**, 21390-21395, doi:10.1073/pnas.1215210110 (2012).

21 Luo, C. *et al.* Soil microbial community responses to a decade of warming as revealed by comparative metagenomics. *Appl. Environ. Microbiol.* **80**, 1777-1786, doi:10.1128/aem.03712-13 (2014).

22 Chaban, B. & Hill, J. E. A 'universal' type II chaperonin PCR detection system for the investigation of Archaea in complex microbial communities. *ISME J.* **6**, 430-439 (2012).

23 Links, M. G. *et al.* Simultaneous profiling of seed-associated bacteria and fungi reveals antagonistic interactions between microorganisms within a shared epiphytic microbiome on *Triticum* and *Brassica* seeds. *New Phytol.* **202**, 542-553, doi:10.1111/nph.12693 (2014).

24 Links, M. G., Dumonceaux, T. J., Hemmingsen, S. M. & Hill, J. E. The chaperonin-60 universal target is a barcode for bacteria that enables *de novo* assembly of metagenomic sequence data. *PLoS One* **7**, e49755, doi:10.1371/journal.pone.0049755 (2012).

25 Links, M. G., Chaban, B., Hemmingsen, S., Muirhead, K. & Hill, J. mPUMA: a computational approach to microbiota analysis by de novo assembly of operational taxonomic units based on protein-coding barcode sequences. *Microbiome* **1**, 23 (2013).

26 Hill, J. E., Town, J. R. & Hemmingsen, S. M. Improved template representation in *cpn60* polymerase chain reaction (PCR) product libraries generated from complex templates by application of a specific mixture of PCR primers. *Environ. Microbiol.* **8**, 741-746, doi:10.1111/j.1462-2920.2005.00944.x (2006).

27 Dumonceaux, T. J., Hill, J. E., Hemmingsen, S. M. & Van Kessel, A. G. Characterization of intestinal microbiota and response to dietary virginiamycin supplementation in the broiler chicken. *Appl. Environ. Microbiol.* **72**, 2815-2823 (2006).

28 Schuenemann, V. J. *et al.* Targeted enrichment of ancient pathogens yielding the pPCP1 plasmid of *Yersinia pestis* from victims of the Black Death. *Proc. Natl. Acad. Sci. U.S.A.* **108**, E746–E752, doi:10.1073/pnas.1105107108 (2011).

29 Wagner, D. M. *et al. Yersinia pestis* and the Plague of Justinian 541-543 AD: a genomic analysis. *Lancet Infect Dis* (2014).

30 Gasc, C. & Peyret, P. Hybridization capture reveals microbial diversity missed using current profiling methods. *Microbiome* **6**, 61, doi:10.1186/s40168-018-0442-3 (2018).

31 Topp, E. *et al.* Accelerated biodegradation of veterinary antibiotics in agricultural soil following long-term exposure, and isolation of a sulfamethazine-degrading *Microbacterium* sp. *J. Environ. Qual.* **42**, 173-178, doi:10.2134/jeq2012.0162 (2013).

32 Martin-Laurent, F., Marti, R., Waglechner, N., Wright, G. D. & Topp, E. Draft Genome Sequence of the Sulfonamide Antibiotic-Degrading Microbacterium sp. Strain C448. *Genome Announcements* **2**, e01113-01113, doi:10.1128/genomeA.01113-13 (2014).

33 Bustin, S. A. *et al.* The MIQE guidelines: minimum information for publication of quantitative real-time PCR experiments. *Clin Chem* **55**, 611-622 (2009).

34 Props, R. *et al.* Absolute quantification of microbial taxon abundances. *ISME J*, doi:10.1038/ismej.2016.117 (2016).

35 Johnson, L. A., Chaban, B., Harding, J. C. & Hill, J. E. Optimizing a PCR protocol for cpn60-based microbiome profiling of samples variously contaminated with host genomic DNA. *BMC research notes* **8**, 253, doi:10.1186/s13104-015-1170-4 (2015).

36 Green, S. J., Venkatramanan, R. & Naqib, A. Deconstructing the polymerase chain reaction: Understanding and correcting bias associated with primer degeneracies and primer-template mismatches. *PLoS ONE* **10**, doi:10.1371/journal.pone.0128122 (2015).

37 Lee, C. K. *et al.* Groundtruthing next-gen sequencing for microbial ecology-biases and errors in community structure estimates from PCR amplicon pyrosequencing. *PLoS ONE* **7**, doi:10.1371/journal.pone.0044224 (2012).

38 Pinto, A. J. & Raskin, L. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. *PLOS ONE* **7**, e43093, doi:10.1371/journal.pone.0043093 (2012).

39 Logares, R. *et al.* Metagenomic 16S rDNA Illumina tags are a powerful alternative to amplicon sequencing to explore diversity and structure of microbial communities. *Environ. Microbiol.* **16**, 2659-2671, doi:10.1111/1462-2920.12250 (2014).

40 Ranjan, R., Rani, A., Metwally, A., McGee, H. S. & Perkins, D. L. Analysis of the microbiome: Advantages of whole genome shotgun versus 16S amplicon sequencing. *Biochem. Biophys. Res. Commun.* **469**, 967-977, doi:http://dx.doi.org/10.1016/j.bbrc.2015.12.083 (2016).

41 Chaban, B., Links, M. & Hill, J. A molecular enrichment strategy based on *cpn60* for detection of Epsilon-Proteobacteria in the dog fecal microbiome. *Microb. Ecol.* **63**, 348-357, doi:10.1007/s00248-011-9931-7 (2012).

42 Demirel, B. & Scherer, P. The roles of acetotrophic and hydrogenotrophic methanogens during anaerobic conversion of biomass to methane: a review. *Rev. Environ. Sci. Biotechnol.* **7**, 173-190 (2008).

43 de Menezes, A. B., Richardson, A. E. & Thrall, P. H. Linking fungal–bacterial co-occurrences to soil ecosystem function. *Curr. Opin. Microbiol.* **37**, 135-141, doi:http://dx.doi.org/10.1016/j.mib.2017.06.006 (2017).

44 Carballa, M., Regueiro, L. & Lema, J. M. Microbial management of anaerobic digestion: exploiting the microbiome-functionality nexus. *Curr. Opin. Biotechnol.* **33**, 103-111, doi:http://dx.doi.org/10.1016/j.copbio.2015.01.008 (2015).

45 Gopal, M. & Gupta, A. Microbiome selection could spur next-generation plant breeding strategies. *Frontiers in microbiology* **7**, doi:10.3389/fmicb.2016.01971 (2016).

46 Muegge, B. D. *et al.* Diet drives convergence in gut microbiome functions across mammalian phylogeny and within humans. *Science* **332**, 970-974, doi:10.1126/science.1198719 (2011).

47 Kau, A. L., Ahern, P. P., Griffin, N. W., Goodman, A. L. & Gordon, J. I. Human nutrition, the gut microbiome and the immune system. *Nature* **474**, 327-336, doi:10.1038/nature10213 (2011).

48 Busby, P. E. *et al.* Research priorities for harnessing plant microbiomes in sustainable agriculture. *PLOS Biology* **15**, e2001793, doi:10.1371/journal.pbio.2001793 (2017).

49 Koch, C., Müller, S., Harms, H. & Harnisch, F. Microbiomes in bioenergy production: From analysis to management. *Curr. Opin. Biotechnol.* **27**, 65-72, doi:10.1016/j.copbio.2013.11.006 (2014).

50 Fierer, N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nature reviews. Microbiology* (2017).

51 Cleary, D. W. *et al.* Long-term antibiotic exposure in soil is associated with changes in microbial community structure and prevalence of class 1 integrons. *FEMS Microbiol Ecol* **92**, doi:10.1093/femsec/fiw159 (2016).

52 Dumonceaux, T. J. *et al.* Multiplex detection of bacteria associated with normal microbiota and with bacterial vaginosis in vaginal swabs by use of oligonucleotide-coupled fluorescent microspheres. *J Clin Microbiol* **47**, 4067-4077, doi:10.1128/jcm.00112-09 (2009).

53 Lee, D. H., Zo, Y. G. & Kim, S. J. Nonradioactive method to study genetic profiles of natural bacterial communities by PCR-single-strand-conformation polymorphism. *Appl. Environ. Microbiol.* **62**, 3112-3120 (1996).

54 Dumonceaux, T. J., Links, M. G., Town, J. R., Hill, J. E. & Hemmingsen, S. M. Targeted capture of *cpn60* gene fragments for PCR-independent microbial community profiling. *Protoc exch* (2017).

598 55 Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of
599 short DNA sequences to the human genome. *Genome Biol* **10**, R25 (2009).
600 56 Rozen, S. & Skaletsky, H. Primer3 on the WWW for general users and for biologist programmers.
601 *Methods in molecular biology (Clifton, N.J.)* **132**, 365-386 (2000).
602 57 Pérez-López, E., Hammond, C., Olivier, C. Y. & Dumonceaux, T. J. in *Diagnostic Bacteriology* Vol.
603 1616 *Methods in Molecular Biology* (ed K.A. Bishop-Lilly) (Humana Press, 2017).
604 58 Caporaso, J. G. *et al.* QIIME allows analysis of high-throughput community sequencing data.
605 *Nature Meth.* **7**, 335-336 (2010).

606 **Authors' contributions**

607 ET, SH, TD and AC performed collection, processing and sequencing of all samples. ML, LM
608 and JT performed bioinformatics analysis of sequencing data. All authors contributed to writing
609 the manuscript.

610 **Acknowledgements**

611 This work was funded through Agriculture and Agri-Food Canada A-base project 1562:
612 Optimizing soil health and protecting environmental quality through judicious manure
613 management, and innovative cover cropping.

614 **Competing interests**

615 The author(s) declare no competing financial or non-financial interests.

616

617 **Figure Legends:**

618 **Figure 1:** CaptureSeq was used to simultaneously profile Bacteria, Archaea, and Eukarya from
619 an ecologically diverse range of samples including soil (n=6), manure (n=3), and a freshwater
620 pond (n=1). The relative abundances of individual Phyla were expressed as a proportion of the
621 entire pan-Domain microbial community.

622 **Figure 2:** Good's coverage estimate reflecting the average total sequencing effort for six soil
623 samples each profiled using amplicon (red), CaptureSeq (blue), or shotgun metagenomic (green)
624 approaches.

625 **Figure 3:** Proportional abundance of taxonomic clusters for type I chaperonins in soil samples
626 profiled using amplicon, CaptureSeq, or shotgun metagenomic approaches. Samples were
627 clustered based on Bray-Curtis distance, and reference clusters composing a minimum of 0.5%
628 of the mapped sequencing reads in any one sample are shown.

629 **Figure 4:** The correlation between input *cpn60* gene copies quantified by species-specific
630 quantitative PCR and the number of mapped sequencing reads was determined for 5 bacteria
631 from the synthetic community.

632 **Figure 5**: Sequencing read abundance for seed wash samples spiked with a synthetic community
633 of 20 bacteria in 10-fold decreasing dilutions and were profiled using UT amplification or
634 CaptureSeq profiling methods. The color scale represents the $\log_{10}$ read abundance in the
635 sequencing library.

636

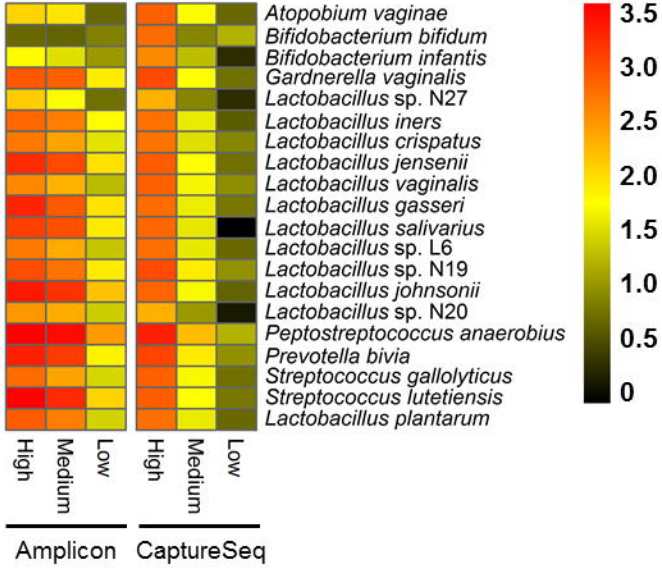
637 **Tables:**

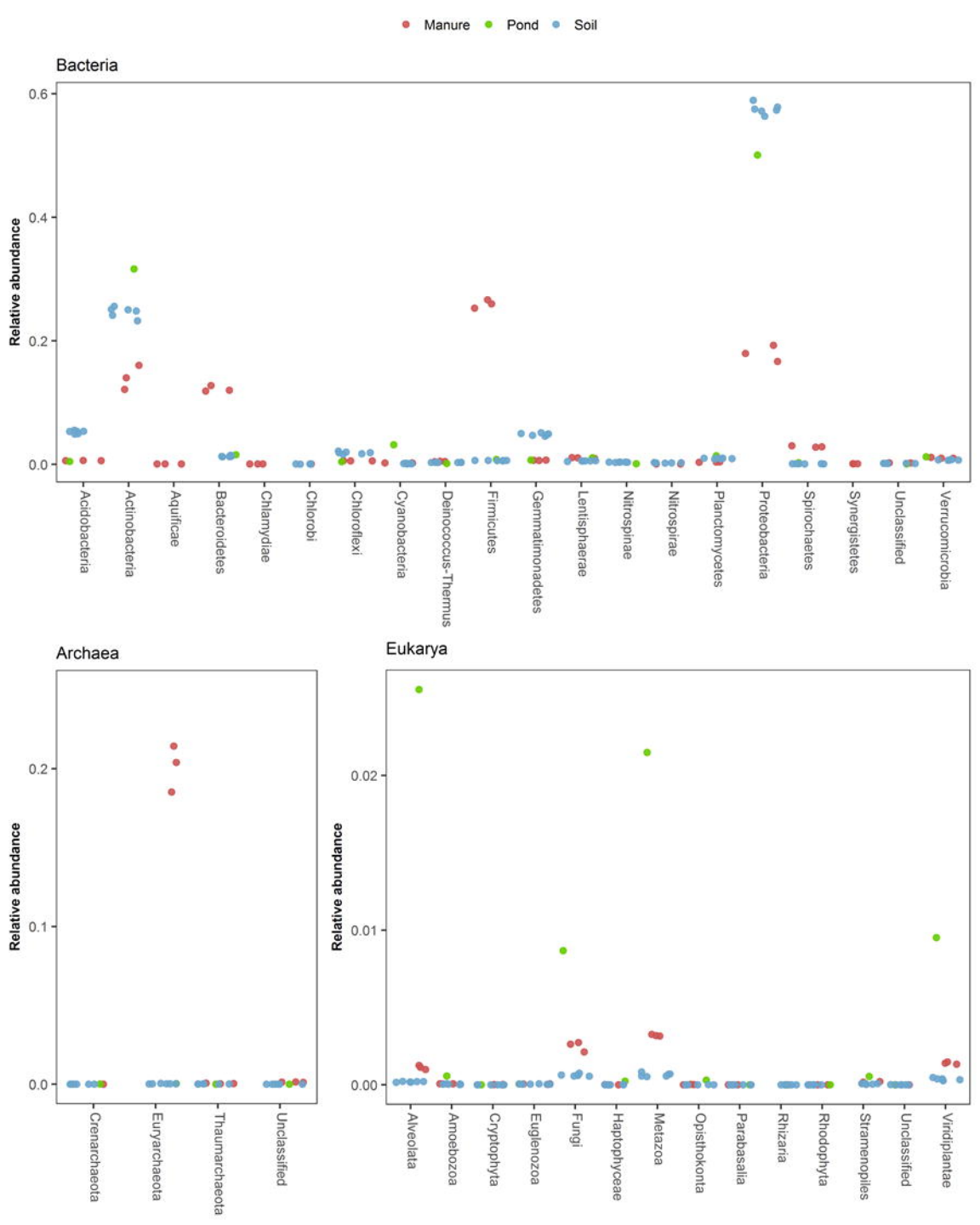
638 **Table 1:** Abundances of selected OTU from each Domain, as determined by quantitative PCR.

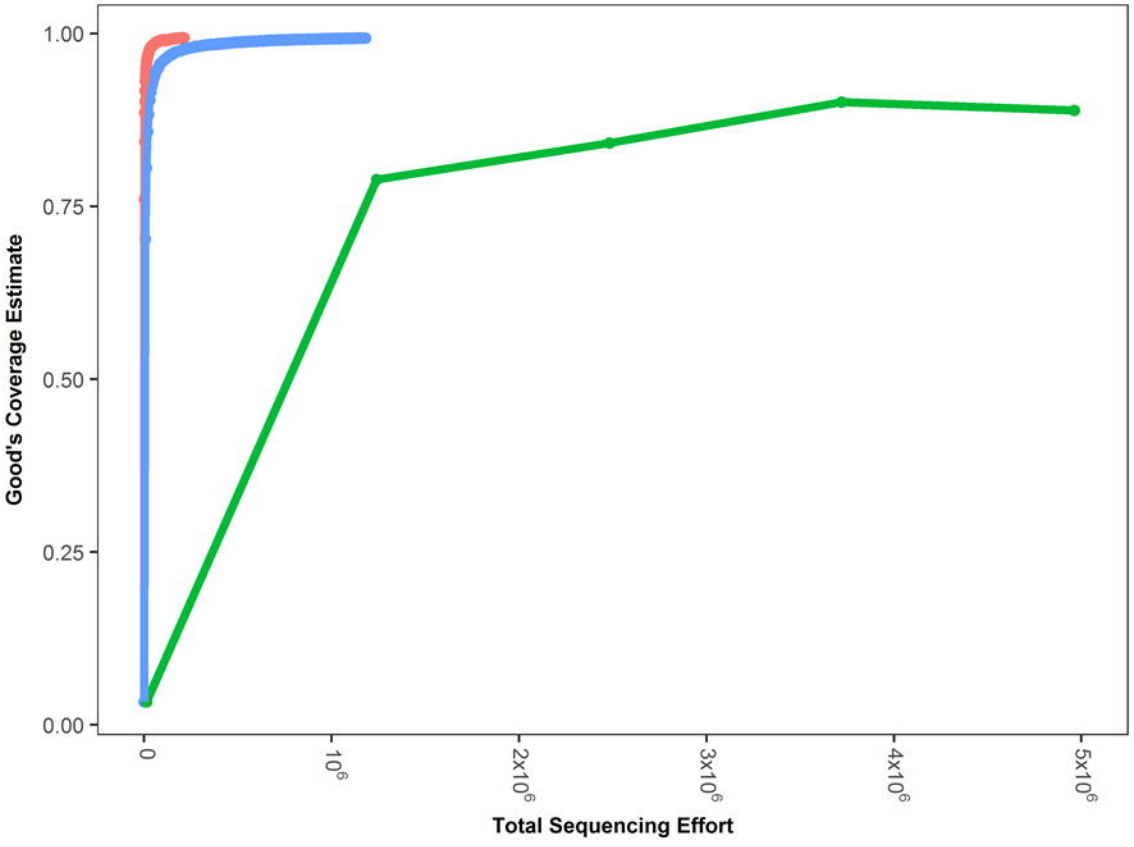
OTU	Domain	cpnDB nearest neighbor	OTU Length (bp)	Sequence identity (%)[a]	Treatment (mg kg^{-1})	soil extract (copies/g soil)	Post-hybridization sample (copies/µl)
XP002901426 DN2_c0_g1_i1	Eurkarya (type I)[b]	*Phytophthora infestans*	539	100	0	ND[c]	1242
					10	ND	3942
WP036300323 DN4_c3_g1_i2	Bacteria (type I)	*Microbacterium* sp. C448	1,066	99	0	6750	1417
					10	38571[d]	8170[d]
KUL05486 DN0_c0_g1_i1	Archaea (type II)	*Methanoculleus marisnigri*	1,029	92	0	495	ND
					10	527	3360

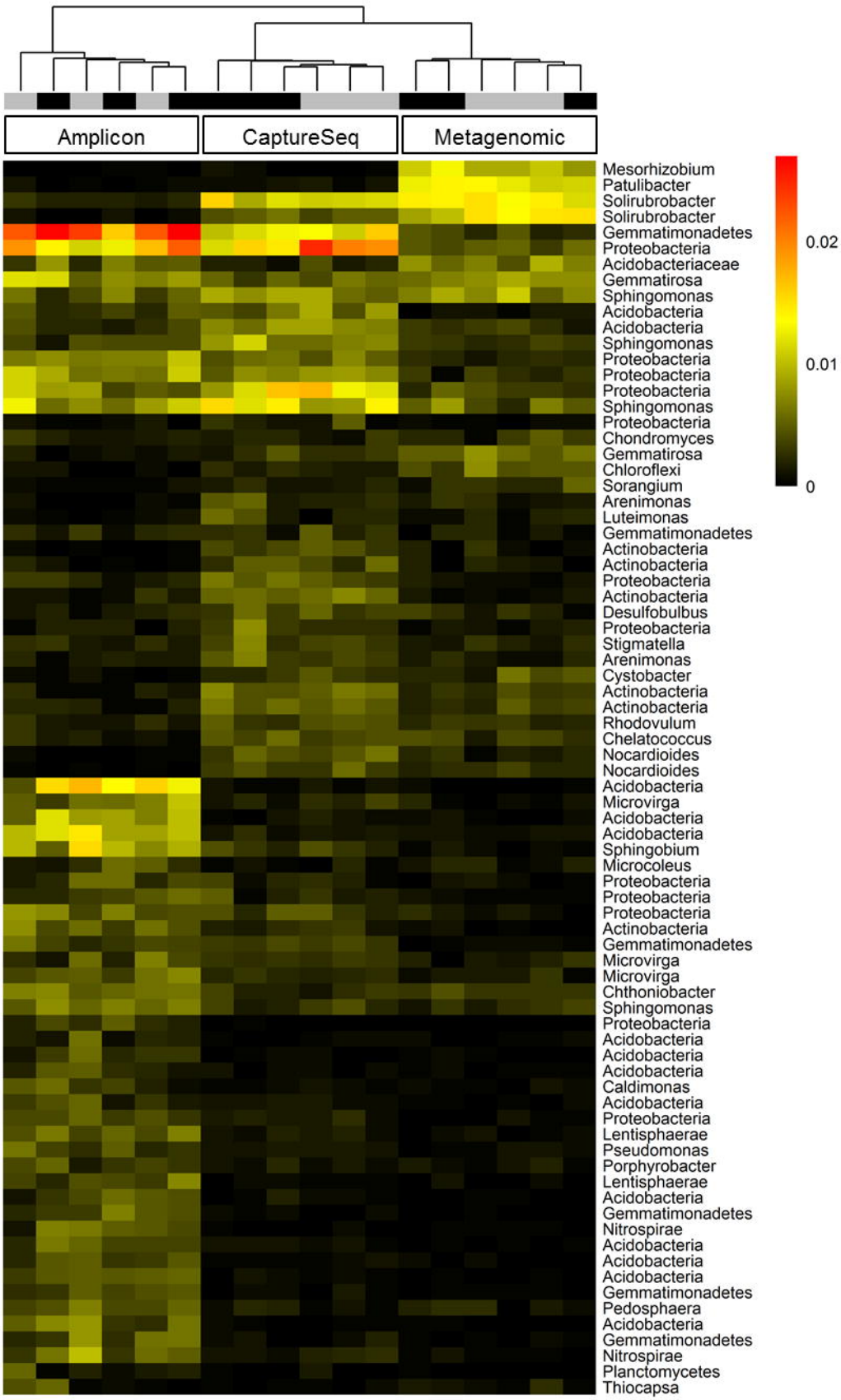
[a]Percent identity to reference sequence in cpnDB.
[b]Type I refers to the ~60 kDa mitochondrial and chloroplast proteins found in Bacteria, Eukarya, and certain Archaea. Type II refers to TCP1, the cytoplasmic orthologue of the group I chaperonins found in Archaea.
[c]ND, not detected. The theoretical detection limit was 3570 copies/g soil, as discussed in the text.
[d]Statistically significant difference (p<0.01) between 0 mg kg^{-1} and 10 mg kg^{-1} groups, using a Mann-Whitney rank sum test

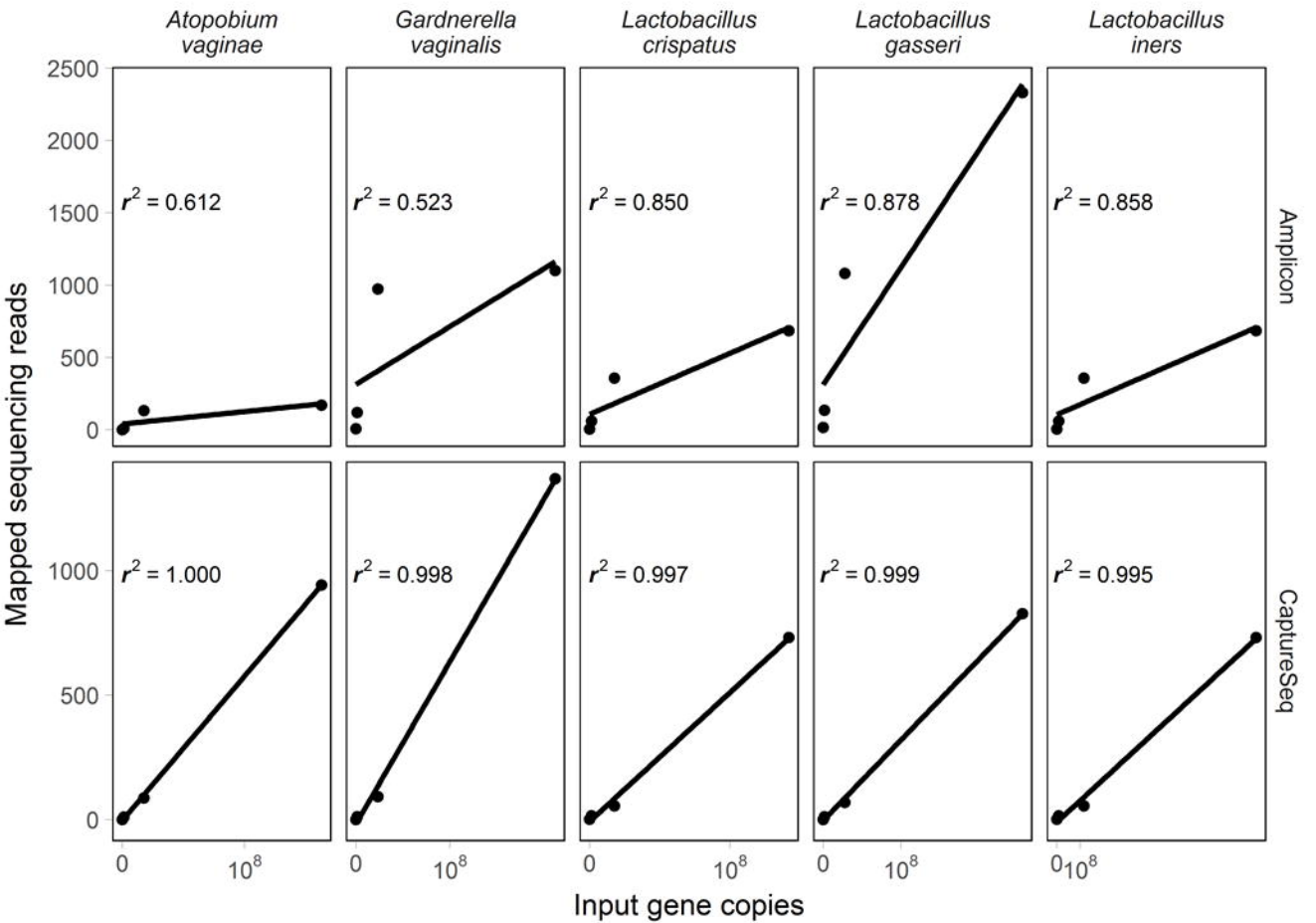
639


Atopobium vaginae
Bifidobacterium bifidum
Bifidobacterium infantis
Gardnerella vaginalis
Lactobacillus sp. N27
Lactobacillus iners
Lactobacillus crispatus
Lactobacillus jensenii
Lactobacillus vaginalis
Lactobacillus gasseri
Lactobacillus salivarius
Lactobacillus sp. L6
Lactobacillus sp. N19
Lactobacillus johnsonii
Lactobacillus sp. N20
Peptostreptococcus anaerobius
Prevotella bivia
Streptococcus gallolyticus
Streptococcus lutetiensis
Lactobacillus plantarum
3.5
3.0
2.5
2.0
1.5
1.0
0.5
0
High
Medium
Low
High
Medium
Low
Amplicon
CaptureSeq


Manure
Pond
Soil
Bacteria
Relative abundance
0.6
0.4
0.2
0.0
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Chlamydiae
Chlorobi
Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Gemmatimonadetes
Lentisphaerae
Nitrospinae
Nitrospirae
Planctomycetes
Proteobacteria
Spirochaetes
Synergistetes
Unclassified
Verrucomicrobia
Archaea
Relative abundance
0.2
0.1
0.0
Crenarchaeota
Euryarchaeota
Thaumarchaeota
Unclassified
Eukarya
Relative abundance
0.02
0.01
0.00
Alveolata
Amoebozoa
Cryptophyta
Euglenozoa
Fungi
Haptophyceae
Metazoa
Opisthokonta
Parabasalia
Rhizaria
Rhodophyta
Stramenopiles
Unclassified
Viridiplantae


1.00
0.75
0.50
0.25
0.00
Good's Coverage Estimate
0
10^6
2x10^6
3x10^6
4x10^6
5x10^6
Total Sequencing Effort


Amplicon
CaptureSeq
Metagenomic
0.02
0.01
0
Mesorhizobium
Patulibacter
Solirubrobacter
Solirubrobacter
Gemmatimonadetes
Proteobacteria
Acidobacteriaceae
Gemmatirosa
Sphingomonas
Acidobacteria
Acidobacteria
Sphingomonas
Proteobacteria
Proteobacteria
Proteobacteria
Sphingomonas
Proteobacteria
Chondromyces
Gemmatirosa
Chloroflexi
Sorangium
Arenimonas
Luteimonas
Gemmatimonadetes
Actinobacteria
Actinobacteria
Proteobacteria
Actinobacteria
Desulfobulbus
Proteobacteria
Stigmatella
Arenimonas
Cystobacter
Actinobacteria
Actinobacteria
Rhodovulum
Chelatococcus
Nocardioides
Nocardioides
Acidobacteria
Microvirga
Acidobacteria
Acidobacteria
Sphingobium
Microcoleus
Proteobacteria
Proteobacteria
Proteobacteria
Actinobacteria
Gemmatimonadetes
Microvirga
Microvirga
Chthoniobacter
Sphingomonas
Proteobacteria
Acidobacteria
Acidobacteria
Acidobacteria
Caldimonas
Acidobacteria
Proteobacteria
Lentisphaerae
Pseudomonas
Porphyrobacter
Lentisphaerae
Acidobacteria
Gemmatimonadetes
Nitrospirae
Acidobacteria
Acidobacteria
Acidobacteria
Gemmatimonadetes
Pedosphaera
Acidobacteria
Gemmatimonadetes
Nitrospirae
Planctomycetes
Thiocapsa


Atopobium vaginae
Gardnerella vaginalis
Lactobacillus crispatus
Lactobacillus gasseri
Lactobacillus iners
Amplicon
CaptureSeq
Mapped sequencing reads
Input gene copies
r² = 0.612
r² = 0.523
r² = 0.850
r² = 0.878
r² = 0.858
r² = 1.000
r² = 0.998
r² = 0.997
r² = 0.999
r² = 0.995
2500
2000
1500
1000
500
0
10⁸