bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Measuring both microbial load and diversity with a single amplicon sequencing library

Derek S. Lundberg[1*], Pratchaya Pramoj Na Ayutthaya[1], Annett Strauß[3], Gautam Shirsekar[1], Wen-Sui Lo[2], Thomas Lahaye[3], Detlef Weigel[1*]

[1]Department of Molecular Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany
[2]Department of Evolutionary Biology, Max Planck Institute for Developmental Biology, 72076 Tübingen, Germany
[3]ZMBP-General Genetics, University of Tübingen, Auf der Morgenstelle 32, 72076 Tübingen, Germany
[*]Corresponding authors weigel@tue.mpg.de (D.W.) and derek.lundberg@gmail.com (D.L.)

Abstract

The microbial population size, or load, in a host is a fundamental metric of colonization by commensals or infection by pathogens. Sequence-based measurement of DNA amount contributed by host and microbes in a single sample provides a simple way of measuring microbial load, and at the same time supports estimation of microbiome diversity. Unfortunately, it is also very costly, especially when host DNA greatly outweighs microbial DNA. We introduce a robust two-step PCR strategy to quantify microbial density of the host-associated microbiome and at the same time describe its composition in a single, cost-effective amplicon library. We demonstrate the accuracy and flexibility of this method across multiple amplicons and hosts, and further present a simple technique that can be used, prior to sequencing, to optimize the host representation in a batch of libraries without a loss of information.

Introduction

Knowing the relative abundance of individual taxa reveals important information about any ecological community, including microbial communities, and an expedient means of learning their composition is to count the copies of 16S or 18S rRNA genes (hereafter rDNA), of the internal transcribed spacer (ITS) of rRNA arrays, or of other amplicons that distinguish microbial species in a sample. However, these common amplicon counting-by-sequencing methods do not provide information on the density or load of the microbes, which is typically defined by other methods such as time-consuming serial dilution plating. Critically, such microbial sequence counts lack a denominator that accounts for the amount of the habitat

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

sampled. Typically, such samples are therefore sequenced to a similar depth and scaled to a common value, for example by subsampling each sample to the same number of sequence reads, or converting all abundances to percentages. A sparsely-colonized and a densely-colonized sample are thus processed in an indistinguishable manner. Another limitation of such compositional data is that the more abundant a microbe is, the more it reduces the relative abundance of all other microbes, because the sum of all microbes is constrained. Most study systems are open systems, and the total number of microbial cells can vary over many orders of magnitude due to nutritional contents, infection status, or other factors. Without knowing to what extent microbial abundances are actually changing in absolute terms, it can be very misleading to make inferences from compositional data alone [1–4]. Unsurprisingly, experimental determination of the microbial density in the environment, for example by relating microbial abundance to sample volume, mass, or surface area, has led to important insights in microbiome research that otherwise would have been missed with relative abundance data [3,5–11].

For many host-associated microbial samples, in particular those from plants [3], nematodes [12], insects [13,14], and other organisms in which it is difficult or impossible to physically separate microbes from host tissues, both host and microbial DNA are co-extracted. For such samples, assuming thorough DNA extraction, the amount of host DNA is directly proportional to the amount of host tissue sampled [15,16], and therefore the ratio of microbial DNA to host DNA is a measure of the microbial load of the sample [3,6,17–19]. Whole metagenome sequencing (WMS), which indiscriminately and quantitatively samples all host and microbial DNA, is conceptually the simplest method to simultaneously describe the microbial community and measure the microbial load. In practice, however, host DNA is often greatly overrepresented in WMS data, and wasteful sequencing of the host's DNA is often necessary to achieve a useful depth of microbial reads [3,17]. For example, WMS sequencing of a leaf extract from wild *Arabidopsis thaliana* typically yields > 95% plant DNA and < 5% microbial DNA. This problem becomes intractable as sample number or host genome size increases. Furthermore, many WMS reads remain unclassifiable in complex samples [3,17].

More commonly, researchers approach host-associated microbial samples with amplicon sequencing, which allows the use of blocking techniques or specially-designed primers to avoid PCR amplification of host DNA [20–22]. They then interpret the amplicon data with orthogonal methods to quantify the microbial load. For example, we previously proposed shallow WMS as a way of measuring microbial load for leaf-associated microbial communities, followed by using this information to scale 16S rDNA amplicon data prepared from the same samples [3]. Alternatively, quantitative PCR (qPCR) of both a microbial and host gene can be used to calculate load [17,23,24], or qPCR of a host gene can be used to calibrate the amount of a synthetic rDNA template that can then be spiked into the sample prior to amplicon library preparation [25]. Sample volume or mass can also be used to calibrate the amount of such a synthetic spike-in [10,26]. Bacteria can be enumerated in samples of known volume or mass by counting colony

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

forming units (CFU), by flow cytometry, or by qPCR or digital PCR of a specific microbial gene [4–7,11,27,28]. While such approaches address the issue of absolute microbial load, combining data from any two methods requires more work, uses more of the sample, and introduces more experimental noise and other caveats. For example, CFU counting only can quantify the culturable fraction of the community, quantitative spike-ins and uniplex qPCR are especially reliant on accurate pipetting, and qPCR does not allow identification or exclusion of off-target amplicons.

Past attempts to implement a single amplicon sequencing protocol for simultaneous analysis of bacterial 16S rDNA and eukaryotic host DNA have met significant challenges. Lebeis and colleagues [19] used a single pair of universal SSU rDNA primers to co-amplify both the genomic 18S rDNA from *Arabidopsis thaliana* and the 16S rDNA from a bacterial community. However, because plant 18S rDNA is present in hundreds or thousands of copies per genome with exact copy number varying between plant accessions [29], it was a poor internal standard. Indeed, to overcome the overabundant plant signal, the authors needed to pre-amplify the bacterial template with a bacteria-specific 16S rDNA oligonucleotide primer, meanwhile blocking amplification of plastid 16S rDNA sequences with peptide nucleic acid (PNA) clamps. Humpfrey and Whiteman [6] attempted to use plant chloroplast rDNA as a proxy for the amount of host. They used bacteria-specific 16S rDNA primers to amplify microbial communities from leaves, employing PNA clamps [20] to reduce the otherwise excessive amplification of SSU rDNA from plant organelles. They showed that the unblocked chloroplast rDNA roughly correlates with the amount of host, and used it to estimate microbial load. The correlation was, however, noisy, likely because the fraction of chloroplast that is blocked by PNAs depends on the total DNA concentration, which varies between samples, and furthermore because the number of chloroplasts in plant cells is variable.

Here, we introduce host-associated microbe PCR (hamPCR), a robust and accurate protocol to co-amplify a low-copy host gene and one or more microbial regions, such as 16S rDNA, in the same reaction. We accomplish this with a two-step PCR protocol similar to that described in established amplicon protocols [20,30–34]. In these two-step protocols, gene-specific primers bind to 'raw' templates in the first step and 'tag' the templates with universal overhangs. In the second step, primers with complementarity to the universal overhangs add barcodes and sequencing adapters. Thus, the same set of barcoded primers can be used regardless of the amplicon(s) tagged in the first step, a major advantage in flexibility.

In hamPCR, we convert the first 'tagging' PCR step into a multiplex reaction that combines two or more template-specific primer pairs in the same reaction tube. To avoid propagating any biases due to differential primer annealing or primer availability across multiple cycles, the tagging step runs for only two PCR cycles, which is the minimum sufficient to make a single copy of each template molecule with universal overhangs at both 5' and 3' ends; a similar strategy of short initial cycling is employed in sequencing protocols that add Unique Molecular Identifiers (UMIs) [26,35–37] and in other two-step multiplex protocols [33,38].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

We then remove the excess gene-specific primers with a clean-up reaction and use the 'tagged' templates in the second step of the PCR reaction, performed with a single barcoded primer pair for 20 or more cycles. Because all templates share the same universal overhangs, primer annealing will not bias their amplification. Such PCR co-amplification of diverse fragments is also commonplace in many RNA-seq and shotgun metagenomics protocols [39,40]. Similar to our work, Carlson and colleagues [33] used a two-step PCR with a multiplexed tagging step to target and quantify both variable and joining segments at human T and B cell receptor loci. They showed through high-throughput sequencing that the method remains quantitative for between five and seven cycles of the first step of PCR. Similarly, Wen and Zhang [38] used three tagging cycles of multiplexed PCR followed by 28-32 cycles of exponential PCR to examine the genetic purity of maize seeds, although they did not sequence their amplicons. Using such methods has enormous potential in microbiome research, as we demonstrate.

We examine the technical reproducibility and limitations of hamPCR and validate our results by independent microbial load quantification methods, using both synthetic and wild samples from a variety of hosts. Because hamPCR relies on a user-defined single- or low-copy host gene, it can work for large genomes, as we demonstrate by quantitative endophytic microbial profiling in *Capsicum annuum* (pepper, ~ 3.5 Gb genome)[41], and *Triticum aestivum* (wheat, ~16 Gb genome)[42]. To further increase flexibility of the method, we designed our host and microbe amplicons to have slightly different lengths, such that they can be resolved on an agarose gel. We show that after pooling finished sequencing libraries, an aliquot of the original pool can be re-barcoded with an unused barcode to form a reference sample that preserves the host-to-microbe ratio in the original pool. The host and microbe fractions in the remainder of the pool can then be separately purified and re-mixed at a favorable ratio for sequencing (for example with host DNA representing an affordable 5-10%). The reference sample is spiked into the re-mixed library before sequencing, and it guides bioinformatic scaling of host sequences in the re-mixed library back to original levels. Thus, in contrast to shotgun sequencing, samples with initially unfavorable host-to-microbe ratios can be easily adjusted *prior to sequencing* without loss of information.

Because of the practical simplicity and flexibility of hamPCR, it has the potential to supplant traditional microbial amplicon sequencing in host-associated samples, requiring only minor adjustments for labs already using a two-step PCR protocol. For labs not using a two-step PCR protocol for amplicon-based microbiome analyses, the ability to perform hamPCR offers yet another compelling reason to make the switch.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Results

Technical reproducibility

The first 2-cycle "tagging" step of hamPCR multiplexes two or more primer pairs in the same reaction, at least one of which targets a single- or low-copy host gene (Supplementary Figure 1). The tagging primers are then cleaned with Solid Phase Reversible Immobilisation (SPRI) magnetic beads [43] (Supplementary Figure 2). Next, an exponential PCR is performed using universal barcoded primers (Figure 1a, Supplementary Figure 1). As a host amplicon in *A. thaliana* samples, we targeted a fragment of the *GIGANTEA* (*GI*) gene, which is well conserved and present as a single copy in *A. thaliana* and many other plant species [44]. As microbial amplicons, we initially targeted widely-used regions of 16S rDNA.

To assess the technical reproducibility of the protocol, we made a titration panel of artificial samples combining varying amounts of pure *A. thaliana* plant DNA with pure bacterial DNA that reflects a simple synthetic community (Methods). These represented a realistic range of bacterial concentrations as previously observed from shotgun metagenomics of wild leaves, ranging from about 0.25% to 24% bacterial DNA [3]. We applied hamPCR to the panel, pairing one of three commonly-used 16S rDNA amplicons for the V4, V3V4, and V5V6V7 variable regions with either a 502 bp or 466 bp *GI* amplicon (Methods, Supplementary Table 1), such that the host and microbial amplicons differed by approximately 80 bp in length and were resolvable by gel electrophoresis. In all pairings, the *GI* band intensity increased as the 16S rDNA band intensity decreased (Figure 1b, Supplementary Figure 3). Encouragingly, we noticed the *GI* and 16S rDNA bands were consistently of similar intensity at around 0.5% to 1% bacterial DNA, which is approximately the concentration at which the templates should have equal copy number, assuming a 135 Mb (1C-value) plant genome with one *GI* copy and a 4 Mb bacterial genome with three or more 16S rDNA copies.

Focusing on the V4 16S rDNA primer set, 515F - 799R, paired with the 502 bp GI amplicon, we amplified the entire titration panel in four independently-mixed technical replicates. In addition to use of the chloroplast-avoiding 799R primer [22], plant organelle-blocking PNAs [20] further prevented unwanted 16S rDNA signal from organelles in the pure plant sample (Figure 1b). We pooled the replicates and sequenced them as part of a paired-end HiSeq 3000 lane. Because the 150 bp forward and reverse reads were not long enough to assemble into full amplicons, we analyzed only the forward reads (Methods), processing the sequences into Amplicon Sequence Variants (ASVs) and making a count table of individual ASVs using *Usearch* [45].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

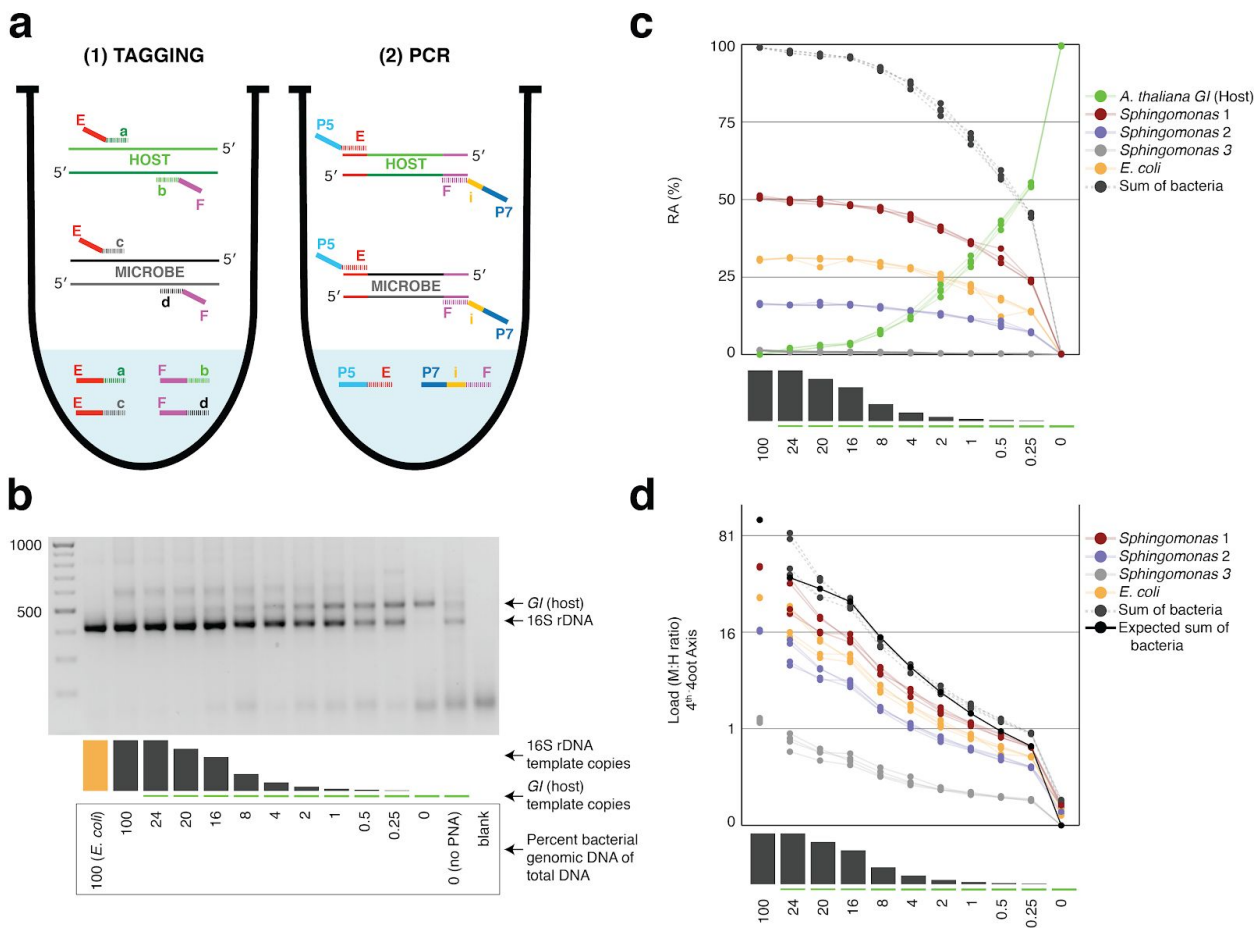

Figure 1. Synthetic samples demonstrate technical reproducibility | a, Schematic showing the two steps of hamPCR. The tagging reaction (left) shows two primer pairs: one for the host (E-a and F-b) and one for microbes (E-c and F-d). Each primer pair adds the same universal overhangs E and F. The PCR reaction (right) shows a single primer pair (P7-E and P5-i-F) that can amplify all tagged products. **b**, Representative 2% agarose gel of hamPCR products from the synthetic titration panel, showing a V4 16S rDNA amplicon at ~420 bp and an *A. thaliana* GI amplicon at 502 bp. The barplot underneath shows the predicted number of original *GI* and 16S rDNA template copies. Numbers boxed below the barplot indicate the percent bacterial genomic DNA of total DNA. **c**, Relative abundance of the host and microbial ASVs in the synthetic titration panel, as determined by amplicon counting. Pure *E. coli*, pure *A. thaliana* without PNAs, and blanks were excluded. **d**, Data in (c) converted to microbial load by dividing by host abundance, with a fourth-root transformed y-axis to better visualize lower abundances.

After identifying the ASVs corresponding to host *GI* and the bacteria in the synthetic community, we plotted the relative abundance of *A. thaliana GI,* the three *Sphingomonas* ASVs, and the single *E. coli* ASV across the samples of the titration panel (Figure 1c). There was high consistency between the four replicates, more than what was visually apparent in the gel (Figure 1c, Supplementary Figure 3). We next divided the ASV abundances in each sample by the abundance of the host ASV in that sample to give the quantity of microbes per unit of host, a measure of the microbial load. Plotting the data with a fourth-root transformed Y axis, used for better visualization of low bacterial loads, revealed consistent and accurate quantification of

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

absolute microbial abundance from 0 up to about 16% total bacterial DNA (Figure 1d). Through this range, the actual sequence counts for total bacteria matched theoretical expectations based on the volumes pipetted to make the titration (solid black line, Figure 1d). At 16% bacterial DNA, bacteria contributed more than 96% of sequences, and the microbe-to-host template ratio was near 25. At higher microbial loads the trend was still apparent, and the decrease in precision was likely exacerbated by the effects of small numbers; when the host ASV abundance is used as a denominator and the abundance approaches 0, load approaches infinity and all sources of error have a greater influence on the quotient. Through simulations, we decided conservatively that loads calculated with a host ASV abundance below 3% should only be classified as "highly colonized", and should not be used for quantitative measurements (Supplementary Figure 4). In our case, only a minority of highly infected plants [3] reached bacterial abundances above this highly-quantitative range.

hamPCR does not distort the detected composition of the microbial community

We amplified the same wild *A. thaliana* phyllosphere template DNA, either with four technical replicates using V4 16S rDNA primers alone, or alternatively with four technical replicates using hamPCR. After sequencing and deriving ASVs, we first compared ASV abundances within identically-prepared replicates of the pure 16S rDNA protocol to demonstrate best-case technical reproducibility of this established technique. As expected, this resulted in a nearly perfect correlation, with a coefficient of determination R^2 of 0.99 and abundance distributions as indistinguishable by a Kolmogorov–Smirnov test (Figure 2b). Next, we removed from the hamPCR data the ASV corresponding to *A. thaliana GI*, and rescaled the remaining microbial ASVs to 100% to give relative abundance data. We then compared microbial ASVs from the four pure 16S rDNA replicates to those from the four rescaled hamPCR replicates. In this comparison as well, R^2 was 0.99 and the distributions were essentially identical (Figure 2c). Thus, the inclusion of a host amplicon in the reaction did not introduce taxonomic biases.

Sensitivity to number of tagging cycles and template concentration

Two tagging cycles minimize amplification biases that might otherwise have compounding effects due to differential primer efficiencies for the host and microbial templates. However, for templates at low concentrations, inefficiencies due to SPRI cleanup could represent a bottleneck in amplification. Additionally, some techniques that prevent off-target organelle amplification [21,46] may benefit from additional tagging cycles. To investigate the sensitivity of the results to additional tagging cycles, we applied hamPCR for 2 through 10 tagging cycles, both on the wild *A. thaliana* phyllosphere DNA described above and on a synthetic plasmid-borne template that contains bacterial rDNA and a partial *A. thaliana GI* gene

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

template in a 1:1 ratio. Surprisingly, for the primers we used, there was no apparent influence of additional tagging cycles, as 7-10 tagging cycles yielded the same distribution of host and 16S rDNA ASV abundances as 2 cycles (Kolmogorov-Smirnov test, $p > 0.47$). This was true for hamPCR and for 16S rDNA primers alone (Figure 2d, 2e). This ideal result may not be the case for all primer pairs and should be tested experimentally, but it is consistent with data that either 5 or 7 tagging cycles gave comparable results for quantifying the human immune receptor repertoire [33], and with the fact that properly-designed multiplex reactions can be used in qPCR carried out with many cycles [47]. We noticed that application of hamPCR to the 1:1 synthetic template yielded an average of 56.5% host *GI* and 43.5% bacteria, invariant with tagging cycle number (Supplementary Figures 5 and 6). This slight and consistent bias in favor of *GI* may be a result of slight differences in tagging primer efficiency or primer concentration, and should be fine-tunable by altering primer concentration [33] (Supplementary Figure 7).

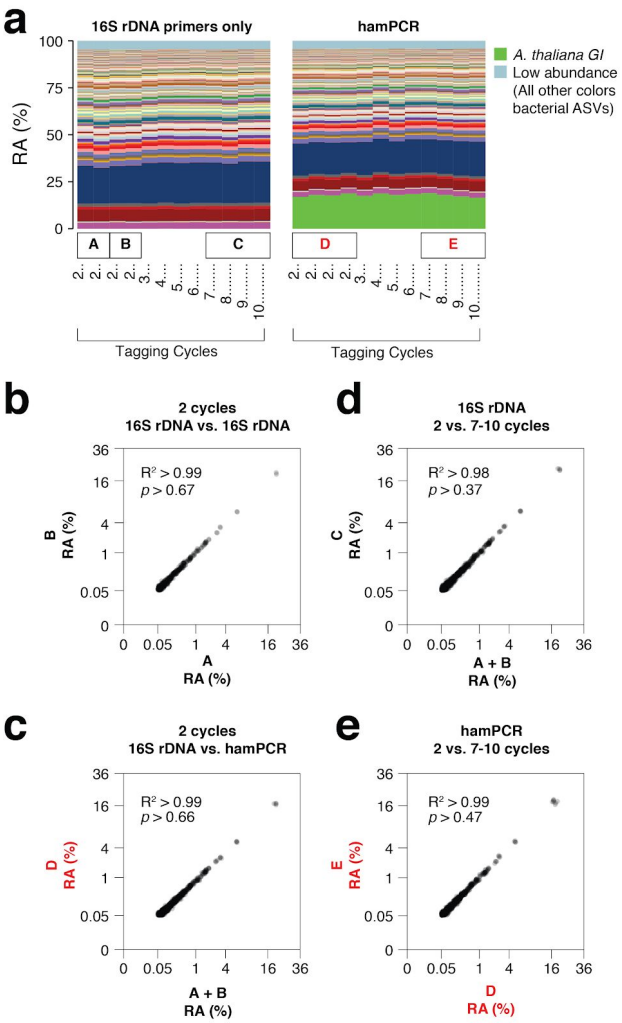

Figure 2. hamPCR is robust and does not distort a complex microbial community | a, DNA extracted from wild *A. thaliana* phyllospheres was used as a template for both V4 16S rDNA PCR (left, 515F and 799R) and hamPCR (right, V4 16S rDNA and *GI* 502 bp primers). Four replicates were produced with 2 cycles of the tagging reaction and 30 cycles of PCR, and additional replicates with 3 to 10 tagging cycles paired with 29 to 22 PCR cycles (for a constant total of 32 cycles). The stacked columns show the relative abundances of major ASVs. Boxed upper case letters demarcate groups of samples compared below. **b**, Correlation of fourth-root transformed ASV abundances for the 16S rDNA samples above panel (a) box [A] to the 16S rDNA samples above box [B]. Only ASVs with a minimum relative abundance of 0.05% were compared. R^2, coefficient of determination. *p*-value from Kolmogorov-Smirnov test. **c**, Same as (b), but for the four 16S rDNA samples above box [A] and [B] compared to the four hamPCR samples above box [D]. For hamPCR, the *A. thaliana GI* ASV was removed and the bacterial ASVs were rescaled to 100% prior to the comparison. **d**, Same as (b) and (c), but for the four 16S rDNA samples above box [A] and [B] compared to the 16S rDNA samples above box [C]. **e**, Same as (b), (c), and (d), **b**ut for the four hamPCR samples above box [D] compared to the four hamPCR samples above box [E].

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

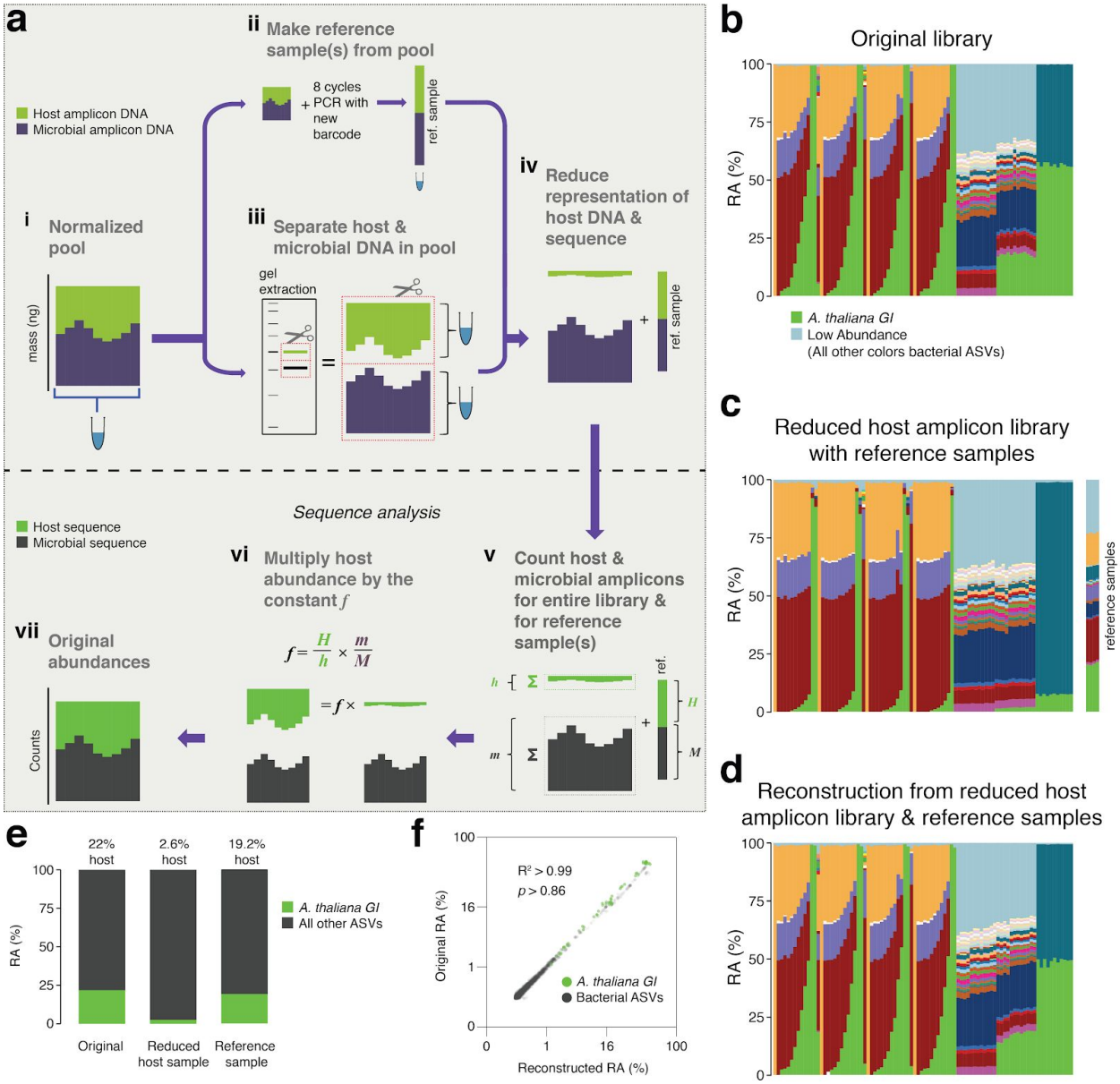

Figure 3. After remixing hamPCR amplicons for efficient sequencing, original abundances can be reconstructed | a, Scheme of remixing process. **i**: Products of individual PCRs are pooled at equimolar ratios into a single tube. **ii**: An aliquot of DNA from the pool in (**i**) is re-amplified with 8 cycles of PCR to replace all barcodes in the pool with a new barcode, creating a reference sample. **iii**: An aliquot of the pool from (**i**) is physically separated into host and microbial fractions via agarose gel electrophoresis. **iv**: The host and microbial fractions and the reference sample are pooled in the ratio desired for sequencing. **v**: All sequences are quality filtered, demultiplexed, and taxonomically classified using the same parameters. **vi**: Host and microbial amplicon counts are summed from the samples comprising the pooled library (*h* and *m* respectively), and from the reference sample (*H and M*). **vi**: *H, h, M, and m* are used to calculate the scaling constant *f* for the dataset. All host sequence counts are multiplied by *f* to reconstruct the original microbe-to-host ratios. **vii**: Reconstructed original abundances. **b**, Relative abundance (RA) of actual sequence counts from our original HiSeq 3000 run. **c**, Relative abundance of actual sequence counts from our adjusted library showing reduced host and 4 reference samples. **d**, The data from (c) after reconstructing original host abundance using the reference samples. **e**, The total fraction of host vs. other ASVs in the original library, reduced host library, and reconstruction. **f**, Relative abundances in the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

original and reconstructed library for all ASVs with a 0.05% minimum abundance, shown on 4th-root transformed axes. R^2, coefficient of determination. *p*-value from Kolmogorov–Smirnov test.

As a further exploration of the robustness of the protocol, we applied hamPCR to a range of total *A. thaliana* leaf template concentrations of between 5 and 500 ng total DNA per reaction, covering a typical template range of 5 to 100 ng. Through the typical range, there was no difference in microbe or host ASV abundances. At 200 ng or above, the host amplicon seemed to be slightly favored, possibly because the 16S rDNA primers started to become limiting at these concentrations (Supplementary Figure 8).

Pre-sequencing adjustment of host-to-microbe ratio

We realized that the size difference between host and microbe bands in hamPCR affords not only independent visualization of both amplicons on a single gel, but also allows convenient and easy adjustment of the host and microbial signals in the pooled library prior to sequencing, in order to improve cost effectiveness. We developed a strategy by which the final hamPCR amplicons are pooled and one aliquot of the pool is rebarcoded to form a reference sample that preserves the original host- to- microbe ratio. The remainder of the pool is run on a gel and the host and microbial bands are separately purified, quantified, and remixed (in our case, to reduce host and gain more microbial resolution). The rebarcoded reference sample, which was not remixed and thereby preserves the original ratio, can be sequenced separately or spiked into the remixed library prior to sequencing. Following sequencing, the reference sample provides the key to the correct host and microbe proportions, allowing simple scaling of the entire library back to original levels (Figure 3a, Supplementary Figure 9).

We made 4 replicate reference samples for our HiSeq 3000 run, which included much of the data from Figure 1 and Figure 2, and then separately purified the host and microbial fractions of the library (Supplementary Figures 9 and 10). Based on estimated amplicon molarities of the host and microbial fractions, we remixed them targeting 5% host DNA, added the reference samples, and sequenced the final mix as part of a new HiSeq 3000 lane. A stacked-column plot of relative abundances for all samples on the original run clearly showed the host *A. thaliana GI* ASV highly abundant in some samples, on average responsible for about 22% of total sequences in the run (Figure 3b, 3e). The remixed reduced host library had nearly 10-fold less host *GI* ASV, 2.6%, slightly lower than our target of 5% (Figure 3c, 3e). The reference samples averaged 19.2% of host *GI* ASV, very close to the 22% host fraction in the original library. After using the reference samples to reconstruct the original host abundance in the remixed dataset, we recreated the shape of the stacked column plot from the original library (compare Figure 3d to 3b). When the fourth-root abundances for ASVs above a 0.05% threshold were compared between the original and reconstructed libraries, the R^2 coefficient of determination was 0.99, with no significant difference between the distributions (Kolmogorov-Smirnov test, $p > 0.86$).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

hamPCR with different 16S rDNA regions compared to shotgun metagenomics

We next applied hamPCR to leaf DNA from eight wild *A. thaliana* plants that we had previously analyzed by WMS, and from which we therefore had an accurate estimate of the microbial load as the number of microbial reads divided by the number of plant chromosomal reads [3]. We applied hamPCR with primer combinations targeting the host *Gl* gene and either the V3V4, V4, or V5V6V7 variable regions of the 16S rDNA. We produced three independent replicates for each primer set, which we averaged for final analysis (Figure 4, Supplementary Figure 11). Across WMS and the three hamPCR amplicon combinations, the relative abundance of bacterial families was consistent (Figure 4a-d: i), with slight deviations likely due to the different taxonomic classification pipeline used for the metagenome reads [3], as well as known biases resulting from amplification or classification of different 16S rDNA variable regions [48]. After converting both WMS and hamPCR bacterial reads to load by dividing by the plant reads in each sample, we recovered a similar pattern despite the quantification method, with decreasingly lower total loads progressing from plant S1 to plant S8, and individual bacterial family loads showing similar patterns (Figure 4a-d: ii, iii). Relative differences in load estimates when comparing the different hamPCR amplicons are likely in part due to different affinities of the 16S rDNA primer pairs for their targets in different bacterial species, and rDNA copy number variation among the microbial families [49]. To quantify the consistency of hamPCR load estimates with WGS load estimates, we plotted the loads against each other and found strong positive correlations, with the highest correlation with the hamPCR using V4 rDNA (Figure 4e-g).

It is important to note that while relative load ratios between samples were consistent across hamPCR primer sets, the total microbe-to-host ratio varied substantially, with the maximum V5V6V7 16S rDNA total load at less than 3 times host, and the maximum V4 16S rDNA total load near 16 times host. This is likely due to variation in *Gl* and 16S rDNA primer efficiencies. To make a statement about the ratio of plant cells to bacterial cells using hamPCR, it would be important to include standard samples with known bacterial load ratios, and to normalize each bacterial taxon by its average rDNA copy number.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

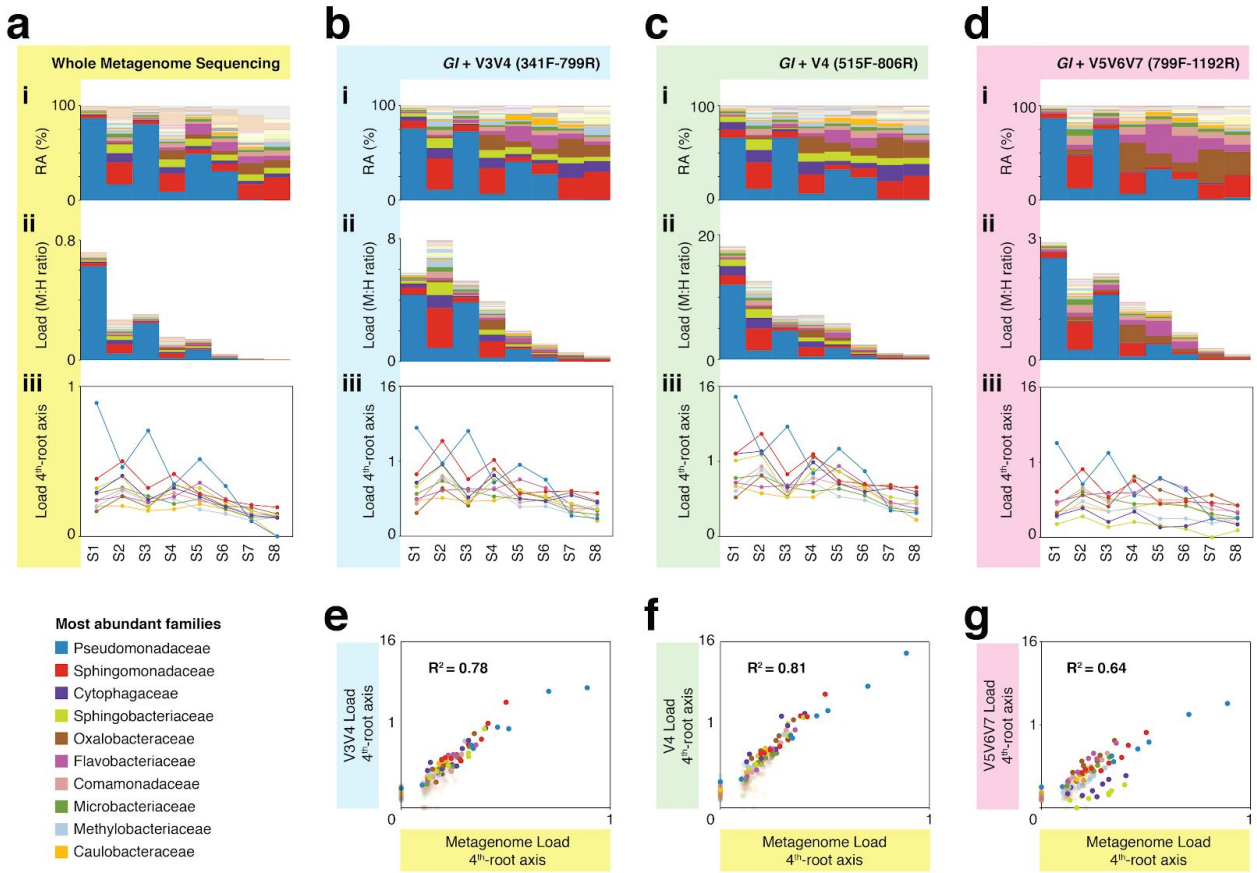

Figure 4. hamPCR with three common 16S rDNA amplicons gives consistent results that agree with shotgun metagenomics | a, i: Stacked-column plot showing the relative abundance (RA) of bacterial families in eight wild *A. thaliana* leaf samples, as determined by shotgun sequencing. The families corresponding to the first 10 colors from bottom to top are shown in reverse order on the bottom left. **ii**: Stacked-column plot showing the bacterial load of the same bacterial families (M:H ratio = microbe-to-host ratio). **iii**: The M:H bacterial load ratios for the 10 major bacterial families shown on a 4th-root transformed y-axis. Lines across the independent samples are provided as a help to visualize patterns. **b**, Similar to (a), but with abundances resulting from hamPCR targeting a 502 bp *A. thaliana GI* amplicon and a ~590 bp V3V4 16S rDNA amplicon. **c**, Similar to (b), but with the 16S rDNA primers targeting a ~420 bp V4 16S rDNA amplicon. **d**, Similar to (b), but with a 466 bp *A. thaliana GI* amplicon and a ~540 bp V5V6V7 16S rDNA amplicon. **e**, 4throot transformed abundance of each bacterial family determined by hamPCR of V3V4 16S rDNA plotted against the 4th-root transformed bacterial load from shotgun metagenomics. R^2 = Coefficient of determination. **f**, Same as (e), but for hamPCR of V4 16S rDNA. **g**, Same as (e), but for V5V6V7 16S rDNA.

Three-amplicon hamPCR for simultaneous determination of oomycete and bacterial load

More than two amplicons can be quantitatively tagged and amplified, although it requires more initial troubleshooting to find compatible primers [33,38,47,50]. We set up hamPCR to co-quantify bacteria and eukaryotic oomycetes on plants. These are both diverse microbial groups that include important *A. thaliana* pathogens and that cannot be captured by the same

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

rDNA primer set. We first tested universal ITSo primers targeting oomycete rDNA (Methods) in combination with *A. thaliana GI* primers and 16S rDNA primers targeting the bacterial V4, V3V4, or V5V6V7 regions, using as template our synthetic plasmid that includes templates for the three primer sets in equal proportion (Methods). A combination of all three amplicons seemed to work efficiently for the V4 region (Supplementary Figure 12), and with this encouraging result, we set up a simple infection experiment. As pathogens, we prepared local strain 466-1 of the obligate biotrophic oomycete *Hyaloperonospora arabidopsidis* (*Hpa*) [51] and the well-described bacterial pathogen *Pseudomonas syringae* pv. *tomato* (*Pto*) DC3000 [52]. We used two *A. thaliana* genotypes: the reference accession Col-0, which is resistant to *Hpa* 466-1 but susceptible to *Pto* DC3000, and an *enhanced disease susceptibility 1* (*eds1-1*) mutant, which has a well-studied defect in a lipase-like protein necessary for many disease resistance responses and which is susceptible to both pathogens [53].

We infected seedlings with either *Hpa* 466-1 alone, a mix of *Hpa* 466-1 and *Pto* DC3000, or a buffer control, and maintained them for 7 days under cool, humid conditions ideal for *Hpa* growth (Methods). The *eds1-1* plants inoculated with *Hpa* 466-1 became heavily infected and sporangiophores were too numerous to count. No visible bacterial disease symptoms were present on any of the plants, likely because the cool temperature decelerated bacterial growth and symptom appearance [54]. We ground pools of 4-5 seedlings in a buffer and used a small aliquot to count *Pto* DC3000 CFUs, and the remainder of the lysate for DNA isolation and hamPCR. Despite the lack of bacterial symptoms, we recovered *Pto* DC3000 CFUs from the inoculated plants.

We applied hamPCR to these samples using the ITSo/16S/*GI* primer set, but due to excessive ITSo product, we repeated library construction replacing the ITSo primers with primers for a single copy *Hpa* actin gene (Supplementary Figure 12)[23]. Intensity of the actin product correlated with visual *Hpa* symptoms (Supplementary Figure 12). Sequencing the libraries confirmed *Hpa* and *Pto* ASVs in the inoculated samples, as expected. A standard bacterial relative abundance plot, as would be obtained from pure 16S rDNA data, confirmed the presence of *Pto* DC3000 in the bacteria-infected samples, and in addition revealed that *Hpa*-infected samples had a different bacterial community than uninfected samples (Figure 5a). Importantly, it failed to detect obvious differences between microbial communities on Col-0 and *eds1-1* plants. However, after including the actin ASV from *Hpa* and converting all abundances to microbial load, a striking difference became apparent between Col-0 and *eds1-1*, with *eds1-1* supporting higher bacterial and *Hpa* abundances. This is expected from existing knowledge [53], and supported by *Pto* DC3000 CFU counts from the same plants (Figure 5b, Figure 5c). The microbial load plot also revealed that *Hpa*-challenged plants supported more bacteria than buffer-treated plants, indicating either that successful bacterial colonizers were unintentionally co-inoculated with *Hpa*, or that *Hpa* caused changes in the native flora (Figure 5b).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To confirm that the sequence abundances for all three amplicons accurately reflected the concentration of their original templates, we prepared a stepwise titration panel with real samples, mixing increasing amounts of DNA from an uninfected *eds1-1* plant (low load) into decreasing amounts of DNA from an *Hpa*-infected *eds1-1* plant (high load). Sequencing triplicate hamPCR libraries revealed a stepwise increase in ASV levels for all amplicons, consistent with the expectation based on pipetting (Figure 5d). These data, combined with the infection experiment, show that hamPCR is quantitative for at least two independent microbial amplicons in real-world samples.

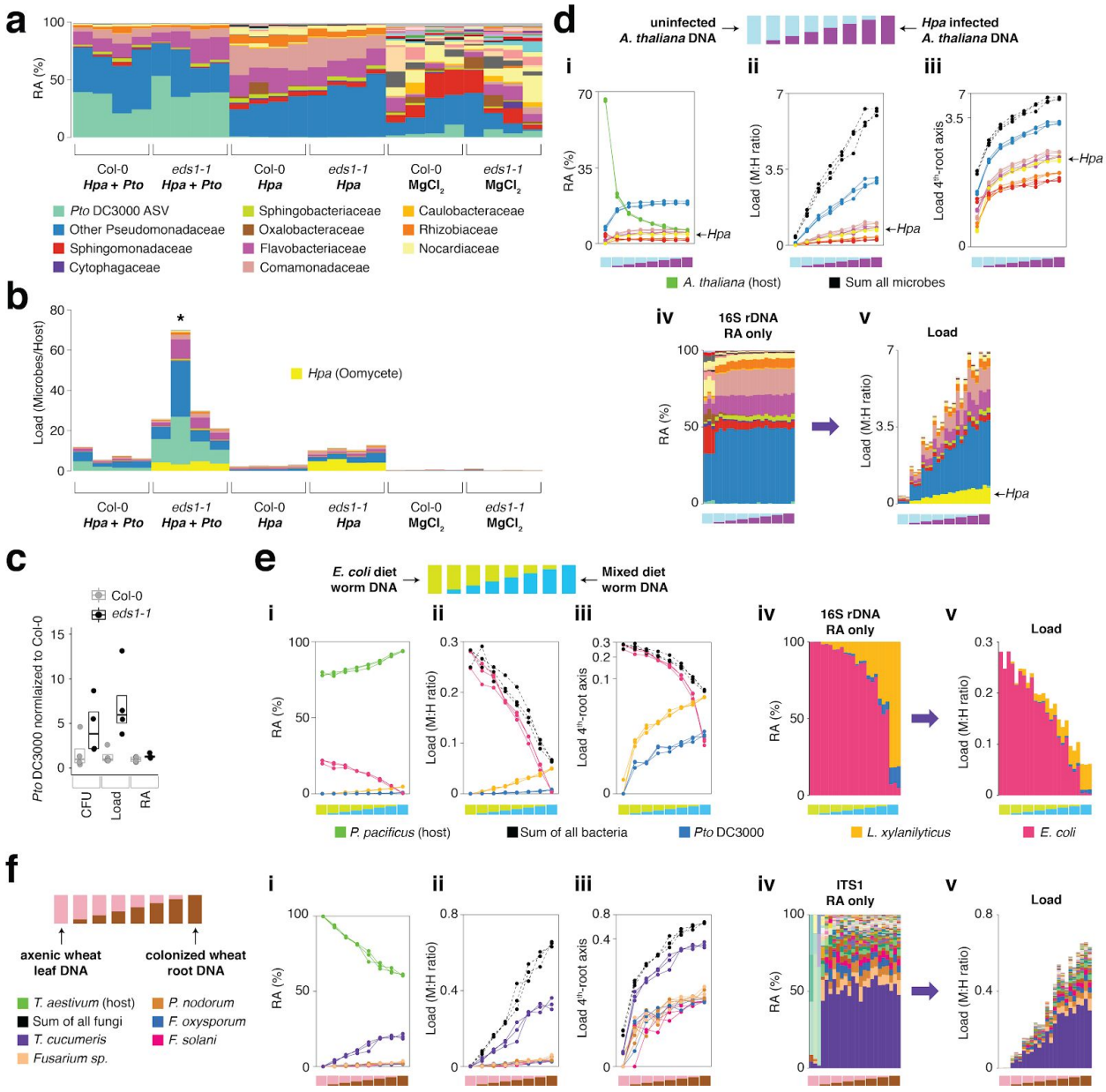

Figure 5. hamPCR can be generalized to more than two amplicons, non-plant hosts, and large host genomes | a, Relative abundance (RA) of only 16S rDNA amplicons for plants co-infected with *Hpa* and *Pto* DC3000 and their controls. The ASV corresponding to *Pto* DC3000 (light green) is shown at the bottom and separately from other Pseudomonadaceae; all other bacteria are classified to the family level, including remaining

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Pseudomonadaceae. **b**, The same data as shown in (a), but converted to microbial load and with the ASV for *Hpa* included. The asterisk indicates what is most likely an unreliable load calculation, because host ASV abundance was below 3% of the total. Same color key as in (a), with an additional color (yellow) for *Hpa* added. *Hpa* amplicon abundance was scaled by a factor of 4 in this panel for better visualization. **c**, *Pto* DC3000 bacteria were quantified in parallel on the Col-0 and *eds1-1* samples infected with *Pto* DC3000 using CFU counts, the microbial load data in (b), or the relative abundance data in (a). The median is shown as a horizontal line and box boundaries show the lower and upper quartiles. **d**, An uninfected plant sample was titrated into an *Hpa*-infected sample to make a panel of eight samples. **i**: the relative abundance of hamPCR amplicons with median abundance above 0.15%. **ii**: after using host ASV to convert amplicons to load. The cumulative load is shown in black. **iii**: the load on a 4th-root transformed y-axis, showing less-abundant families. **iv**: stacked column visualization of all ASVs for the panel as it would be seen with pure 16S rDNA data. **v**: stacked-column plot of the panel corrected for microbial load. Same color key as in (b), but with colors for *A. thaliana* and sum of microbes added. **e**, Similar to (d), but with the nematode worm *P. pacificus* as host, and V5V6V7 16S rDNA primers. Instead of bacterial families, specific ASV abundances are shown. **f**, Similar to (e), but with hexaploid wheat *T. aestivum* as host, and fungal ASV abundances from ITS1 amplicons.

Utility in diverse hosts and crops with larger genomes

To demonstrate the utility of hamPCR outside of plants, we prepared samples of the nematode worm *Pristionchus pacificus*, fed on a diet of either pure *E. coli* OP50, or alternatively a mix of *E. coli* OP50 with *Pto* DC3000 and *Lysinibacillus xylanilyticus*. The worms were washed extensively with PBS buffer to remove epidermally-attached bacteria, enriching the worms for gut-associated bacteria, and we prepared DNA from each sample. In the same manner as described in the previous section, we titrated the two DNA samples into each other to create a panel of samples representing a continuous range of colonization at biologically-relevant levels. Over three replicates, hamPCR accurately captured the changing bacterial loads of the gut microbes (Figure 5e, Supplementary Figure 13). We similarly validated the technique for fungal and bacterial microbes of *Triticum aestivum* (bread wheat) the most widely grown crop in the world and one of the most difficult to study due to a 16 Gb haploid genome size [42]. To simulate different levels of infection, we titrated DNA from axenically-grown wheat leaves into DNA from wheat roots that had been cultivated in non-sterile soil and applied hamPCR, using as a host gene RNA polymerase A1 (*PolA1*), which is present as a single copy in each of the A, B, and D subgenomes [55]. We recovered expected abundance patterns in the panel both for ITS1 rDNA primers (Figure 5f, Supplementary Figures 14 and 15) and for V4 16S rDNA primers (Supplementary Figures 14 and 15). We noticed the ratio of ITS1 to *PolA1* sequences recovered was low; because ITS primers produce amplicons that are highly variable in length some of which may co-migrate with the host amplicon on a gel, the cut-and-mix approach described in Figure 3 could not be used to improve ITS1 representation. However, increasing the ratio of the ITS1:*PolA1* tagging primers from 1:1 to 2:1 (Methods) successfully enriched the ITS1 amplicon without sacrificing relative load determination between samples (Supplementary Figures 14 and 15).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

As a final proof-of-concept experiment for the potential of hamPCR for microbiome research in species with larger genomes, we used bell pepper (*Capsicum annuum*), which has a 3.5 Gb genome [41], approximately 25× larger than *A. thaliana*, and the pepper pathogenic bacterium *Xanthomonas euvesicatoria* (*Xe*) strain 85-10 [56]. We first constructed a simple infiltration panel. After diluting *Xe* 85-10 in $MgCl_2$ buffer to final concentrations of 10^4, 10^5, 10^6, 10^7 and 10^8 CFU / mL, we infiltrated four replicate leaves with each concentration. Immediately afterwards, without further bacterial growth, we harvested leaf discs within inoculated areas using a cork borer (7 mm diameter). From each sample, we used fresh lysate for *Xe* 85-10 CFU counting and processed the remaining lysate into DNA. We used the DNA both for hamPCR, targeting the V4 16S rDNA and the pepper *GI* gene (*CaGI*), and qPCR, targeting the *xopQ* gene for a *Xe* type III effector [57], and the *C. annuum UBI-3* gene for a ubiquitin-conjugating protein [58].

Sequencing the hamPCR libraries revealed that as the *Xe* 85-10 infiltration concentration increased, so did the resulting load of the ASV corresponding to *Xe* 85-10 (Figure 6a, Supplementary Figure 16). The other major bacterial classes detected in the infiltration experiment, comprising commensal bacteria already present in the leaves, were Actinobacteria as well as Alpha-, Beta-, and Gammaproteobacteria. These had similar, low abundances, regardless of the amount of infiltrated *Xe* 85-10 (Figure 6b). When aligned with CFU counts recovered from the same lysates (Methods), the hamPCR *Xe* 85-10 ASV loads showed nearly the same exponential differences between samples, although at lower infiltration concentrations, hamPCR gave a slightly higher estimate than CFU counts (Figure 6c). Like hamPCR, qPCR also gave a higher estimate than CFU counts at low microbial loads. The presence of a low level of native *Xe* on the leaves could potentially explain this discrepancy, because they would be detected by DNA-based methods, but would not have the antibiotic resistance necessary to be counted in the CFU measurements.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

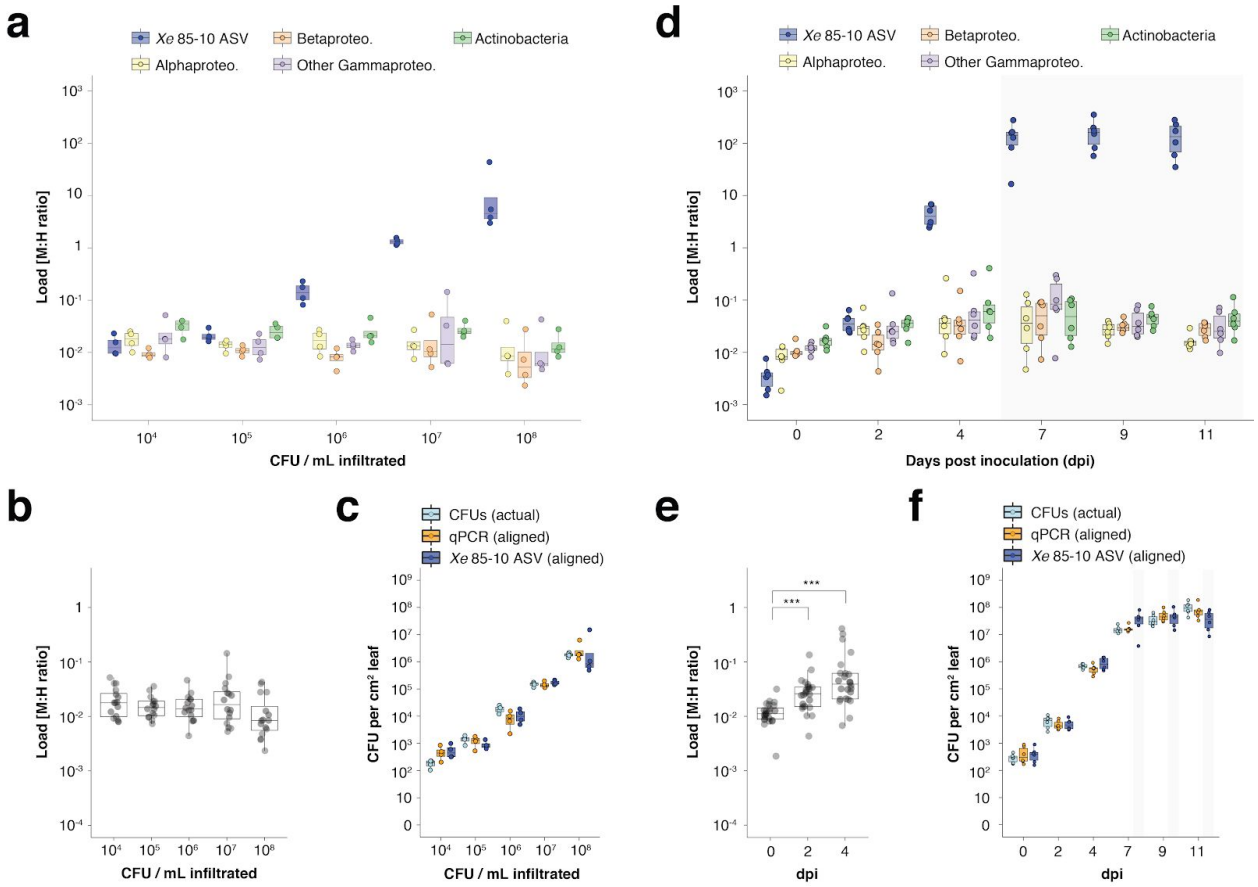

Figure 6. hamPCR can be generalized to *Xanthomonas* infection of *C. annuum* | a-c, Infiltration experiment. All y-axes are on a base-10 logarithmic scale. In all boxplots, the median is represented by a horizontal line and box boundaries show the lower and upper quartiles. Whiskers extend from the box up to 1.5 times the interquartile range. **a**, *Xe* 85-10 was inoculated into *C. annuum* leaves at five concentrations, leaves were harvested immediately afterwards, and hamPCR was performed. The bacterial load is shown for the particular ASV corresponding to *Xe* 85-10, as well as for the major bacterial classes. **b**, The total load for all bacterial classes shown in (a) at each concentration. **c**, Actual CFU counts for *Xe* 85-10 in the infiltration experiment juxtaposed with aligned qPCR and hamPCR loads. **d-f**, Growth curve experiment. **d**, *Xe* 85-10 was inoculated into *C. annuum* leaves at 10^4 CFU/mL. Leaf samples were taken at 0, 2, 4, 7, 9, and 11 days post inoculation (dpi), and hamPCR performed. The corrected load is shown for the particular ASV corresponding to *Xe* 85-10, as well as for the major bacterial classes. **e**, The total load for all bacterial classes shown in (d) at 0, 2, and 4 dpi (*** = $p < 0.001$). **f**, Actual CFU counts for *Xe* 85-10 in the growth curve experiment juxtaposed with aligned qPCR and hamPCR loads.

Having confirmed that hamPCR could capture absolute changes in pathogen abundance in our system, we set up a growth-curve experiment to monitor the pathogen and the commensal microbes through time. We infiltrated six *C. annuum* leaves of six different plants with *Xe* 85-10 at a concentration of 10^4 CFU / mL, and took samples from each plant at 0, 2, 4, 7, 9 and 11 dpi for CFU counting, qPCR, and hamPCR. We observed a rapid increase in *Xe* 85-10 ASV abundance as a result of rapid bacterial growth, leveling off at 7 dpi (Figure 6d). By 7 dpi, bacterial growth had reduced the host *GI* amplicon abundance to below 3%, making the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

load corrections unreliable from day 7 on (gray box, Figure 6d). Notably, the other major bacterial classes, Actinobacteria and the Alpha-, Beta-, and Gammaproteobacteria, also increased in microbial load through time, a trend significant even comparing 2 dpi to 0 dpi (Figure 6e, Mann Whitney U-test, $p < 0.001$). This increase in load for the other classes was not an artifact due to high *Xe* 85-10 titers, because in the infiltration experiment, measurements for these classes had not changed even at higher pathogen concentrations (Figure 6b). This subtle but biologically significant effect would be invisible in a pure 16S rDNA amplicon analysis.

In the growth curve, aligned *Xe* 85-10 ASV loads compared very closely to CFU counts and to aligned qPCR abundances up to 7 dpi (Figure 6f). Unlike in the infiltration experiment, the three methods agreed well at low *Xe* 85-10 concentrations. After 7 dpi, CFU counts and qPCR continued to show slight increases in load. However, hamPCR-derived estimates flatlined; above the concentration reached at 7 dpi, the number of reads comprising the host ASV became too low for accurate load estimates (vertical gray bars, Figure 6f).

Discussion

With hamPCR, we developed a simple and robust method to quantitatively co-amplify one or more microbial amplicons along with an unrelated single-copy host gene in the same reaction, allowing accurate determination of microbial community composition and microbial load from a single sequencing library (Figure 1, Figure 2). Furthermore, we also developed a method to predictably optimize the amount of sequencing effort devoted to microbe vs. host, without losing information about the original microbe to host ratio (Figure 3). This is an important advance with our approach that greatly increases cost-efficiency.

The principle behind hamPCR builds on a body of literature describing related, firmly established techniques, which bodes well for robust and wide-spread adoption of our approach. The two-step PCR protocol on which hamPCR relies is well-established for microbial amplicon sequencing [20,30–32]. The use of multiple primer pairs in the first step of a two-step PCR protocol and limiting the cycles of the first step to reduce priming bias also has precedent [38], as does the use of high-throughput sequencing for quantifying amplicons generated through such a protocol [33]. Co-amplification of diverse fragments from multiple organisms by PCR has proven reliable in low-input shotgun metagenomics and RNA-seq protocols, albeit with some PCR-related biases such as differential amplification based on GC content [33,59,60]. Multiplexed reactions for fluorometric qPCR analyses are also well described, including in a protocol to quantitatively co-amplify microbe and host DNA [61]. This rich literature should increase confidence when implementing hamPCR in microbiome research, and it also provides rich resources for optimization and further development. For example, hamPCR tagging primers could be fitted with UMIs for higher precision, and the protocol could be adapted for sequencing platforms with longer read lengths.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

To perform hamPCR, a universal two-step PCR protocol for amplicon sequencing is required [20,30–34]. A two-step protocol already provides the major advantage that only a one-time investment in a set of universal barcoding primers for a flexible step two is required, which then can be easily adapted to any amplicon by simply swapping in different template-specific primers for step one. For labs already equipped for two-step PCR, hamPCR is simple to implement, requiring only slight adjustments to cycling conditions, and ordering a handful of template-specific host and microbe primers for the first step.

Throughout our reactions, we have used 30 PCR cycles for the exponential step for consistency. However, reducing the PCR cycle number whenever possible will reduce noise and bias. We have demonstrated that the tagging step in hamPCR is sensitive to the relative concentrations between the host and microbe primers (Supplementary Figure 7), consistent with the effects of primer concentration on amplification efficiency in qPCR [62–64]. This property was exploited by Carlson and colleagues, who adjusted primer concentrations until multiple amplicons gave the correct product ratio [33]. We have also shown that it can be exploited to intentionally increase the representation of a microbial amplicon for more efficient sequencing (Supplementary Figure 16a). The effect of primer concentration has important implications for how a large project should be prepared. We recommend that the tagging primers should be carefully pipetted into a multiplexed primer master mix sufficiently large to be used for the entire project. If such a mix is not possible, resequencing of the same control samples across sample batches can allow detection and calibration of slight batch differences in the microbe-to-host ratios.

A limitation of hamPCR is reduced accuracy at very high microbial loads (Supplementary Figure 4). As the number of host sequences used as a denominator approaches 0, the quotient, that is the calculated load, quickly approaches infinity, and small sources of noise become amplified. Higher loads also require deeper sequencing to ensure that enough host reads are sequenced. For our samples, only the most highly-infected samples reached a level that interfered with accurate quantification. We expect that a simple application of hamPCR will be precise and accurate for the majority of colonized hosts. There are three straightforward adjustments, however, that can increase host signal and thus reduce noise for samples with the highest microbial loads. First, adjusting the host and microbe amplicon ratio prior to sequencing, as demonstrated in Figure 3, could be used to increase the overall host representation. If only a subset of the samples has very high loads, one could increase host representation in only this subset, as long as a reference sample is also made from it to reconstruct the original microbe-to-host ratio in that subset. Second, a host gene with a higher copy number could be chosen for template tagging throughout the entire project. Finally, carefully increasing the concentration of the host primer in the tagging reaction could also increase the representation of host (Supplementary Figure 7) [33].

In summary, we have demonstrated that hamPCR is agnostic to the taxonomic identities of the organisms studied – on both the host and microbe side –, their genome sizes, or the

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

functions of the genomic regions amplified. We have also shown that hamPCR can quantitatively monitor three amplicons at the same time, and in principle more amplicons can be multiplexed with careful design. Our focus here has been on tracking hosts and their closely-associated microbes, but the protocol could be adapted to quantitatively relate different amplicons targeting archaea, bacteria, and fungi in diverse "host-free" environments like soil. Besides whole organisms, hamPCR would also enable quantitative monitoring of bacterial populations and sub-genomic elements, such as plasmids or pathogenicity islands that might not be shared by all strains in a population. An exciting application of hamPCR is for the study of endophytic microbial colonization and infection in crop plants, many of which have very large genomes that preclude the analysis of any sizable number of samples by shotgun sequencing. In a previous study, we sequenced leaf metagenomes from over 200 *A. thaliana* plants, at not insignificant costs [3]. In wheat, assuming comparable microbial loads, the same investment would barely be sufficient for two samples due to the size of the wheat genome of over 16 Gb. Other exciting applications are the recognition of cryptic infections [65], tracking of mixed infections, and measurement of pathogen abundances on hosts showing quantitative disease resistance - this could even be accomplished by spiking hamPCR amplicons into the same sequencing run used to genotype the hosts [66]. In sparsely-colonized samples, hamPCR will help prevent inflating the abundance of ultra-low abundance microbes, such as reagent contaminants. Finally, for projects with many samples, the fact that hamPCR derives microbial composition and load from the same library not only saves cost and uses less of the sample, but also simplifies analysis and project organization.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Methods

hamPCR protocol

hamPCR requires two steps: a short 'tagging' reaction of 2 cycles, and a longer 'exponential' reaction. We used 30 cycles throughout this work, although fewer can and should be used if the signal is clear. The primers employed in the tagging reaction were used at ⅛ the concentration of the exponential primers, as this still represents an excess in a reaction run for only two cycles, prevents waste, and reduces dimer formation. See Supplementary Information and Supplementary Table 1 for detailed information about the primers.

Tagging reaction

We used Taq DNA Polymerase (NEB, Ipswich, MA, USA) for the first tagging step, and set up 25 µL reactions as follows.

Master Mix	**For 25 µL**
10x Taq buffer	2.5 µL
Taq polymerase	0.2 µL
dNTPs (10 mM)	0.5 µL
* Tagging primer mix	1.25 µL
** PNAs (mix of mPNA and pPNA each at 50 µM)	0.375 µL
*** DNA (2-10 ng/µL)	5 µL
Water	(to 25 µL)

* tagging primer mix is an equimolar mix of tagging primers with each at a partial concentration of 1.25 µM

** PNA was used to block chloroplast and mitochondrial amplification in reactions involving the V4 or V3V4 region of 16S rDNA. For other reactions, it is not helpful and is omitted.

*** not part of master mix

Each well received 20 µL of master mix and 5 µL of DNA (around 50 ng). Completed reactions were thoroughly mixed on a plate vortex and placed into a preheated thermocycler. We used the following standard cycling conditions:

1) **94° C** for 2 min. ***Denature***
2) **78° C** for 10 sec. ***PNA annealing***
3) **58° C** for 30 sec. ***Primer annealing***
4) **55° C** for 30 sec. ***Primer annealing***
5) **72° C** for 2 min. ***Extension***
6) **GO TO STEP 2** for **1** additional cycle
7) **16° C** forever ***Hold***

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

The tagging reaction was cleaned with Solid Phase Reversible Immobilization (SPRI) beads [43]. All ITS amplicons were cleaned with a 1.1:1 ratio of SPRI beads to DNA, or 27.5 µL beads mixed in 25 µL of tagged template. After securing beads and DNA to a magnet and removing the supernatant containing primers and small fragments, beads were washed twice with 80% ethanol, air dried briefly, and eluted in 17 µL of water. For primer sequences, see Supplementary Information and Supplementary Table 1.

Exponential reaction

15 µL of the tagged DNA from step one was used as template for the exponential reaction. To reduce errors during the exponential phase, we used the proof-reading enzyme Q5 from NEB, with its included buffer. We prepared reactions in either 25 µL for technical tests with replicated samples. For samples prepared without sequenced replicates, we prepared bias-reducing triplicate reactions in which a 40 µL mix was split into 3 parallel reactions of ~13 µL prior to PCR.

Master Mix	**For 25 µL**	**For 40 µL**
5x Q5 buffer	5 µL	8 µL
Q5 polymerase	0.25 µL	0.4 µL
dNTPs (10 mM)	0.5 µL	0.8 µL
100 µM F universal PCR primer	0.0625 µL	0.1 µL
* PNAs (mix of mPNA and pPNA each at 50 µM)	0.375 µL	0.6 µL
** 5 µM reverse barcoded primer	1.25 µL	2 µL
** DNA (from previous reaction)	15 µL	15 µL
Water	(to 25 µL)	(to 40 µL)

* PNA was used to block chloroplast and mitochondrial amplification in reactions involving the V4 or V3V4 region of 16S rDNA. For other reactions, it is not helpful and is omitted.

** not part of master mix

We first distributed 8.75 µL (or 23 µL for 40 µL mixes) of master mix to each well. We then added 15 µL of the DNA from the tagging reaction and 1.25 µL (or 2 µL for 40 µL mixes) of 5 µM barcoded reverse primer. For the 40 µL mixes, 13 µL was pipetted into two new PCR wells. The PCR reactions were placed into a hot thermocycler and cycled with the following standard conditions:

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

1) **94° C** for 2 min. ***Denature***
2) **94° C** for 30 sec. ***Denature***
3) **78° C** for 5 sec. ***PNA annealing***
4) **60° C** for 1 min. ***Primer annealing***
5) **72° C** for 1 min. ***Extension***
6) **GO TO STEP 2** for **29** additional cycles
7) **16° C** forever ***Hold***

Following PCR, sets of three 13 µL reactions were recombined to 40 µL. For primer sequences, see Supplementary Information and Supplementary Table 1.

Library quality control and pooling

For visualization, 5 µL of PCR product was mixed with 3 µL of 6x loading dye and all 8 µL loaded on a 2% agarose gel and stained with ethidium bromide. The remaining PCR products were cleaned with a SPRI-to-DNA ratio of 1.1:1.0 (v/v). The DNA concentrations in the cleaned products were measured with PicoGreen (Invitrogen, Carlsbad, CA, USA) and samples were pooled at equimolar total DNA ratios. We note that because host and microbial fractions are independently visible on the gel, it would also be possible to measure the quantity of microbial products with image analysis software such as ImageJ [67] and pool at equimolar microbial ratios.

The pooled library was diluted to ~1 ng/uL and run on a Bioanalyzer High Sensitivity DNA chip (Agilent, Santa Clara, CA, USA) to check library purity and to estimate the expected ratio of host to microbial amplicons in the sample.

Pre-sequencing adjustment of host : microbe Ratio

To adjust the host-to-microbe ratio in the "synthetic template panel" and "cycle number test" prior to sequencing on a HiSeq3000 instrument (Illumina, San Diego, CA, USA), four reference samples were first made by rebarcoding the original pooled library (Supplementary Figures 9 and 10). To accomplish this, ~5 ng of of the pooled library was used in a 30 µL PCR reaction as follows:

Master Mix	**For 30 µL**
5x Q5 buffer	6 µL
Q5 polymerase	0.3 µL
dNTPs (10 mM)	0.6 µL
100 µM F universal PCR primer	0.075 µL
** 5 µM reverse barcoded primer	1.5 µL
** 5 ng original pooled library	5 µL
Water	16.45 µL

** not part of master mix

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

After distributing 23.5 µL of master mix to each well, 5 µL of the diluted original library was added to each well (5 ng total), along with 1.5 µL of 5 µM barcoded reverse primer. Just prior to placing the reactions in the thermocycler, a 5 µL pre-PCR aliquot was removed from each one and kept on ice to preserve the pre-PCR concentrations. The remaining 25 µL reaction was placed into a preheated thermocycler and run for 8 cycles, using the following cycling conditions:

1) **94° C** for 2 min. ***Denature***
2) **94° C** for 30 sec. ***Denature***
3) **78° C** for 5 sec. ***PNA annealing***
4) **60° C** for 1 min. ***Primer annealing***
5) **72° C** for 1 min. ***Extension***
6) **GO TO STEP 2** for **7** additional cycles
7) **16° C** forever ***Hold***

Following PCR, the pre-PCR aliquots were run alongside 5 µL of post-PCR product on a 2% gel to confirm successful amplification of the reference libraries (Supplementary Figure 10). The remaining 20 µL of PCR reactions were then cleaned with SPRI beads (1.5 : 1.0 [v/v]) and set aside. A large aliquot of the original library (approximately 50 ng) was also run on a 2% gel to separate the host and microbe bands for individual purification. The bands were cut out of the gel and each band was put into a separate Econospin spin column (Epoch Life Sciences, Missouri City, TX, USA) without any other liquids or binding buffer. The gel slices were centrifuged at maximum speed to force the liquid containing the DNA into the bottom chamber, leaving the dried gel on top. The eluted DNA was cleaned with SPRI beads at 1.5 : 1.0 (v/v) and eluted in EB.

The purified pooled host library fraction, pooled microbe library fraction, and each of the four reference libraries were quantified with Picogreen and the molarity of each was estimated. The pools were then mixed together, targeting host molarity at 5% of the total and each reference library at 1% of the total.

Illumina sequencing

Pooled and quality-checked sequencing libraries were cleaned of all remaining dimers and off-target fragments using a BluePippin (Sage Science, Beverly, MA, USA) set to a broad range of 280 to 720 bp. The libraries were then diluted for Illumina sequencing following manufacturers' protocols. Libraries were first diluted to 2.5 - 2.8 nM in elution buffer (EB, 10 mM Tris pH 8.0) and spiked into a compatible lane of the HiSeq3000 instrument (2 x 150 bp paired end reads) to occupy 2-3% of the lane. Samples were sequenced across four total lanes (Supplementary Table 1)

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Sequence processing

The sequences were demultiplexed first by the 9 bp barcode on the PCR primers (Supplementary Table 1), of which there are 96, not allowing for any mismatches. In some cases in which two samples differed in both their host and microbe primer sets, we amplified both samples with the same 9 bp barcode to increase multiplexing; such samples were further demultiplexed using regular expressions for the forward primer and reverse primer sequences. Following demultiplexing, all samples were filtered to remove sequences with any mismatches to the expected primers. With HiSeq3000 150 bp read lengths, overlap of read 1 and read 2 was not possible for our amplicons, and therefore only read 1 was processed further.

All primer sequences were removed. Additional quality filtering, removal of chimeric sequences, preparation and Amplicon Sequence Variant (ASV) tables, and taxonomic assignment were done with a combination of VSEARCH [68] and USEARCH10 [45]. ASVs were prepared as 'zero-radius OTUs' (zOTUs) [45]. The 16S rDNA taxonomy was classified based on the RDP training set v16 (13k seqs.)[69], and ITS1 taxonomy of the top 10 most abundant fungal ASVs was classified manually using the UNITE database[70] (https://unite.ut.ee/). To reduce memory usage, data from the five lanes was processed into four independent ASV tables (Supplementary Data), as described in the sample metadata (Supplementary Table 1).

ASV tables were analyzed statistically and graphically using custom scripts in R[71], particularly with the help of packages "ggplot2"[72] and "reshape2"[73]. Custom scripts are available on GitHub at (https://github.com/derekLS1/hamPCR).

Samples

Synthetic titration panel

Seeds from the *Arabidopsis thaliana* accession Col-0 were surface sterilized by immersion for 1 minute in 70% ethanol with 0.1% Triton X-100, soaking in 10% household bleach for 12 minutes, and washing three times with sterile water. Seeds were germinated axenically on ½ strength MS media with MES, and about 2 g of seedlings were harvested after 10 days. DNA was extracted in the sterile hood as in [3] and diluted to 10 ng/µL in elution buffer (10 mM Tris pH 8.0, hereafter EB). Pure *E. coli* and *Sphingomonas* sp. cultures were likewise grown with LB liquid and solid media respectively, and DNA was extracted using a bead beating protocol [3]. *E. coli* DNA was used separately, or alternatively pooled with the mixed *Sphingomonas* DNA, and diluted to 10 ng/µL. The plant DNA and microbial DNA were then combined according to the following table:

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

	100E	100	24	20	16	8	4	2	1	0.5	0.25	100P	100P	blank
µL *A .thaliana* DNA (10 ng/µL)			760	800	840	920	960	980	990	995	997.5	1000	1000	
µL mixed bacterial DNA (10ng/µL)		1000	240	200	160	80	40	20	10	5	2.5			
µL E.coli DNA (10 ng/µL)	1000													
µL water														1000

V4 tagging was performed with 515_F1_G-46603 and 799_R1_G-46601 (V4 16S rDNA) and At.GI_F1_G-46602 and At.GI_R502bp_G-46614 (*A. thaliana GI*). Each exponential PCR reaction was completed in a single reaction of 25 µL.

Synthetic equimolar plasmid template

The ITS1 region from *Agaricus bisporus*, a fragment of the *GI* gene from *A. thaliana* Col-0 accession, the 16S rRNA gene from *Pto* DC3000, and the ITS1 region from *H. arabidopsidis* were PCR amplified individually, combined into one fragment via overlap extension PCR, and cloned into pGEM®-T Easy (Promega, Madison, WI, USA). The sequences of these templates can be found in Supplementary Information.

Wild *A. thaliana* samples

DNA from chosen samples previously analyzed by conventional 16S rDNA-only sequencing of the V4 region and shotgun metagenomics [3] was reused, chosen to capture a wide range of realistic bacterial loads. The samples were individually assayed with hamPCR using 5 µL DNA template (approximately 50 ng). V4 tagging was performed with 515_F1_G-46603 and 806_R1_G-46631 (V4 16S rDNA) and At.GI_F1_G-46602 and At.GI_R502bp_G-46614 (502 bp *A. thaliana GI*). These were the only V4 samples tagged with the 806R primer instead of the nearby 799R primer, and it was used to enable direct comparison to the dataset in [3]. V3V4 tagging was performed with 341_F1_G-46605 and 799_R1_G-46601 (V3V4 16S rDNA) and At.GI_F1_G-46602 and At.GI_R502bp_G-46614 (502 bp *A. thaliana GI*). V5V6V7 tagging was performed with 799_F1_G-46628 and 1192_R1_G-46629 (V5V6V7 16S rDNA) and At.GI_F1_G-46602 and At.GI_R466bp_G-46652 (466 bp *A. thaliana GI*). Each exponential PCR reaction was completed in a single reaction of 25 µL; each sample was replicated three times.

Wild *A. thaliana* mixed sample

DNA from samples previously analyzed by conventional 16S rDNA-only sequencing of the V4 region and shotgun metagenomics [3] were pooled to prepare a single abundant mixed sample to be used repeatedly for technical tests.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Hyaloperonospora arabidopsidis and *Pseudomonas syringae pv. tomato* DC3000 co-infection

Both wildtype *A. thaliana* seedlings in the Col-0 genetic background and *enhanced disease susceptibility 1* mutants in the Ws-0 genetic background (*eds1-1)* were grown from surface-sterilized seeds. Seedlings were raised in ED73 potting mix (Einheitserdewerke, Sinntal-Altengronau, Germany) in 5 cm pots for 10 days under short-day conditions (8 hours light, 16 hours dark). Each pot contained 4 to 5 seedlings, and for each genotype, four pots were used for each infection condition. Plants were treated with either 10 mM $MgCl_2$ (buffer only), *H. arabidopsidis* (*Hpa*) isolate 466-1 alone (5 x 10^4 spores / mL), or *Hpa* 466-1 with *P. syringae pv. tomato* (*Pto*) DC3000 (OD600 = 0.25, a gift from El Kasmi lab, University of Tübingen).

The infected plants were grown at 16°C for 8 days (10 hours light, 14 hours dark) and harvested by pooling all seedlings in each pot into a sterile pre-weighed tube, which was again weighed to find the mass of the seedlings. Three 5 mm glass balls and 300 µL 10 mM $MgCl_2$ were added to each tube and the plant cells were lysed at a speed of 4.0 m/s for 20 seconds in a FastPrep-24™ Instrument (MP Biomedicals, Illkirch-Graffenstaden, France) to release the live bacteria from the leaves. From the pure lysate, 20 µL was used for a serial log dilution series, and 5 µL of each dilution was plated on LB agar supplemented with 100 µg/mL rifampicin. Colony forming units (CFUs) were counted after 2 days of incubation at 28°C. The remaining 280 µL of lysate were combined with 520 µL DNA lysis buffer, 0.5 mL of 1 mm garnet sharp particles (BioSpec, Bartlesville, OK, USA). 60 µL of 20% SDS was added to make a final SDS concentration of 1.5%, and DNA was extracted using a bead beating protocol [3]. The number of *Hpa* sporangiophores was too high to be accurately quantified by visual counting.

The DNA preps were individually assayed with hamPCR using 5 µL DNA template (approximately 30 ng); tagging was performed with three primer sets: Ha.Actin_F1_G-46716 and Ha.Actin_R1_G-46717 (Hpa Actin), At.GI_F1_G-46602 and At.GI_R502bp_G-46614 (502 bp *A. thaliana GI*), and 515_F1_G-46603 and 799_R1_G-46601 (V4 16S rDNA). Each exponential PCR reaction was completed in three parallel reactions of 13 µL, which were recombined prior to sequencing.

Titration with plant DNA infected with *H. arabidopsidis*

A titration panel was made combining different amounts of DNA from uninfected plants (*eds1-1* treated only with 10 mM $MgCl_2$) and DNA from *Hpa*-infected plants (*eds1-1* infected with *Hpa* as described above). Infected and uninfected pools were each diluted to 6 ng/µL, and combined in 0:7, 1:6, 2:5, 3:4, 4:3, 5:2, 6:1, and 7:0 ratios. These were tagged using the same three primer sets described above for *Hpa* actin, *A. thaliana GI*, and V4 16S rDNA above. Each

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

exponential PCR reaction was completed in a single reaction of 25 µL; hamPCR was replicated on the titration three times.

Capsicum annuum infections with *Xanthomonas*

Leaf infiltration log series: Using pressure infiltration with a blunt-end syringe, *C. annuum* cultivar Early Calwonder (ECW) leaves were inoculated with *Xanthomonas euvesicatoria* (*Xe*). *Xe* strain 85-10 [56] was resuspended in 10 mM $MgCl_2$ to final concentration of 10^8 CFU / mL (OD_{600}=0.4) and further diluted to 10^7, 10^6, 10^5 and 10^4 CFU / mL. Upon infiltration, 5 leaf discs (7 mm diameter) were punched from each leaf per sample and placed in a 2 mL round-bottom tube with two SiLibeads (type ZY-S 2.7-3.3 mm, Sigmund Lindner GmbH, Warmensteinach, Germany) and 300 µL 10 mM $MgCl_2$. The samples were ground by bead beating for 25 sec at 25 Hz using a Tissue Lyser II machine (Qiagen, Hilden, Germany). For CFU-based bacterial enumeration, 30 µL of the lysate or 30 µL of serial dilutions were plated on NYG medium (0.5 % peptone, 0.3 % yeast extract, 0.2 % glycerol and 1.5 % agar [74] containing rifampicin (100 µg/ml). *Xe* bacteria were counted 3 days post incubation at 28°C. The remaining 250 µL of lysate was combined with 600 µL of DNA lysis buffer containing 2.1% SDS (for a 1.5% final SDS concentration) and transferred to screw cap tubes filled with 1 mm garnet sharp particles, for a bead-beating DNA prep as previously described [3].

Growth curve: *Xe* strain 85-10, resuspended in 10 mM $MgCl_2$ to a final concentration of 10^4 CFU / mL, was infiltrated via a blunt end syringe into 6 *C. annuum* (ECW) leaves of 6 different plants. Upon 0, 2, 4, 7, 9 and 11 dpi (days post inoculation) 4 leaf discs (7 mm diameter) from each inoculated leaf were harvested and bacterial numbers were determined as described above. 250 µL of leaf lysates were used for a bead-beating DNA prep as described for all other samples above.

Each hamPCR template tagging reaction used 5-10 µL template (approximately 50 ng each); tagging was performed with primers 515_F3_G-46694 and 799_R1_G-46601 (V4 16S rDNA), and Ca.GI_F1_G-46626 and Ca.GI_R1_G-46627 (*C. annuum GI*). Each exponential PCR reaction was completed in three parallel reactions of 13 µL, which were recombined prior to sequencing.

Pristionchus pacificus titration panel

Pristionchus pacificus strain PS312 [75] was grown on nematode growth media (NGM) plates supporting a bacterial lawn of either pure *E. coli* OP50 or alternatively a mix of *E. coli* OP50, *Pto* DC3000, and *Lysinibacillus xylanilyticus* (a strain isolated from wild *P. pacificus*). The worms were washed extensively with PBS buffer to remove epidermally-attached bacteria, and DNA was prepared from whole worms using the same bead beating protocol as described for *A. thaliana* [3]. Worm DNA from the pure culture and the mixed culture were each diluted to 6 ng/µL, and combined in 0:7, 1:6, 2:5, 3:4, 4:3, 5:2, 6:1, and 7:0 ratios to create a titration panel.

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Each hamPCR template (5 µL template or 30 ng total) was used to perform the tagging reaction, using primers 799F1_G-46628 and 1192R1_G-46629 (V5V6V7 16S rDNA), and Pp_csq-1_F1_G-46691 and Pp_csq-1_R1_G-46692 (*P. pacificus csq-1*). Each exponential PCR reaction was completed in a single reaction of 25 µL; the titration was replicated three times.

Triticum aestivum titration panel

Triticum aestivum (wheat) seeds (Rapunzel Naturkost, Legau, Germany) were surface-sterilized by immersion in 70% ethanol and 0.1% Triton X-100 for 1 minute, soaking for 15 minutes in 10% household bleach, and finally washing three times in sterile autoclaved water. Axenic plants were grown on 1% agar supplemented with 1/2 strength MS medium buffered with MES. About 1 g of sterile leaf tissue was harvested after 10 days, and DNA was extracted in the sterile hood as described in ref. [3]. Roots that had been spontaneously colonized by microbes were obtained by growing by transplanting germinated seeds outdoors into potting soil. Roots were harvested from approximately 4-week old plants and surface-sterilized by immersion in 10% household bleach with 0.1% Triton X-100 for 5 minutes, followed by 3 washes with sterile water. Axenic leaf DNA and spontaneously-colonized root DNA were each diluted to 60 ng/µL and combined in 0:7, 1:6, 2:5, 3:4, 4:3, 5:2, 6:1, and 7:0 ratios to create a titration panel of eight samples. Each hamPCR tagging reaction used 3 µL (~180 ng) template; fungal ITS1 tagging was performed with primers ITS1_F1_G-46622 and ITS2_R1_G-46623 (ITS1 rDNA), and PolA1_F1_G-46750 and PolA1_R1_G-46751 (*T. aestivum* RNA polymerase 1 gene, *PolA1*). Bacterial 16S rDNA tagging was performed with the same *PolA1* primers and with 515_F1_G-46603 and 799_R1_G-46601 (V4 16S rDNA). To make an additional ITS1 library enriched for ITS1 amplicons, the ITS1 primer pair concentration was increased by a factor of 1.33 and the *PolA1* primer pair concentration was decreased by a factor of 0.66, giving a 2:1 ratio instead of the standard 1:1 ratio, and the tagged products were amplified with 7 tagging and 25 PCR cycles instead of the standard 2 tagging and 30 PCR cycles.

Test of tagging step cycle numbers

As templates, we used a pool of mixed wild *A. thaliana* leaf DNA (~ 50 ng / reaction) and the "synthetic equimolar plasmid template" (~ 0.05 ng / reaction). For the wild *A. thaliana* leaf DNA, we tested V4 16S rDNA primers alone in the tagging step vs. hamPCR with V4 16S rDNA primers plus primers for the host *GI* gene. For the "synthetic equimolar plasmid template", we used only hamPCR. Specifically, we used 515_F1_G-46603 and 799_R1_G-46601 (V4 16S rDNA) and At.GI_F1_G-46602 and At.GI_R502bp_G-46614 (502 bp *GI* gene).

We applied hamPCR for 2, 3, 4, 5, 6, 7, 8, 9, or 10 tagging cycles, paired with 30, 29, 28, 27, 26, 25, 24, 23, or 22 PCR cycles, respectively. All tagging and PCR reactions were started together, and fewer tagging cycles than 10, or fewer PCR cycles than 30, were achieved by

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

taking PCR tubes out of the thermocycler at the end of the appropriate extension steps and placing them on ice.

Tests of template concentrations

A panel of 8 concentrations of wild *A. thaliana* leaf DNA was prepared, ranging from 5 to 500 ng per reaction. Primers for the wild *A. thaliana* leaf DNA were 515_F1_G-46603 and 799_R1_G-46601 (V4 16S rDNA) and At.GI_F1_G-46602 and At.GI_R502bp_G-46614 (502 bp Gigantea gene), with both primer pairs in equal ratio.

Quantitative real time PCR on *C. annuum* samples

A primer set targeting the gene for the type III effector XopQ of pathogenic *Xanthomonas* was used to measure abundance of *Xe* 85-10 [57]. For *C. annuum*, primers targeting the *UBI-3* gene encoding a ubiquitin-conjugating protein were used [58]. Two reagent mastermixes were prepared, one for each primer set, to help improve primer dose consistency. Each sample was amplified using three 10 µL technical replicates per primer set that were averaged for analysis. Each 10 µL reaction included 2.5 µL of DNA, to which was added, as a mastermix, 5 µL SYBR® Green PCR Master Mix (Life Technologies, Carlsbad, California), 1.5 µL water, 0.5 µL of 100 µM forward primer, and 0.5 µL of 100 µM reverse primer. qPCR was performed on a BioRad CFX384 Real-time System and analyzed with the CFX Manager Software. The following conditions were used for the amplification of both target genes:

1) **94° C** for 5 min. ***Denature***
2) **94° C** for 30 sec. ***Denature***
3) **55° C** for 30 sec. ***PNA annealing***
4) **68° C** for 45 sec. ***Extension***
5) **Image fluorescence**
6) **GO TO STEP 2** for **39** additional cycles

The ratio of microbial to host DNA was initially calculated as 2^(-mean *xopQ* Cq value) / 2^(-mean *UBI-3* Cq value). See alignment to CFU counts below.

Alignment of *Xanthomonas* qPCR and hamPCR load with CFU counts

Log10-transformed *Xe* 85-10 ASV loads from hamPCR were regressed onto log10-transformed *xopQ* loads from qPCR (least squares method), and the slope (*m*) and y-intercept (*b*) of the best-fit line were used to transform and align the qPCR loads to hamPCR loads with the following formula: $\text{Load_qPCR}_{\text{hamPCR-aligned}} = m \times \text{Load_qPCR} + b$. Next, log10-transformed CFU counts were regressed onto the log10-transformed hamPCR loads, and

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

the slope and y-intercept of the resulting best-fit line were used similarly to align both $Load_qPCR_{hamPCR\text{-}aligned}$ and hamPCR loads to the CFU counts.

Data availability

All data in this manuscript has been deposited in the European Nucleotide Archive (ENA). It can be accessed under the project number PRJEB38287. At https://www.ebi.ac.uk/ena

Author contributions

DL and PP planned the study. DL led the data analysis. DL and PP performed all DNA extractions and prepared all hamPCR libraries for sequencing. PP constructed the synthetic plasmid template. AS and TL planned, and AS conducted, the pepper and *Xanthomonas* infections. DL and GS planned and conducted the *Hpa* and *Pto* DC3000 co-infection of *A. thaliana*. WL grew and harvested the colonized *P. pacificus* samples. DL, PP, and DW wrote the manuscript.

Acknowledgements

We thank Lei Li and Fiona Beitel for discussions and suggestions regarding PCR and qPCR, Talia Karasov for suggestions on presentation, Katrin Fritschi and Heike Budde for technical help sequencing, Ilja Bezrukov for assistance with demultiplexing, and Dor Russ and Jeff Dangl for supportive discussions and comments on the manuscript. Thanks to Victor Schmidt, Nick Youngblut, and Hua Wang for giving us space in their HiSeq 3000 lanes. Supported by a Human Frontiers Science Program (HFSP) Long-Term Fellowship (LT000565/2015-L to DL), the DFG through SPP Priority Program DECRyPT, and the Max Planck Society (DW).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

Competing interests

The authors declare no competing interests.

References

1. Tsilimigras, M. C. B. & Fodor, A. A. Compositional data analysis of the microbiome: fundamentals, tools, and challenges. *Ann. Epidemiol.* **26**, 330–335 (2016).
2. Gloor, G. B., Macklaim, J. M., Pawlowsky-Glahn, V. & Egozcue, J. J. Microbiome Datasets Are Compositional: And This Is Not Optional. *Front. Microbiol.* **8**, 2224 (2017).
3. Regalado, J. *et al.* Combining whole-genome shotgun sequencing and rRNA gene amplicon analyses to improve detection of microbe–microbe interaction networks in plant leaves. *ISME J.* (2020) doi:10.1038/s41396-020-0665-8.
4. Barlow, J. T., Bogatyrev, S. R. & Ismagilov, R. F. A quantitative sequencing framework for absolute abundance measurements of mucosal and lumenal microbial communities. *Nat. Commun.* **11**, 2590 (2020).
5. Vandeputte, D. *et al.* Quantitative microbiome profiling links gut community variation to microbial load. *Nature* **551**, 507–511 (2017).
6. Humphrey, P. T. & Whiteman, N. K. Insect herbivory reshapes a native leaf microbiome. *Nature Ecology & Evolution* 1–9 (2020).
7. Props, R. *et al.* Absolute quantification of microbial taxon abundances. *ISME J.* **11**, 584 (2016).
8. Tkacz, A., Hortala, M. & Poole, P. S. Absolute quantitation of microbiota abundance in environmental samples. *Microbiome* **6**, 110 (2018).
9. Smets, W. *et al.* A method for simultaneous measurement of soil bacterial abundances and community composition via 16S rRNA gene sequencing. *Soil Biol. Biochem.* **96**, 145–151 (2016).
10. Stämmler, F. *et al.* Adjusting microbiome profiles for differences in microbial load by spike-in bacteria. *Microbiome* **4**, 28 (2016).
11. Niu, B., Paulson, J. N., Zheng, X. & Kolter, R. Simplified and representative bacterial community of maize roots. *Proc. Natl. Acad. Sci. U. S. A.* **114**, E2450–E2459 (2017).
12. Ogier, J.-C., Pagès, S., Frayssinet, M. & Gaudriault, S. Entomopathogenic nematode-associated microbiota: from monoxenic paradigm to pathobiome. *Microbiome* **8**, 25 (2020).
13. Parker, E. S., Newton, I. L. G. & Moczek, A. P. (My Microbiome) Would Walk 10,000 miles: Maintenance and Turnover of Microbial Communities in Introduced Dung Beetles. *Microb. Ecol.* (2020) doi:10.1007/s00248-020-01514-9.
14. Gendrin, M. *et al.* Antibiotics in ingested human blood affect the mosquito microbiota and capacity to transmit malaria. *Nat. Commun.* **6**, 5921 (2015).
15. Davies, D. R. DNA contents and cell number in relation to seed size in the genus Vicia. *Heredity* **39**, 153 (1977).
16. Massonnet, C. *et al.* New insights into the control of endoreduplication: endoreduplication could be driven by organ growth in Arabidopsis leaves. *Plant Physiol.* **157**, 2044–2055 (2011).
17. Karasov, T. L. *et al.* The relationship between microbial biomass and disease in the Arabidopsis thaliana phyllosphere. *bioRxiv* 828814 (2019) doi:10.1101/828814.
18. Karasov, T. L. *et al.* Arabidopsis thaliana and Pseudomonas Pathogens Exhibit Stable Associations over Evolutionary Timescales. *Cell Host Microbe* **24**, 168–179.e4 (2018).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

19. Lebeis, S. L. *et al.* Salicylic acid modulates colonization of the root microbiome by specific bacterial taxa. *Science* **349**, 860–864 (2015).
20. Lundberg, D. S., Yourstone, S., Mieczkowski, P., Jones, C. D. & Dangl, J. L. Practical innovations for high-throughput amplicon sequencing. *Nat. Methods* **10**, 999–1002 (2013).
21. Agler, M. T., Mari, A., Dombrowski, N., Hacquard, S. & Kemen, E. M. New insights in host-associated microbial diversity with broad and accurate taxonomic resolution. *bioRxiv* 050005 (2016) doi:10.1101/050005.
22. Chelius, M. K. & Triplett, E. W. The Diversity of Archaea and Bacteria in Association with the Roots of Zea mays L. *Microb. Ecol.* **41**, 252–263 (2001).
23. Anderson, R. G. & McDowell, J. M. A PCR assay for the quantification of growth of the oomycete pathogen Hyaloperonospora arabidopsidis in Arabidopsis thaliana. *Mol. Plant Pathol.* **16**, 893–898 (2015).
24. Nadkarni, M. A., Martin, F. E., Jacques, N. A. & Hunter, N. Determination of bacterial load by real-time PCR using a broad-range (universal) probe and primers set. *Microbiology* **148**, 257–266 (2002).
25. Guo, X. *et al.* Host-Associated Quantitative Abundance Profiling Reveals the Microbial Load Variation of Root Microbiome. *Plant Communications* 100003 (2019).
26. Tourlousse, D. M. *et al.* Synthetic spike-in standards for high-throughput 16S rRNA gene amplicon sequencing. *Nucleic Acids Res.* **45**, e23 (2017).
27. Chen, T. *et al.* A plant genetic network for preventing dysbiosis in the phyllosphere. *Nature* (2020) doi:10.1038/s41586-020-2185-0.
28. Jian, C., Luukkonen, P., Yki-Järvinen, H., Salonen, A. & Korpela, K. Quantitative PCR provides a simple and accessible method for quantitative microbiota profiling. *PLoS One* **15**, e0227285 (2020).
29. Rabanal, F. A. *et al.* Unstable Inheritance of 45S rRNA Genes in Arabidopsis thaliana. *G3* **7**, 1201–1209 (2017).
30. Gohl, D. M. *et al.* Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **34**, 942–949 (2016).
31. Holm, J. B. *et al.* Ultrahigh-Throughput Multiplexing and Sequencing of >500-Base-Pair Amplicon Regions on the Illumina HiSeq 2500 Platform. *mSystems* **4**, (2019).
32. de Muinck, E. J., Trosvik, P., Gilfillan, G. D., Hov, J. R. & Sundaram, A. Y. M. A novel ultra high-throughput 16S rRNA gene amplicon sequencing library preparation method for the Illumina HiSeq platform. *Microbiome* **5**, 68 (2017).
33. Carlson, C. S. *et al.* Using synthetic templates to design an unbiased multiplex PCR assay. *Nat. Commun.* **4**, 2680 (2013).
34. Symeonidi, E., Regalado, J., Schwab, R. & Weigel, D. CRISPR-finder: A high throughput and cost effective method for identifying successfully edited *A. thaliana* individuals. *bioRxiv* 171538 (2020) doi:10.1101/2020.06.25.171538
35. Fields, B., Moeskjær, S., Friman, V.-P., Andersen, S. U. & Young, J. P. W. MAUI-seq: Metabarcoding using amplicons with unique molecular identifiers to improve error correction. *bioRxiv* 538587 (2020) doi:10.1101/538587.
36. Karst, S. M. *et al.* Enabling high-accuracy long-read amplicon sequences using unique molecular identifiers with Nanopore or PacBio sequencing. *bioRxiv* 645903 (2020) doi:10.1101/645903.
37. Kou, R. *et al.* Benefits and Challenges with Applying Unique Molecular Identifiers in Next Generation Sequencing to Detect Low Frequency Mutations. *PLoS One* **11**, e0146638 (2016).
38. Wen, D. & Zhang, C. Universal Multiplex PCR: a novel method of simultaneous amplification of multiple DNA fragments. *Plant Methods* **8**, 32 (2012).
39. Quince, C., Walker, A. W., Simpson, J. T., Loman, N. J. & Segata, N. Shotgun metagenomics, from sampling to analysis. *Nat. Biotechnol.* **35**, 833–844 (2017).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

40. Kukurba, K. R. & Montgomery, S. B. RNA Sequencing and Analysis. *Cold Spring Harb. Protoc.* **2015**, 951–969 (2015).
41. Kim, S. *et al.* Genome sequence of the hot pepper provides insights into the evolution of pungency in Capsicum species. *Nat. Genet.* **46**, 270–278 (2014).
42. International Wheat Genome Sequencing Consortium (IWGSC) *et al.* Shifting the limits in wheat research and breeding using a fully annotated reference genome. *Science* **361**, (2018).
43. Rohland, N. & Reich, D. Cost-effective, high-throughput DNA sequencing libraries for multiplexed target capture. *Genome Res.* **22**, 939–946 (2012).
44. Duarte, J. M. *et al.* Identification of shared single copy nuclear genes in Arabidopsis, Populus, Vitis and Oryza and their phylogenetic utility across various taxonomic levels. *BMC Evol. Biol.* **10**, 61 (2010).
45. Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460–2461 (2010).
46. Song, L. & Xie, K. Engineering CRISPR/Cas9 to mitigate abundant host contamination for 16S rRNA gene-based amplicon sequencing. *Microbiome* **8**, 80 (2020).
47. Vet, J. A. *et al.* Multiplex detection of four pathogenic retroviruses using molecular beacons. *Proc. Natl. Acad. Sci. U. S. A.* **96**, 6394–6399 (1999).
48. Graspeuntner, S., Loeper, N., Künzel, S., Baines, J. F. & Rupp, J. Selection of validated hypervariable regions is crucial in 16S-based microbiota studies of the female genital tract. *Sci. Rep.* **8**, 9678 (2018).
49. Kembel, S. W., Wu, M., Eisen, J. A. & Green, J. L. Incorporating 16S gene copy number information improves estimates of microbial diversity and abundance. *PLoS Comput. Biol.* **8**, e1002743 (2012).
50. Shuber, A. P., Grondin, V. J. & Klinger, K. W. A simplified procedure for developing multiplex PCRs. *Genome Res.* **5**, 488–493 (1995).
51. Coates, M. E. & Beynon, J. L. Hyaloperonospora Arabidopsidis as a pathogen model. *Annu. Rev. Phytopathol.* **48**, 329–345 (2010).
52. Xin, X.-F. & He, S. Y. Pseudomonas syringae pv. tomato DC3000: a model pathogen for probing disease susceptibility and hormone signaling in plants. *Annu. Rev. Phytopathol.* **51**, 473–498 (2013).
53. Bhandari, D. D. *et al.* An EDS1 heterodimer signalling surface enforces timely reprogramming of immunity genes in Arabidopsis. *Nat. Commun.* **10**, 772 (2019).
54. Huot, B. *et al.* Dual impact of elevated temperature on plant defence and bacterial virulence in Arabidopsis. *Nat. Commun.* **8**, 1808 (2017).
55. Rai, B., Takahashi, H., Kato, K., Sato, Y.-I. & Nakamura, I. Single-copy nuclear PolA1 gene sheds light on the origin of S genome with relationships to B and G genomes of polyploid wheat species. *Genet. Resour. Crop Evol.* **59**, 1713–1726 (2012).
56. Thieme, F. *et al.* Insights into genome plasticity and pathogenicity of the plant pathogenic bacterium Xanthomonas campestris pv. vesicatoria revealed by the complete genome sequence. *J. Bacteriol.* **187**, 7254–7266 (2005).
57. Doddaraju, P. *et al.* Reliable and early diagnosis of bacterial blight in pomegranate caused by Xanthomonas axonopodis pv. punicae using sensitive PCR techniques. *Sci. Rep.* **9**, 10097 (2019).
58. Wan, H. *et al.* Identification of reference genes for reverse transcription quantitative real-time PCR normalization in pepper (Capsicum annuum L.). *Biochem. Biophys. Res. Commun.* **416**, 24–30 (2011).
59. Bowers, R. M. *et al.* Impact of library preparation protocols and template quantity on the metagenomic reconstruction of a mock microbial community. *BMC Genomics* **16**, 856 (2015).
60. Rinke, C. *et al.* Validation of picogram- and femtogram-input DNA libraries for microscale metagenomics. *PeerJ* **4**, e2486 (2016).

bioRxiv preprint doi: https://doi.org/10.1101/2020.05.19.103937; this version posted June 25, 2020. The copyright holder for this preprint (which was not certified by peer review) is the author/funder, who has granted bioRxiv a license to display the preprint in perpetuity. It is made available under aCC-BY 4.0 International license.

61. Weller, S. A., Elphinstone, J. G., Smith, N. C., Boonham, N. & Stead, D. E. Detection of Ralstonia solanacearum strains with a quantitative, multiplex, real-time, fluorogenic PCR (TaqMan) assay. *Appl. Environ. Microbiol.* **66**, 2853–2858 (2000).
62. Bustin, S. & Huggett, J. qPCR primer design revisited. *Biomol Detect Quantif* **14**, 19–28 (2017).
63. Pierce, K. E., Sanchez, J. A., Rice, J. E. & Wangh, L. J. Linear-After-The-Exponential (LATE)-PCR: primer design criteria for high yields of specific single-stranded DNA and improved real-time detection. *Proc. Natl. Acad. Sci. U. S. A.* **102**, 8609–8614 (2005).
64. Sanchez, J. A., Pierce, K. E., Rice, J. E. & Wangh, L. J. Linear-after-the-exponential (LATE)-PCR: an advanced method of asymmetric PCR and its uses in quantitative real-time analysis. *Proc. Natl. Acad. Sci. U. S. A.* **101**, 1933–1938 (2004).
65. Stergiopoulos, I. & Gordon, T. R. Cryptic fungal infections: the hidden agenda of plant pathogens. *Front. Plant Sci.* **5**, 506 (2014).
66. St Clair, D. A. Quantitative disease resistance and quantitative resistance Loci in breeding. *Annu. Rev. Phytopathol.* **48**, 247–268 (2010).
67. Rueden, C. T. *et al.* ImageJ2: ImageJ for the next generation of scientific image data. *BMC Bioinformatics* **18**, 529 (2017).
68. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source tool for metagenomics. *PeerJ* **4**, e2584 (2016).
69. Cole, J. R. *et al.* Ribosomal Database Project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res.* **42**, D633–42 (2014).
70. Nilsson, R. H. *et al.* The UNITE database for molecular identification of fungi: handling dark taxa and parallel taxonomic classifications. *Nucleic Acids Res.* **47**, D259–D264 (2019).
71. R Core Team. R: A Language and Environment for Statistical Computing. (2019).
72. Wickham, H. ggplot2: Elegant Graphics for Data Analysis. (2016).
73. Wickham, H. Reshaping Data with the reshape Package. *Journal of Statistical Software* vol. 21 1–20 (2007).
74. Daniels, M. J., Barber, C. E., Turner, P. C., Cleary, W. G. & Sawczyc, M. K. Isolation of Mutants of Xanthomonas campestris pv. campestris Showing Altered Pathogenicity. *Microbiology* **130**, 2447–2455 (1984).
75. Dieterich, C. *et al.* The Pristionchus pacificus genome provides a unique perspective on nematode lifestyle and parasitism. *Nat. Genet.* **40**, 1193–1198 (2008).