www.sciencemag.org/cgi/content/full/333/6038/58/DC1

Supporting Online Material for

Probing Individual Environmental Bacteria for Viruses by Using Microfluidic Digital PCR

Arbel D. Tadmor,* Elizabeth A. Ottesen, Jared R. Leadbetter, Rob Phillips*

*To whom correspondence should be addressed. E-mail: arbel@caltech.edu (A.D.T.); phillips@pboc.caltech.edu (R.P.)

Published 1 July 2011, *Science* **333**, 58 (2011)
DOI: 10.1126/science.1200758

This PDF file includes:

Materials and Methods
SOM Text
Figs. S1 to S8
Tables S1 to S13
References

Supplemental Online Materials

MATERIALS AND METHODS

Termite collection

Reticulitermes hesperus specimens were collected from Chilao Flats Campground in the Angeles National Forest (Table S3). Throughout the experiment, starting in the field, different colonies were kept in separate tip boxes and never came in contact with each other. Colonies thereafter were maintained in the laboratory (*12*). Microfluidic array experiments were carried out days to weeks (< 4 weeks) thereafter.

PCR on the microfluidic array

Microfluidic array multiplex PCR reactions contained Perfecta multiplex qPCR master mix (Quanta Biosciences), 0.1% Tween 20 (Sigma Aldrich Incorporated), 100nM ROX (Quanta Biosciences). Universal 16S SSU rRNA primers and probes used were (*12*): forward 357F 5'-CTCCTACGGGAGGCAGCAG-3' (300nM), reverse 1492RL2D 5'-TACGGYTACCTTGTTACGACTT-3' (300nM), 1389 probe HEX-GTGCCAGCMGCCGCGGTAA-BHQ1 HPLC purified (300nM). Unprobed terminase primers used were: forward ter7F 5'-CATTTGATTTGCCGTTACCGIGCYAARGAYGC-3' (200nM) and reverse ter5eR 5'-CICCWCCAGCCGGATCRCARTAMAC-3' (100nM). The probed terminase reverse primer used was: ter5eR.L 5'-CAGCCACACICCWCCAGCCGGATCRCARTAMAC-3' (100nM). The universal probe used for the terminase primer set was: Roche Universal Probe #5 (250 nM). The primers and the rRNA probe were ordered from Integrated DNA Technologies and resuspended in sterile TE buffer (10 mM Tris-HCl, 1 mM EDTA, pH 8) filtered with a 0.02 µm sterile Anotop syringe filter (Whatman). Primers and probes were diluted in DEPC-treated sterile filtered water (Sigma) and then sterile filtered again (prior to dilution) with a 0.02 µm syringe filter.

Preparation of termite hindguts

In each experiment three *Reticulitermes hesperus* worker termites from the same colony (and same tip box) were incubated for several minutes at 4°C to immobilize the specimens and whole guts were subsequently extracted using sterilized forceps on a disposable sterile petri dish. Guts were resuspended in 897 µL of 4°C "synthetic gut fluid" (SGF) salt solution (*42*) pre-filtered with a 0.02 µm sterile syringe filter containing 0.5 µg/mL final concentration of DNase free RNase (Roche) to prevent inhibition by ribosomal RNA. Guts were repeatedly disrupted with a sterile 1 ml filter pipette tip and suspensions were briefly vortexed and allowed to settle for 30 seconds to sediment large particles. Samples were then diluted to working concentrations using the SGF diluent. For microfluidic arrays C through G the resuspended gut fluid was further filtered with an Acrodisc 5 µm sterile syringe filter (Pall Life Sciences) to remove inhibiting large particles such as wood fragments and protists. Samples were then mixed 1:10 with the PCR reaction mix (above) for immediate loading onto the primed microfluidic array once the dilutions were completed. Termite bodies were frozen for later analysis of their COII sequences (see below).

Microfluidic array thermocycling and fluorescence analysis

BioMark 12.765P peelable microfluidic arrays from Fluidigm were loaded with the samples described above and PCR was performed using the BioMark system (Fluidigm Corporation) as recommended by Fluidigm. The cycling protocol was 95ºC 5 min, (95ºC 15 s, 60ºC 90 s)×45, 10 min at 60ºC, 20ºC 10 sec. Amplification curves were evaluated using BioMark Digital PCR analysis software (Fluidigm, v.2.0.6) applying ROX normalization and a linear baseline correction. FAM fluorescence threshold was set to detect any increase in fluorescence, while the HEX threshold was set above the fluorescence leakage of the FAM channel into the HEX channel, detectable in both a no-16S rRNA-primer control panel (dedicated for this purpose) and the no-template-control panel. Both panels were included in every microfluidic array. To minimize diffusion from neighboring chambers after pressure release, only chambers displaying fluorescence in both channels that were flanked by chambers displaying no fluorescence in both channels were selected for retrieval. An example of end-point fluorescence of an array panel is shown in Fig. 1A. In this figure only fluorescence from within chambers is shown, detected based on the reference dye fluorescence measurement. To illustrate the nature of colocalizations, we mask the chambers in such a way that half of each chamber shows one fluorescence channel and the other half shows the other. This way the left half of each chamber showed only the FAM/viral channel fluorescence and the right half of each chamber showed only the HEX/SSU rRNA channel fluorescence. Fluorescence is shown on a logarithmic scale with background subtracted.

Sample retrieval

Microfluidic arrays were peeled shortly after the end of the PCR run and pressure in the arrays was released by depressing the pressure valves. Samples were retrieved into 10µl TE buffer (that was pre-filtered with a 0.02µm sterile Anotop syringe filter) using disposable sterile 30.5G needles (*12*) (one disposable needle per chamber) and subsequently evaluated for the presence of target genes via conventional simplex PCR. In addition, for each array, with the exception of array B, at least five chambers were also retrieved from the no-template-control panel to test for possible cross-contamination (all control retrievals were negative - see below). The PCR reaction mix consisted of perfecta qPCR multiplex master mix with the SSU rRNA primers at 300nM concentration and terminase primers at 200nM concentration. The SSU rRNA probe, the Universal Probe #5 and the probe binding primer ter5eR.L were omitted from these reactions. The cycling protocol for conventional PCR for the simplex terminase reaction was 95ºC 3 min, (95ºC 15 s, 60ºC 60 s, 72ºC 60 s)×40, 72ºC 10 min and for the simplex SSU rRNA reaction was the same but with 32 cycles of amplification to prevent amplification of contaminates associated with the Taq master mix. The presence or absence of product was evaluated using agarose gel electrophoresis. Samples that displayed a band at the expected fragment size for both simplex reactions were deemed successful.

The majority of successful retrievals from the microfluidic arrays were amplified for cloning and/or sequencing in two 30 μL reactions using 3.5 U of EXPAND high fidelity polymerase (Roche), Fail-Safe PCR PreMix D (Epicentre), and primers and cycling conditions as above. In the case of microfluidic array A, terminase sequences were amplified with Perfecta qPCR multiplex master mix instead. For each reaction 1.5 μL of retrieved sample was used. PCR products were purified using the Qiagen PCR purification kit, and sequenced using the terminase ter7F and ter5eR primers and SSU rRNA gene internal primers 1100R (3'-AGGGTTGCGCTCGTTG-5') and 533F (3'-GTGCCAGCMGCCGCGGTAA-5'). Sequencing reactions of microfluidic array amplicons were carried out by the USC DNA core facility (Los Angeles, CA) using an annealing temperature of 50 or 55ºC.

Sequences that contained a mixture of SSU rRNA sequences were discarded from further analysis. Sequences that contained a mixture of terminase sequences, or in which the trace quality was poor were cloned for sequencing using the TOPO TA cloning kit (Invitrogen). At least eight colonies from each cloning reaction were picked and used as templates for PCR reactions. PCR reaction mix included Fail-Safe PCR PreMix H (Epicentre), Taq polymerase (New England Biolabs) and standard T3/T7 primers at 250 nM. Cycling conditions were 95ºC 3 min, (95ºC 15 s, 55ºC 30 s, 72ºC 60 s)×35, 72ºC 10 min. Sequences with different restriction fragment length polymorphism (RFLP) patterns were chosen for sequencing. For the RFLP analysis, 6 μl of each reaction was digested at 37ºC for 4 hr with 3 units HinPI1 from New England Biolabs followed by an inactivation step at 65ºC for 20 min. A representative of each RFLP type (with the correct product band) was sequenced with the high fidelity polymerase and standard T3 and T7 primers. PCR products were purified using the Qiagen PCR purification kit and sequenced with standard T3/T7 primers. Sequencing reactions for cloning were carried out by Laragen Inc. (Los Angeles, CA).

Identification of termite species

The mitochondrial cytochrome oxidase II (COII) gene was used to identify the termite specimens analyzed in this study (*43,44,22*). For each of the three colonies that were collected, either heads or bodies of three to five worker termites frozen on the day of the microfluidic array experiments were used as a template for a PCR amplification of the COII gene. Primers used were A-tLeu (5'-ATGGCAGATTAGTGCAATGG -3') and B-tLys (5'-GTTTAAGAGACCAGTACTTG-3')(*45-46*). For colonies 1 and 2 the PCR product was cloned and sequenced. For colony 3 the product was directly sequenced. Colonies 1, 2, and 3 shared 99.3% nt identity with 0 gaps (0.003% SD; *n*=3 over 680 unambiguous nt) and 100% amino acid identity (over 226 residues) with the COII sequence of *Reticulitermes hesperus* isolate LBL2 (accession number AY623445.1).

Sequence analysis

Sequence traces were converted into a nucleotide sequence using Lasergene SeqMan Pro v8.1.2. Representatives of the SSU rRNA nucleotide sequence of hosts I through IV were then screened for chimeras using Pintail (*47*) and Bellerophon (*48*), the latter implemented in Greengenes (*49*), returning negative results. All terminase sequences from all 41 colocalizations were also tested for amplification related chimeras using Bellerophon (*48*). Cases where both chimera parents belonged to the same PCR batch (E2iii) were eliminated from further analysis.

SSU rRNA sequences were aligned by SILVA (*50*) incremental aligner SINA and subsequently analyzed in ARB (*51*) version 07.12.07org using SILVA release 100 (SSURef_100_SILVA_02_08_09_opt). jModelTest 0.1.1 (*52-53*) was used to find the optimal nucleotide substitution model for the rRNA sequences in Fig. 2 testing 40 different models on an alignment of 898 unambiguous nucleotides without gaps, estimating a maximum likelihood (ML) tree for each model. The optimal nucleotide substitution model (based on the AICc criterion with sample size set to the number of sites in the alignment) was a Tamura-Nei model (*54*) +I+Γ with unequal base frequencies. A maximum likelihood tree was then computed for this alignment with PhyML 2.4.5 (*53*) implemented in ARB using the Tamura-Nei model +I+Γ (nCat=4), with all parameters estimated from the data and with 1000 non-parametric bootstrap iterations. Other treeing methods such as Phylip DNAPARS v3.6a3 (*55*) and Fitch-Margoliash (*56*) distance method implemented in ARB predicted very similar topologies (Fig. 2). In Fig. 2 solid circles represent significant nodes supported by ML, parsimony (Phylip DNAPARS v3.6a3 (*55*)), and distance (Fitch-Margoliash (*56*)) methods. Half circles represent nodes supported by ML and either parsimony or distance methods. Open circles represent nodes supported by only ML. In addition, support values greater than 50% for 1000 bootstrap iterations are shown. We note that the topological relation between phage host clades I–IV appeared to be sensitive to the addition of other *Treponema* sequences from public databases, and to the particular outgroup chosen as well, and therefore the topology in Fig. 2, though robust, may not be definitive.

Nucleotide sequences of the large terminase subunit gene present in *R. hesperus*, Z. *angusticollis* and *Nasutitermes* sp. termites were translated in reading frame and aligned with ClustalW (*57*) in MEGA4 (*58*) (the alignment used in the analysis was straightforward and involved a single insertion event of a highly conserved five amino acid sequence in some of the sequences). Subsequently 705 unambiguous aligned nucleotides without gaps were tested for the presence of recombination with RDP3 v3.44 (*59*). Methods used to scan for recombinant sequences included Geneconv (*60*), Maxchi (*61*), and RDP (*62*) (as recommended in the RDP3 manual and shown to be the preferable tests for non-redundant sequences (*63-64*)) as well as the Bootscan method (*65*). Since each recombination detection method individually is error prone (*63-64, 66*) several methods are required to explore recombination (*63, 66*). Similar sequences (≤3.3%) were removed prior to analysis as recommended in the RDP3 manual. The first two events found by RDP3 implicated by all four methods alleles A13ii and B1 as recombinants, confirmed by manual phylogenetic inspection in RDP3. A NeighborNet analysis with SplitsTree4 (*67*) using optimal substitution parameters estimated by FindModel (*68*) confirmed the reticulate nature of these alleles and consequently these alleles were excluded from the phylogenetic tree in Fig. 2 (see Fig. S5). The following two events detected by RDP3 (H5, B2) were only supported by Maxchi, however the NeighborNet network showed these putative recombinants were also associated with significant reticulate patterns, which were eliminated upon removal of these sequences. Consequently these two samples were also excluded from the phylogenetic tree. The remaining events detected by RDP3 with lower confidence exhibited either a small degree of local reticulate patterns or no reticulate patterns and were therefore kept in the analysis. Eliminating potential recombinant alleles resulted in a largely tree-like network suitable for phylogenetic analysis (Fig. S5B). A likelihood-mapping analysis (*69-70*) with TREE-PUZZLE 5.0 using 10000 quartets and the optimal model found by jModelTest (see below) showed that 95.7% percent of the quartets fell in the triangle corners (A_1,A_2,A_3) suggesting that a phylogenetic tree should fit the data (*70*).

After recombinant sequences were removed, jModelTest was used to find the optimal nucleotide substitution model testing 40 different models, estimating a ML tree for each model. The optimal model (based on the AICc criterion as described above) was a Tamura-Nei model (*54*) +I+Γ with the base frequencies having little effect on the AICc score. A ML tree was then computed with PhyML 2.4.5 implemented in ARB using the Tamura-Nei model with +I+Γ (nCat=4), with all parameters estimated from the data and with 1000 non-parametric bootstrap iterations. Other treeing methods such as DNAPARS v3.6a3 and Fitch-Margoliash distance method implemented in ARB predicted very similar topologies (Fig. 2). Tree topology was also similar to the ML estimated tree topology of the corresponding 235 amino acid residues, with the main differences being a slight repositioning of the higher termite clade and sequence A2. Since the terminase gene is comprised of two functional domains, an ATPase domain and a nuclease domain (see Fig. S3), we also compared the ML estimated topology of 495 unambiguous aligned nucleotides of the N-terminal domain of the gene (see Fig. S2 for alignment) with the nucleotide tree of the entire gene and found the topologies to be nearly identical. p-distances were measured in MEGA4 and standard deviations were calculated in Matlab.

Survey of SSU rRNA ribotypes on the microfluidic array

In order to assess the frequency of putative host ribotypes I through IV on the microfluidic array as well as the frequency of other rRNA ribotypes, we constructed a library of 118 randomly sampled rRNA hits from the microfluidic arrays. To this end, for two microfluidic arrays (F and G) and for every panel on these arrays (except the two control panels), 10 chambers for which the HEX (rRNA) fluorescence exceeded the detection threshold (irrespective of florescence in the FAM/terminase channel) were randomly selected for retrieval. The identities of the chambers for retrieval were obtained by a random number generator implemented in Matlab 7.4. These sequences were then post-amplified for sequencing using Perfecta multiplex qPCR master mix (Quanta Biosciences) as described in the Methods section. Sequencing was performed by the USC DNA core facility using internal SSU rRNA primers 533F and 1100R (see Methods). A total of 118 sequences were successfully sequenced and assembled using Lasergene SeqMan Pro v8.1.2. In Fig. 3 we plot the rank abundance curve of just *Treponema* phylotypes from the reference library. The frequency of each phylotype is given in Table S5. Each column in Fig. 3 can be thought of as a random variable sampled from a binomial distribution with mean $n \cdot p$ and standard deviation $SD = \sqrt{n \cdot p \cdot (1-p)}$, where p is the probability to sample this phylotype and n is the total number of trials (here n=78 trials). The error bars in Fig. 3 are ±SD, with p estimated for each phylotype as the number of occurrences of that phylotype divided by n.

Degenerate primer design and testing

Terminase phage primers were designed to target several conserved regions of the large terminase subunit gene found in the four prophage-like elements in *Treponema primitia* (ZAS-2) (*23*) and *Treponema azotonutricium* (ZAS-9) (*24*), and in 46 contigs found in the metagenome of a *Nasutitermes* species termite (*22*). The primers were designed with CODEHOP (*17*), selecting candidates with melting temperatures matching the all-bacterial SSU rRNA primer set (primer candidates were required to be different by at least five base pairs to be considered different candidates). The primer sequences in both the degenerate core region and the clamp region were manually tweaked to offer the best coverage for the conserved region (matching the codon bias in these sequences) and to minimize primer dimers. In addition, inosines were incorporated at certain positions instead of mixed bases to reduce primer degeneracy. Several forward and reverse primer candidates were chosen and the nucleotide regions were further adjusted to minimize forward/reverse primer-dimers and dimers with the all-bacterial primers and probe. Multiplex PCRs for various forward and reverse primers were performed on a dilution series of purified genomic DNA from ZAS-2 and ZAS-9. PCR products were analyzed by agarose gel electrophoresis and primers yielding the strongest bands and having the lowest detection limit (<100 copies) were selected. The chosen primers were further screened on genomic DNA extracted from *Zootermopsis nevadensis* by agarose gel electrophoresis.

To allow us to do quantitative PCR (qPCR) with these primers without having to design a degenerate probe we implemented a universal-template probe strategy first suggested by Zhang *et al.* (*71*) and adapted for degenerate primers by Ottesen *et al.* (*42*). In this method a short universal nondegenerate probe sequence is attached to the 5' end of the forward and/or reverse primers. The probe-binding sequence is incorporated into the amplicon during the first round of amplification, allowing the probe to detect amplification of that product. A short nondegenerate 8 base probe incorporating locked nucleic acids (LNAs) then binds to the probe-binding sequence and is subsequently cleaved by the DNA polymerase like in a standard TaqMan chemistry. The locked nucleic acids increase the melting temperature of the probe allowing usage of a very short probe. A probe yielding the minimal interaction with the SSU rRNA amplicon and other oligos in the master mix was chosen for this task. A linker sequence was incorporated between the probe-binding sequence and the degenerate primer to further reduce dimers.

Multiplex qPCR standard curves were obtained for all probe binding sequence combinations (probe binding sequence on the forward primer, probe binding sequence on the reverse primer and probe binding sequence on both the forward and the reverse primers) and for all the candidate primer sets. In all cases, primers with LNA probe binding sequences were mixed 50% with primers lacking the probe binding sequence as this seemed to enhance the PCR reaction. Primer sets yielding the best standard curves, highest end-point amplification for positive templates and highest Cts for the no-template-controls were selected. Primer sequences for the best candidates were fine tuned to further reduce dimers and then screened again using the same metric described above. The best candidates were then tested on ZAS DNA on the digital PCR microfluidic array. Primers yielding the best amplification curves, highest end-point amplification, and lowest number of no-template-control hits were selected. Finally, primer and probe concentrations were optimized on the microfluidic array for the chosen primer set. All benchtop qPCRs were performed on a Stratagene Mx3000P. Cycling conditions were as described in the Methods section.

Measures to prevent and test for contamination

To prevent contamination from the environment, from termites and from post-PCR products, several precautions were taken. Experiments were conducted in five different laboratories that were physically separated (different laboratories within the same building or different buildings). All PCR master mixes for dPCR runs, PCR master mixes for post-amplification of retrieved microfluidic array samples, and tubes loaded with 10 µl TE buffer for retrieved sample resuspension were prepared in laboratory #1 that never came in contact with termites or related samples thereafter. In addition, pipettes and benches were always thoroughly cleaned with EtOH or EtOH and bleach prior to setup. Termite handling and microfluidic array loading were conducted in laboratory #2, where each of these two procedures took place in well-separated designated areas. Sample retrieval was performed in a separate room within laboratory #2 using disposable syringes. Sample loading for post-amplification was performed in laboratory #3. Master mixes for cloning-related PCR reactions were prepared in laboratory #3 (which was designated as a PCR cloning "clean area") and loading of samples for cloning-related PCR was performed in laboratory #4. All subsequent manipulations of samples or cloned PCR products (such as RFLP analysis, agarose gel electrophoresis, PCR purification, etc.) were performed in laboratory #5.

To test that no contamination occurred, every microfluidic array contained a no-template control panel and for each array (except B) at least five chambers from the no-template-control panel on the array were retrieved and processed with the rest of the samples to insure there was no cross-contamination during the retrieval process. No-template-control chambers retrieved for this purpose were selected such that these chambers and their flanking chambers on either of their sides did not exhibit fluorescence in both the FAM and HEX channels (this was done to prevent possible diffusion of targets from adjacent chambers into the sampled chamber after pressure release). All no-template-control samples that were retrieved from the microfluidic arrays were post-amplified with the rest of the retrievals and tested by agarose gel electrophoresis. All negative controls were always negative for both channels (*). Background amplification in the no-template-control-panels never exceeded 2.6% of positive chambers for both channels (1.25 ± 0.75% SD for the terminase channel and 1.35 ± 0.7% SD for the SSU rRNA channel). Some background amplification using all-bacterial SSU rRNA primers is expected (*12*) and is commonly attributed to DNA fragments present in commercial enzyme preparations (*72*). The positive hits for the FAM channel in the microfluidic panels are expected to be a consequence of the modified TaqMan chemistry employed: since the universal LNA probe can spuriously bind to a terminase primer, primer-dimers will lead to amplification of a spurious product and fluorescence (similar to primer-dimers observed in SYBR Green assays), however no actual contaminating target is present, verified by agarose gel electrophoresis (see Fig. S7, Table S11, and supporting text for further discussion). Finally, every post-array amplification was always executed with several no-template-controls.

(*) One of the five SSU rRNA control chambers in array G was positive in a diagnostic post-amplification (not for sequencing), however this turned out to be an artifact of the diagnostic run as post-amplification of the same sample a second time was negative (with the positive control being positive).

Measurement of PCR and cloning error rates

To measure the sequence error rate of samples retrieved from the microfluidic dPCR array, genomic DNA from ZAS-9 was used as a reference template in a microfluidic dPCR array. Vortexed genomic DNA from ZAS-9 was loaded onto a microfluidic dPCR array and cycled as described in the Methods section. Samples were then retrieved and the rRNA and terminase gene fragments were post-amplified using EXPAND high fidelity polymerase (Roche) as described in the Methods section. To measure the error rate, sequenced array retrievals were aligned against the known sequence of ZAS-9 rRNA and terminase genes. The error rate of the rRNA gene was 0 with 0 gaps (*n*=8, 905 ± 20bp SD) and the error rate of the terminase gene was 0 with 0 gaps (*n*=16, 711 ± 14bp SD). Post-amplification of the terminase gene fragment with the Quanta master mix resulted in a small number of ambiguous bases, however correcting these artifacts resulted in perfect matches. To test cloning associated errors, a retrieved ZAS-9 terminase sequence post-amplified with Roche high fidelity polymerase was cloned and several colonies were picked, amplified with the Roche high fidelity polymerase and sent for sequencing, as described in the Methods section. The measured error rate was 0.59 ± 0.29% SD (*n*=9, 759 ± 4bp SD) with 1 gap for 1 out of 9 cases. A similar cloning error rate was found when comparing the nucleotide sequences of 12 terminase amplicons in Fig. 2 sequenced directly from retrieved samples with their corresponding TOPO clones (0.55% ± 0.32% SD, *n*=12). In some cases single nucleotide deletions were also observed (see below). To check that clone errors were not sequencing related, five samples of the same terminase clone were amplified and sent for sequencing, however all sequences were found to be identical. To check that these errors are not introduced by *E. coli* during the growth phase, a single terminase colony was re-streaked and five colonies were amplified and sent for sequencing. All colonies yielded 100% identical sequences. Consequently, the origin of the terminase sequence errors appears to be the cloning step.

Out of 31 terminase sequences in Fig. 2, 10 were sequenced from the original retrieval, 12 were sequenced from a combination of the original retrieval and a TOPO clone, and 9 were sequenced from the TOPO clone alone. When sequences from the original retrieval were available and unambiguous, to minimize cloning errors these sequences were used in the consensus sequence in overlapping regions. Therefore for these sequences the error rate is expected to be lower. TOPO clones A9ii and E2i initially contained a frame shift mutation and E2i contained in addition an errant stop codon. These mutations were suspected to be cloning-related errors, confirmed by sequencing additional TOPO clones for each sample and calling base pairs by majority consensus. TOPO clone A11 also contained a frame shift mutation outside the alignment region considered in Fig. 2. This frame shift mutation also appears to be a cloning artifact as similar (though not identical) clones from the same retrieval did not contain this frame shift mutation. Consequently an N was inserted at this position. In the absence of TOPO clones, if an ambiguous base was declared (one such case) the degeneracy was arbitrarily broken to facilitate translation.

Measurement of primer efficiency

To measure SSU rRNA primer efficiency, five panels of a microfluidic dPCR array were loaded with ZAS-9 genomic DNA. Genomic DNA was titrated to achieve a final expected number of 400 (*n*=1), 300 (*n*=2), and 200 (*n*=2) SSU rRNA targets that were uniformly distributed across a panel containing 765 microfluidic chambers. Expected number of targets was estimated based on genomic DNA concentration measured using a Hoefer DynaQuant 200 fluorimeter. Digital PCR chemistry and cycling conditions were as described in the Methods section. The genomic DNA was vortexed upon extraction and therefore the genome is expected to be sheared to 10–20kb fragments. Since the two copies of the rRNA and terminase genes were located 689 kbs and 939 kbp apart, respectively, each genome was assumed to contribute two separate copies of each gene. After subtraction of noise, estimated from the no-template-control panels, the average rRNA and terminase primer efficiencies were calculated to be 59 ± 6% SD (*n*=5) and 74 ± 7% SD (*n*=5).

Selection pressure analysis

The program HyPhy 2.0 (*73*) was used to estimate the relative rate of non-synonymous (β) and synonymous (α) substitutions (ω=β/α) for all 28 retrievals associated with hosts I through IV using a maximum likelihood approach with a codon substitution model (*74*). An alignment comprising 705 unambiguous nucleotides without gaps was used to generate a maximum likelihood (ML) tree with phyml assuming a TN93 (*54*) nucleotide substitution model +Γ(nCat=4)+I+F. Given the above alignment and ML tree, HyPhy was used to find an optimal nucleotide substitution model out of all possible time-reversible models using the AIC criterion for selection. Finally, HyPhy was used to obtain the ML estimates of the independent model parameters of an MG94(*75*)xREV_3X4(*74*) substitution model with the optimal constraints found above (012032) assuming global parameters, the above ML tree, and the above in-frame alignment. Equilibrium frequencies were estimated from the partition. The global estimated ω was found to be 0.079. The 95% profile likelihood confidence interval was 0.071 to 0.088. This range is significantly lower than ω=1 (the case of neutral evolution) indicating that the terminase gene is under substantial negative selection pressure. A likelihood ratio test (LRT) comparing the null hypothesis model (ω=1) to the above alternative model strongly rejects the null hypothesis of neutral evolution with LR=754 and a P value (likelihood ratio test) predicted by HyPhy to be 0. In Table S6 the selection pressure was estimated for individual bacterial hosts using several additional methods and resulted in the same conclusion.

Analysis of viral genes in the metagenome

We were interested in finding the more abundant viral genes in the metagenome to identify a viral marker gene for this environment. In order to make this method widely accessible we designed an automated tool called MetaCAT that screens all gene objects in a metagenome and clusters them based on homology to genes in a reference database of known viral genes. The number of metagenome gene objects in a given cluster is then interpreted as the relative frequency of the corresponding known viral reference gene in the metagenome. This method is capable of assessing the relative frequency of viral-related metagenome gene objects in an annotation independent way. We refer to the implementation of this algorithm as the Metagenome Cluster Analysis Tool (MetaCAT), available upon request.

The MetaCAT algorithm is as follows: we first BLAST a list of known (viral) reference genes against all metagenome gene objects using BLAST v2.2.22+ (*76*) (wrapped by Matlab) with a cutoff E value of 10^{-3}. As a reference list of known viral genes we use NCBI's viral RefSeq database v37 (*32*). The number of metagenome gene objects homologous to each of the known reference genes is defined to be the *abundance* of that known reference gene in the metagenome. Since the list of known reference genes is long (~80,000 genes) we wished to filter this list based on several criteria. First, we retain only known reference genes whose best E value score is $\leq 10^{-7}$. This filtering step is performed to retain only known reference genes that yield reasonable alignments to metagenome gene objects. The second filtering step, implemented in Matlab, was designed to take out redundancy in the RefSeq database itself with respect to the metagenome using a dedicated clustering algorithm. For example, if two known reference genes are homologous to similar lists of metagenome gene objects, we would like to report only one of the two known reference genes, choosing the one with the lower E value. More generally, we wish to find for every known reference gene all the other known reference genes to which it is *related* (a known reference gene is always *related* to itself; see definition below). Therefore each known reference gene belongs to a *group* of *related* known reference genes. Finally, for each *group* of *related* known reference genes we only report the known reference gene with the lowest E value to represent that *group*. The combined list of reported known reference genes is then the final list of viral genes. The frequency of each reported viral gene is defined as the *abundance* of that known reference gene in the metagenome (see above). To complete the definitions: two known reference genes are said to be *related* if the *signatures* of both known reference genes is *similar*. A *signature* of a known reference gene is defined as the list of metagenome gene objects to which that known reference gene is homologous (E ≤ 10^{-3}). Two signatures are then said to be *similar* if they share 50% of the elements in their lists. That is, if list A has L_i elements and list B has L_j elements, lists A and B are said to be similar if $50\% \geq 100 \cdot \min\left(L_i \cap L_j / L_i, L_i \cap L_j / L_j\right)$, with the symbol $\cap$ denoting the intersection between the two lists.

Note that the final reported known reference genes can still be *related*. Nevertheless, this filtering step is effective at removing a considerable amount of redundancy in the

RefSeq database. A third manual filtering step is applied to retain only viral genes related to building a virion. Such genes are considered to be virus-specific genes (*29*). Examples of such genes include capsid proteins, portal proteins, terminase proteins, tail proteins, baseplate proteins, and so on (*29*). The list of the most abundant viral genes in the metagenome (*abundance* ≥10) is given in Table S1.

SUPPORTING TEXT

Statistical analysis of colocalization in digital PCR microfluidic arrays

Origin of a random colocalization component

We wish to see if *k* repeated colocalizations of a particular 16S rRNA ribotype with the terminase gene can be explained by chance colocalization on the microfluidic array (referred hereto as a "chip"). The reason there is a finite probability for chance colocalization is that typical array panels usually contain a certain fraction of FAM hits (the channel of the terminase marker) that are not colocalized with HEX hits (the channel of the 16S rRNA marker) as is shown in Fig. S6. If a fraction of these non-colocalized FAM hits contains the terminase target there is finite probability they may colocalize by random chance with a 16S rRNA gene and be mistaken for a true (host/terminase) colocalization. The number of these types of chance events determines the probability for false colocalization. Non-colocalized FAM hits (which do not always contain an actual terminase product) can arise for several reasons:

(1) Since the universal LNA probe binds to a terminase primer, primer-dimers can lead to amplification and spurious fluorescence, i.e., fluorescence in the absence of a terminase target. These types of hits are apparent in the no-template-control panel and can account for roughly half of the non-colocalized hits on a typical panel (see Table S11 and Table S13 discussed below). To verify that FAM hits in the no-template-control panel do not contain a target and are not the result of a contamination, four positive FAM chambers were retrieved from a no-template-control panel, post amplified for the terminase gene and analyzed by agarose gel electrophoresis, however no bands were detected. In addition, for several panels for two chips all FAM hits (both colocalized and non-colocalized) were retrieved, post amplified for the terminase gene and analyzed by agarose gel electrophoresis (Table S11). For each panel there were several samples that did not display any band (see Fig. S7 for a representative example), a finding that is consistent with the presence of spurious products observed in the no-template-control (NTC) panel. Furthermore, the average number of samples that did not display a band agreed well with the number of FAM hits in the no-template-control panels for these chips (Table S11), confirming that there is a noise component of spurious amplification on the panels similar to the no-template-control panel. For the seven chips in this study the average number of FAM hits in the no-template-control panel was 9.6 ± 5.7. These types of non-colocalized FAM hits will not lead to chance colocalization with a 16S rRNA gene since there is no actual terminase target present.

(2) If the end-point fluorescence generated by a 16S rRNA target did not exceed the HEX threshold, this chamber would seemingly appear as a non-colocalized event (even though there is a 16S product present). Since the HEX threshold is set high enough to filter out cross-talk from the FAM channel into the HEX channel, some potential HEX hits may have been omitted. Indeed, when retrieving all FAM hits from a panel and amplifying all retrievals for the 16S rRNA gene, usually some wells whose HEX end point fluorescence did not pass the detection threshold did have a 16S rRNA band (data not shown). These types of non-colocalized FAM hits should not contribute to false colocalization or contribute minimally because samples with mixed/chimera 16S rRNA traces are discarded from analysis and the probability of repeatedly amplifying the same wrong 16S rRNA is negligibly small (see discussion below).

(3) The 16S rRNA qPCR efficiency was measured to be ~60% for ZAS-9 genomic DNA (see Materials and methods). These types of events could potentially lead to false colocalization if a 16S rRNA amplification product is not generated (but the terminase gene in this cell was amplified) and this target colocalized by chance with another bacterial cell whose 16S rRNA gene was amplified. If an amplicon was generated (but for some reason fluorescence was inhibited) then these types of non-colocalized FAM hits will not contribute to false colocalization because samples with mixed 16S rRNA traces are discarded.

(4) Some cells may potentially prematurely lyse and their DNA may get sheared (for example when crushing the gut or during the loading process onto the chip). If this happens there is a possibility that free floating terminase targets are released into the mix.

(5) There may be assembled viruses present or free floating viral DNA, which can be regarded as free floating terminase targets.

As mentioned above, approximately half of the non-colocalized FAM hits on a given panel can be explained by the spurious noise and do not contribute to random colocalization. Of the remaining non-colocalized FAM hits, the fraction relating to (2), if present, will not lead to false colocalization. Therefore the probability for false colocalization estimated below, which is based on fluorescence measurements alone, is an upper bound on the true probability for false colocalization.

<u>Statistical model of random colocalization (P value estimation)</u>

In Fig. 2 we see that certain 16S rRNA ribotypes are repeatedly colocalized, giving rise to 16S rRNA clades I–IV. The null hypothesis is that these 16S rRNA ribotypes are not true hosts and that the observed repeated colocalizations are due to chance associations, that is, these 16S rRNA ribotypes are simply colocalized many times by chance with free floating terminase targets. We therefore wish to estimate the probability (P value) that out of n=41 successful retrievals from the chip, i.e., retrievals that resulted in obtaining a 16S rRNA and terminase sequence after post-amplification, we will retrieve k or more instances of a particular ribotype $\boldsymbol{S}$ colocalized with a terminase (any terminase). This probability is given by

$$\text{Prob}\left(\text{number of chance co-localizations of } \boldsymbol{S} \text{ with a terminase } \geq k \,|\, n \text{ successful retrievals}\right) =$$

$$= \sum_{k'=k}^{n} \text{Prob}\left(\text{number of chance co-localizations of } \boldsymbol{S} \text{ with a terminase } = k' \,|\, n \text{ successful retrievals}\right) =$$

$$= 1 - \sum_{k'=0}^{k-1} \text{Prob}\left(\text{number of chance co-localizations of } \boldsymbol{S} \text{ with a terminase } = k' | \, n \text{ successful retrievals}\right) =$$

$$= 1 - \sum_{k'=0}^{k-1} \text{Prob}\left(\text{succeed } k' \text{ times with probability } p_F | \, n \text{ trials}\right) =$$

$$= 1 - \sum_{k'=0}^{k-1} \binom{n}{k'} p_F^{k'} \left(1 - p_F\right)^{n-k'} = 1 - \mathbf{binocdf}\,(k-1, n, p_F)$$

where **binocdf** is the cumulative distribution function of the binomial distribution and p_F is the probability that when we successfully retrieve a colocalized well from a panel it contains the particular ribotype ***S*** and any terminase gene by pure chance. Given *k*, *n* and p_F (estimated below) the P value can be calculated. We find that the P values (*n*=41; one-tailed) for hosts I–IV are all highly statistically significant (P < 10^{-4}; see Table 1 and Table S13) allowing us to reject the null hypothesis.

A model for a typical panel

Each panel loaded with a template is assumed to have the following species: *Y* HEX hits ("blue" hits), *X* FAM hits ("red" hits), out of which "*noise"* FAM hits are due to spurious amplification (no actual target). We assume that out of the *X* FAM hits there is a fraction of FAM hits that are free floating targets, that is a DNA fragment coding for a terminase gene but not for a 16S rRNA gene. The number of free floating targets is defined to be $X_T - noise$. These free floating targets would be the source of false colocalizations events. Thus colocalization events observed on the chip can be due to three possible causes: **(1)** genuine colocalization of a host SSU rRNA with its terminase, **(2)** chance colocalization of a free floating terminase gene with a 16S rRNA gene, **(3)** chance colocalization of a spurious FAM amplification (no actual terminase amplicon present) with an rRNA gene. See Table S12 for a definition of all the variables used in the model.

Estimation of p_F

To calculate the P value above, one must estimate p_F, i.e., the probability that a successful retrieval from a panel contains our particular ribotype ***S*** and any terminase gene by pure chance. This probability can be estimated as follows: let X_T be defined as the sum of the total number of free floating terminase targets and spurious targets leading to spurious FAM amplification (i.e., noise). We will see how to estimate X_T later on but for the time being let's assume it is given. The average number of free floating terminase targets to colocalize with a *particular* 16S rRNA ribotype ***S*** on a panel, defined as I_S, is given by multiplying the number of wells on a panel (765) by (a) the probability that a given well will contain a free floating terminase target p_{ter} and (b) the probability that that well will also contain ribotype ***S*.** The probability that a given well will contain a free floating terminase target is

$$\text{(S1)} \qquad p_{ter} = \left(\frac{X_T - noise}{765} \right)$$

where *noise* is the number of FAM hits that are due to spurious amplification and are not associated with an actual terminase target. Thus $X_T - noise$ is the number of free floating terminase targets on the panel. Note that $X_T - noise$ will lead to an upper bound on the number of free floating terminase targets (leading to an upper bound on p_F) since $X_T - noise$ may include wells with a genuine 16S rRNA amplicon that simply did not pass the HEX detection threshold and are thus wrongly labeled as free-floating terminase targets (as described above). The value for *noise* can be estimated from the no-template-control panel for a given chip (see for example Table S11).

The average number of free floating terminase targets to colocalize with a *particular* 16S rRNA ribotype ***S*** on a panel is therefore given by

(S2a) $$I_{\mathrm{S}} = 765 \cdot p_{ter} \cdot \left(\frac{f_{\mathrm{S}} \cdot Y}{765} \right) = p_{ter} \cdot f_{\mathrm{S}} \cdot Y$$

where Y is the total number of HEX hits on a panel, f_S is the frequency of ribotype ***S*** on the chip so that $f_{\mathrm{S}}Y$ is the number of ribotypes ***S*** on a given panel. I_S is an estimate of the number of false colocalizations on a panel. This number is smaller than the number of observed colocalization on the panel, which we designate by I (=number of HEX and FAM intersections on a given panel). The number actual colocalizations on a panel of any 16S rRNA target with any terminase target (i.e., the total pool from which we draw successful retrievals) would be on average

(S2b) $$I_{\text{all 16S-ter}} = I - \frac{noise \cdot Y}{765}$$

taking out random colocalization of spurious FAM hits from I. The probability p_F is therefore given by the ratio of the number of random colocalization on a panel, I_S, and $I_{\text{all 16S-ter}}$, the number of actual colocalizations on the panel (i.e., of any 16S rRNA and any terminase target, both true and false colocalizations). Thus

(S3) $$p_F = \frac{I_{\mathrm{S}}}{I_{\text{all 16S-ter}}} = f_{\mathrm{S}} \cdot \frac{p_{ter} \cdot Y}{I_{\text{all 16S-ter}}}.$$

Since $p_{ter} \cdot Y / I_{\text{all 16S-ter}}$ can vary somewhat from panel to panel, to calculate p_F we use Bayes' theorem:

$$p_F = P(\text{false} \mid \text{panel A}) P(\text{panel A}) + P(\text{false} \mid \text{panel B}) P(\text{panel B}) + \ldots \ .$$

We therefore replace $p_{ter} \cdot Y / I_{\text{all 16S-ter}}$ in Eq. S3 by its panel averaged value, weighted by the number of times each panel was sampled (making at total of n=41 trials). The estimated values of p_F per host type are given in Table S13.

Estimation of X_T

Let us assume that a given panel has X FAM hits, Y HEX hits, and I intersections. The number of non-colocalized terminase hits is then $X_f = X - I$. X_T is slightly larger than X_f since some of the free floating targets or spurious targets may have colocalized with HEX hits. This difference ($X_T - X_f$) is estimated by multiplying the number of wells on a panel by (a) the probability that a well will contain a free floating target *or* a spurious target and (b) the probability that that well will contain *any* HEX hit. Thus $X_T - X_f = (765 \text{ wells}) \left(\frac{X_T}{765} \right) \left(\frac{Y}{765} \right)$, or

$$X_T = X_f + (765 \text{ wells}) \left(\frac{X_T}{765} \right) \left(\frac{Y}{765} \right).$$

Solving for X_T we find that

(S4) $$X_T = X_f \left(1 - \frac{Y}{765}\right)^{-1} = \left(X - I\right)\left(1 - \frac{Y}{765}\right)^{-1}.$$

Note that since typically Y~50, $X_T \approx X - I$.

Estimation of f_s

f_s, the frequency of ribotypes ***S*** on the chip, is estimated based on the number of the particular REP ribotypes that grouped with the corresponding host ***S*** (e.g., five REP4 ribotypes out of 118 grouped with host I in Fig. S4, therefore f_s=5/118). Operational taxonomical units for REP/host clades were determined by a DOTUR analysis (Table S5 and Fig. S4).

Given f_S and X_T (Eq. S4) we can calculate p_F (Eq. S3), and given k (Table 1) we can calculate the P value. Table S13 summarizes the frequencies f_S, probabilities p_F and P values for hosts I though IV. As mentioned in the beginning of this section, the P values calculated for hosts I through IV were very small ($P < 10^{-4}$) allowing us to reject the null hypothesis, i.e., the repeated ribotypes I–IV cannot be explained by random colocalization of these ribotypes with free floating terminase targets.

Bound on false colocalization in the dataset

We would like to estimate the average number of retrievals where one of the observed hosts colocalized by chance with a terminase (resulting in either two terminases—the host's and the free floating terminase, or, in the case the host's terminase did not amplify or was not present, one wrong terminase). The probability that we retrieve from a given panel any of the host ribotypes with the wrong terminase is given by summing the individual false colocalization probabilities for each host:

$$p_{F,tot} = \sum_{\text{host I-IV}} p_F = \left(p_{ter} \cdot \sum_{\text{host I-IV}} f_{\mathrm{S}} \cdot Y \right) \Big/ I_{\text{all 16S-ter}} .$$

The average number of false colocalizations in a dataset of n=41 retrievals would therefore be

(S5) $$N_{false} = p_{F,tot} \cdot n.$$

We find that N_{false} =0.6. Thus out of 28 repeated colocalizations of our hosts, on average ~0.6 are expected to be false (an error of 2%). The fact that no colocalized pairs were retrieved with the most abundant phylotypes on the array (see Table S5 and Fig. S4) and that the three most abundant phylotypes on the array comprising 49% of all treponemes in only one out of 38 cases colocalized with an rRNA gene (see discussion on non-hosts below) confirms that erroneous colocalization was indeed very rare.

Numerical simulation to test the statistical model

To check our statistical analysis (Eq. S1–S7) we conducted a Monte Carlo simulation of retrieval from the microfluidic panels based on the model presented above (Fig. S8). The numerical simulation results were predicted precisely by the statistical model described above.

Model for Monte Carlo simulation

In the simulation Y rRNA templates were loaded randomly onto a panel of 765 chambers ($Y \sim U[Y_{min}, Y_{max}]$). Each panel was also randomly loaded with *noise* spurious FAM hits ($noise \sim U[noise_{min}, noise_{max}]$) and *free* free floating terminase targets ($free \sim U[free_{min}, free_{max}]$). A fraction f (i.e., probability) of the Y rRNA templates was assumed to be genuine hosts (i.e., hosts that genuinely harbor a terminase gene). The terminase gene within these hosts was assumed to be amplified with probability e_{ter}. Each retrieval trial consisted of loading a single panel of 765 chambers with the above elements and retrieving one sample that contained both a 16S rRNA sequence and a terminase sequence. If the retrieval failed (i.e., the rRNA was colocalized with a spurious FAM target) a new retrieval trial would be attempted until successful (these mute trials would not be counted as successful iterations). For each successful retrieval trial it was registered if the retrieval was a false colocalization (i.e., a host 16S rRNA sequence was colocalized with a free floating terminase). In addition for each successful retrieval trial the probability of false colocalization p_F was calculated. This probability is given by the ratio of number of false colocalizations on the panel (i.e., a 16S rRNA gene that colocalized with a free-floating terminase) and the total number of colocalization on the panel (any 16S and any terminase gene). A single Monte Carlo iteration ended when $N_{retrievals}$ (=41) successful retrievals were obtained. At the end of each Monte Carlo iteration, the total number of false colocalizations (N_{false}) was tallied and the average value for p_F was calculated. In total there were 1000 Monte Carlo iterations.

To compare with the statistical model above, after each Monte Carlo iteration, p_F and N_{false} were estimated based on Eq. S3 and Eq. S5 assuming $f = f_S$ and given the random values for X, Y, I, and *noise* generated for each of the 41 panels in the simulation. At the end of the simulation the average value of p_F and N_{false} (averaged over 1000 iterations) was compared to the predicted values of p_F and N_{false} based on the statistical analysis.

Simulation parameters

Simulation parameters were chosen to mimic the experiments in this study as closely as possible: $N_{retrievals} = 41$; all hosts were assumed to be indistinguishable so that f_S was given as the sum of all the rates f_S in Table S13 (i.e., $f_S = 9/118$, where 9 is the total number of occurrences of hosts I–IV phylotypes in the reference library, and 118 is the size of the reference library—see Table 1). All other parameters followed the distributions in Table S13 with $Y \sim U(20,80)$, $noise \sim U(5,15)$, $free \sim U(0,20)$, and $e_{ter} = 0.74$ (see Materials and methods).

Simulation results

We found that the predictions for p_F (Eq. S3) and N_{false} (Eq. S5) closely matched the numerical simulation:

$$\begin{cases} p_F\text{(simulation)=0.014} \pm 0.011 \\ \hat{p}_F\text{(Eq. S3)=0.018} \pm 0.022 \end{cases} \quad \begin{cases} N_{false}\text{(simulation)=0.6} \pm 0.8 \\ \hat{N}_{false}\text{(Eq. S5)=0.7} \pm 0.9 \end{cases} .$$

The errors are standard deviations. The simulation presented here shows that the statistical model presented above (Eq. S1–S7) is consistent with the numerical simulations.

Chambers with multiple cells

Since the average number of targets loaded per panel was small (~50), the chance of obtaining multiple cells in a given chamber was small (1.7 chambers out of 50 on average (*15*)). However cells can also potentially "stick" together upon loading as well. If a chamber contains multiple 16S rRNA genes and more than one gene is amplified then the sequence trace will be mixed. Such samples were automatically discarded in this study. If a 16S rRNA chimera is formed, chimera products are screened with Pintail (*47*) and Bellerophon (*48*) and discarded from further analysis (no such chimeras were found in this study). The chance however that the same ribotypes would repeatedly colocalize and either form a chimera or amplify the wrong rRNA gene are extremely small. To estimate the chance for such an event, we shall consider the case where the host 16S rRNA gene, ***S***, repeatedly colocalized with the same rRNA gene ***S'***, and that the foreign 16S rRNA gene (***S'***) was amplified while the host 16S rRNA gene (***S***) was not amplified. The average number of such chance events per panel where the host terminase was also amplified is given by $I_{SS'} = \varepsilon_{ter}\varepsilon_{16S}(1-\varepsilon_{16S})\left(f_sY\right)\left(f_{s'}Y\right)/765$, where ε_{ter} and ε_{16S} are the amplification efficiencies of the terminase gene and the 16S rRNA gene, respectively (see Materials and methods for an estimation of these efficiencies), $f_{s'}$ is the frequency of the ***S'*** ribotype, and $\left(f_sY\right)\left(f_{s'}Y\right)/765$ is the number of chance colocalizations of ***S*** and ***S'*** cell types on a given panel. The probability therefore of retrieving such events is $p_F^{mixed} = I_{SS'}/I_{\text{all 16S-ter}}$. Assuming $f_{s'} \sim 0.2$ (corresponding to the worst case scenario of co-localizing with the most frequent ribotype on the chip, REP1) then based on Table S13 we have $p_F^{mixed} << p_F$ (where p_F is given in Eq. S3) and therefore these events can be neglected (the P values for such events would be much smaller than those in Table 1).

Uniformity of panel loading

On a few occasions, panels were loaded by the NanoFlex somewhat nonuniformly. This has the consequence of reducing the effective number of wells available for the cells. The samples affected for host I were C2 and G1. The terminases of samples C2 and G1 fell in the main clade of host I of highly similar terminases (clade V1 in Fig. 2), lending support for these colocalizations. Sample G2 (host III) was taken from a slightly nonuniform panel, however the terminase of sample G2 was 100% identical at the amino acid level (235 aa alignment) to F2 also associated with host III, lending support for this colocalization. Samples affected for host II were A4 and A7, however the terminase of A4 was 99.6% identical at the amino acids level (235 aa alignment) to the terminase of A9i also of host II, lending support for this sample. The terminase of A7 was 95.3% identical at the amino acids level to the terminase of A13i also of host II, lending support for this sample.

Estimation of the P value for putative *Treponema* non-host (REPs1–3)

The phylotypes REP1, REP2, and REP3 were highly repeated in the random rRNA reference library ($f_{\mathrm{S}} = 23/118, 8/118, 7/118$, respectively) but were never sampled in the colocalization library (*n*=41). The null hypothesis is therefore that ribotype ***S*** is a genuine host but was not sampled *n*=41 times by chance. We wish to calculate the probability for this event. The fraction of colocalizations in a given panel that contain host ***S*** is given on average by

(S6) $$p_S = \frac{\varepsilon_{ter} \cdot f_{\mathrm{S}} \cdot Y}{I_{\text{all 16S-ter}}}$$

where ε_{ter} is the efficiency of amplification for the terminase gene (see Materials and methods), f_{S} the frequency of host ***S*** on the chip, *Y* the number of 16S rRNA hits on a given panel, and $I_{\text{all 16S-ter}}$ is the number colocalizations on a panel of a 16S rRNA target with an actual terminase target (Eq. S2b). Therefore $\varepsilon_{ter} \cdot f_{\mathrm{S}} \cdot Y$ is the number of expected genuine colocalizations for ribotype **S**, and $\varepsilon_{ter} \cdot f_{\mathrm{S}} \cdot Y / I_{\text{all 16S-ter}}$ would be the probability to sample this colocalization. The probability (P value, one tailed, *n*=41) for not retrieving ***S*** (*k*=0) after *n*=41 trials is given by

$$\text{P value} = \text{Prob}\left(k = 0 \,|\, n = 41 \text{ successful retrievals}\right) = (1 - p_S)^n$$

where p_S is averaged using Bayes' theorem as described above (i.e., a panel-weighed average based on Table S13 for all 41 retrievals). For $\varepsilon_{ter} \approx 0.8$ (measured value) we find that the P value (one tailed test with *n*=41) for not retrieving a host with a frequency of $f_{\mathrm{S}} \geq 7/118$ is $\leq 4.8 \cdot 10^{-20}$ allowing us to reject this hypothesis. If REPs-1, 2, and 3 are infected in only >5%, 14%, and 16% of the cases respectively, then the P value for not retrieving these infected strains is 0.01 (one tailed test with *n*=41). Therefore based on statistical grounds we conclude that the majority of REP1–3 cells are not infected. Furthermore 21 out of 23 REP-1 ribotypes, 8 out of 8 REP-2 ribotypes, and 7 out of 7 REP-3 ribotypes were not associated with a terminase hit on the microfluidic chips. Of the two positive hits for REP-1, post-amplification followed by agarose gel electrophoresis showed that just one of these samples contained a terminase target. Statistically, out of *n*=38 occurrences of REPs1–3, $p_{ter} \cdot n$ should randomly colocalize with a terminase target on the chip, or 0.4 ± 0.2 random colocalizations, as observed. This is consistent with the hypothesis that REPs1–3 are indeed non-hosts.

The viral marker gene and its genetic context

Requirements for a viral marker gene

Since certain viral genes can be of bacterial origin, and some viral genes may not be associated with an actual functional virus, a genuine viral marker should satisfy certain requirements (*29*). We were therefore interested in choosing as a viral marker a gene that (a) was unique to viruses, (b) was present in a larger viral context, (c) was prevalent in the ecosystem we were investigating, (d) contained multiple conserved regions that could be used to design degenerate primers, and (e) is active or has been active in recent evolutionary history in this system. The large terminase subunit chosen as a viral marker gene fulfilled all of the above requirements:

(1) The large terminase subunit is considered to be one of the most universally conserved phage genes and best phage identifiers (*29*), exhibiting certain conserved residues and motifs (see Figs. S2 and S3). Furthermore, since typically different phages exhibit little overall sequence similarly (see main text), the terminase gene also appears to be system specific (*77*), thereby potentially serving as a good differentiating marker (*36*).

(2) Bioinformatic analysis of the ZAS-2 and ZAS-9 genomes revealed four prophage-like elements (two in each genome) that were related to tailed phages based on their sequence homology. The largest of these elements (ZAS-2A) spanned 43.5 kb, which is a typical size for tailed phages (*78*). Furthermore, all four copies of the terminase gene in the ZAS genomes had homologs in the higher termite metagenome with 77 to 79% amino acid identity. The largest of these elements, ZAS-2A, appeared to be associated with the *Caudovirales* order: When BLASTing each of the 41 identified genes in this prophage-like element against NCBI's viral RefSeq (v37) database, 16 genes had significant hits ($E < 0.005$), with 15 out of the 16 genes being associated with homologs present in viruses belonging to the *Caudovirales* order. The viral genes also follow a typical tailed-phage gene organization pattern (*79*). For example genes ZA3, ZA4, ZA5, ZA7, ZA8 are the head related genes (homologous to the small and large terminase subunit genes, portal protein gene, prohead protease gene, and capsid protein gene, respectively), whereas genes ZA32 and ZA33 towards the end of the cassette exhibited a weak homology to a tail fiber gene and a tail tape measure protein gene, respectively ($E = 0.16$, 0.29, respectively). Among the 15 hits above, 11 were associated with the *Siphoviridae* family, two with the *Podoviridae* and two with the *Myoviridae* family. The last four genes appear to be less diagnostic than the *Siphoviridae*-related genes as they are not signature phage genes and the E value for three of these genes was low ($E \geq 0.001$). Although it is possible that the prophage-like elements are mosaics of *Caudovirales* families (*80*), based on the above analysis it appears that these elements are mostly closely related to the *Siphoviridae* family.

(3) Bioinfomratic analysis of the metagenome (Table S1) identified the large terminase subunit as one of the most abundant viral-unique genes in the metagenome (though this may not reflect absolute abundance in the sample due to assembler bias). In addition, more generally, the ZAS prophage-like elements appear to be ubiquitous to the termite environment as certain cassettes within the ZAS prophage-like elements were found to be abundant in the higher termite metagenome. For example, the large terminase subunit and its adjacent portal protein from ZAS-2A had a maximum percent amino acid identity of 78% and 70%, respectively, when BLASTed against the metagenome (Table S2) and were homologous to 46 and 43 metagenome gene objects, respectively (E ≤ 1e-5). Furthermore, these two genes, that are adjacent to each other in the ZAS genomes (a typical organization in viruses (*29*)) were also found to be next to each other in the metagenome contigs.

(4) Alignment of the terminase alleles from the ZAS genomes and the higher termite metagenome revealed multiple conserved regions that could be used for primer design (Fig. S2).

(5) Viral-specific genes encoded by ZAS-2 and ZAS-9 prophage-like elements (the portal protein, the capsid protein, the large terminase subunit and the prohead protease protein) exhibited substantial negative selection pressure (data not shown). In addition, the terminase genes retrieved from *R. hesperus* specimens also exhibited substantial negative selection pressure (see Materials and methods and Table S6). This evidence suggests that the terminase gene in the termite system if not functional, has been functional in recent evolutionary history (see discussion below). In addition, there is some anecdotal evidence suggesting the terminase is part of an active viral entity. In one of the earlier experiments with the microfluidic arrays (prior to execution of arrays A through G from which samples were retrieved), where chilling of samples to 4°C was not strictly enforced, a dilution series of a *Zootermopsis nevadensis* termite hindgut fluid was loaded onto a microfluidic array. The panel on the array corresponding to the largest gut dilution exhibited 34.9 times the number of expected terminase hits (384 observed versus 11 expected), where the expected number of hits was estimated based on the number of hits from more concentrated dilutions loaded onto the same microfluidic array. At the same time, the rRNA channel displayed the expected number of hits (72 observed versus 74 expected) for this dilution. Since the degenerate terminase primers that were used in the qPCR chemistry were designed based on the terminase alleles in the ZAS-2 and ZAS-9 prophage-like elements (among other alleles), this induction event is specific to the terminase gene investigated in this study. This result indicates that a lytic event associated with the prophage-like element may have taken place in the tube containing the largest gut dilution, suggesting that this putative prophage is functional. We note that earlier experiments to induce the ZAS-2 and ZAS-9 cultures using mitomycin C were not successful, suggesting that mitomycin C may not be the inducing agent of this element.

Functionality of the terminase gene

Given the fact that the terminase gene is under negative selection pressure and in the absence of obvious frame shift mutations or errant stop codons in the alignment, there are several options regarding the nature of the prophage-like element in which it resides and the functionality of the terminase gene within these elements: **(1)** the terminase is part of an active prophage (for which there is some evidence, as discussed in point 5 above) **(2)** the terminase is part of a defective prophage but it remained functional because there was not enough time for point mutations to have accumulated. This can happen because "prophage-debilitating deletions can accumulate more rapidly than gene-inactivating point mutations" (*29*). **(3)** The prophage indeed decayed and the terminase gene degraded over time, but was subsequently repaired by a recombination event with another phage that was likely functional (since it infected the cell in the first place) (*29*). Finally, **(4)** the terminase was recruited by the bacterium because it confers on the bacterium some competitive advantage and is therefore under negative selection pressure.

To further elaborate on the last point (4), phage genes that are adopted by the cell are typically lysogenic conversion genes (*29*)—genes that change the phenotype of the cell and confer some selective advantage on the cell. In this context, known possibilities may be (*29*) tail-like bacteriocins and genetic transfer agents (GTAs). Bacteriocins are devices that kill other bacteria and some bacteria can produce bacteriocins that resemble phage tails (*29, 81*). However since these entities do not have heads or package DNA it seems unlikely they would encode a terminase gene. For example, type F and type R tail-like bacteriocins of *Pseudomonas aeruginosa* PAO1 do not appear to encode a terminase gene or any other head related proteins (*82-83*). GTAs are tailed phage-like particles that encapsidate random fragments of the bacterial genome and can transfer them to other bacteria of the same species (*29*). GTAs are thought to be adopted by the host cell to facilitate genetic exchange under the control of the host (*84-86*). The GTA coding region is typically short (~14–16 kb (*86*)) and appears to contain the genes required for assembly of the GTA head and tail structures and the genes required for DNA packaging (including a terminase gene) (*84, 86*). Phage DNA-specific replication functions and phage DNA-specific integration or excision functions are in principle not required by the GTA (*84*). Although it cannot be ruled out that the terminase genes retrieved from *R. hesperus* are part of a GTA, this possibility appears to be unlikely since the predicted prophage-like element identified in ZAS-2 spans ~43.5 kb (a typical length for a functional phage), which is much longer than a typical GTA length (14–16 kb—see above). In addition, unlike GTAs, the ZAS-2 prophage-like element encodes both integration genes and several DNA replication machinery genes.

To summarize, the fact that the *R. hesperus* terminase alleles are under substantial negative selection pressure suggests that this terminase is either active or has been active in recent evolutionary history and was the direct or indirect result of a viral infection (options 1, 2, or 3 above). The possibility that the terminase was adopted by the cell and is part of a GTA appears to be unlikely. Thus the associations between the hosts and the terminase genes revealed by the microfluidic assay should be a valid proxy for interaction of these hosts with genuine infecting phages, reflecting either current or recent infections.

SUPPORTING FIGURES

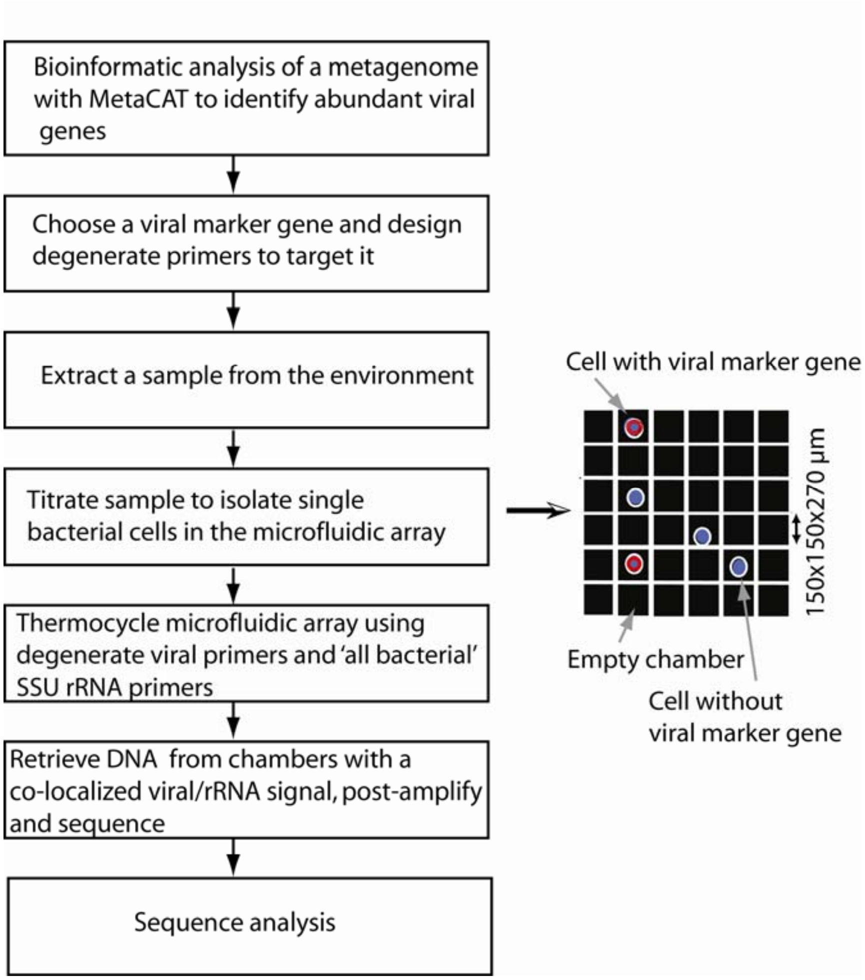

Bioinformatic analysis of a metagenome with MetaCAT to identify abundant viral genes
Choose a viral marker gene and design degenerate primers to target it
Extract a sample from the environment
Titrate sample to isolate single bacterial cells in the microfluidic array
Thermocycle microfluidic array using degenerate viral primers and 'all bacterial' SSU rRNA primers
Retrieve DNA from chambers with a co-localized viral/rRNA signal, post-amplify and sequence
Sequence analysis
Cell with viral marker gene
150x150x270 μm
Empty chamber
Cell without viral marker gene

Figure S1. Workflow using the microfluidic digital PCR array for host-virus colocalization in a novel environmental sample. See Materials and methods for further details.

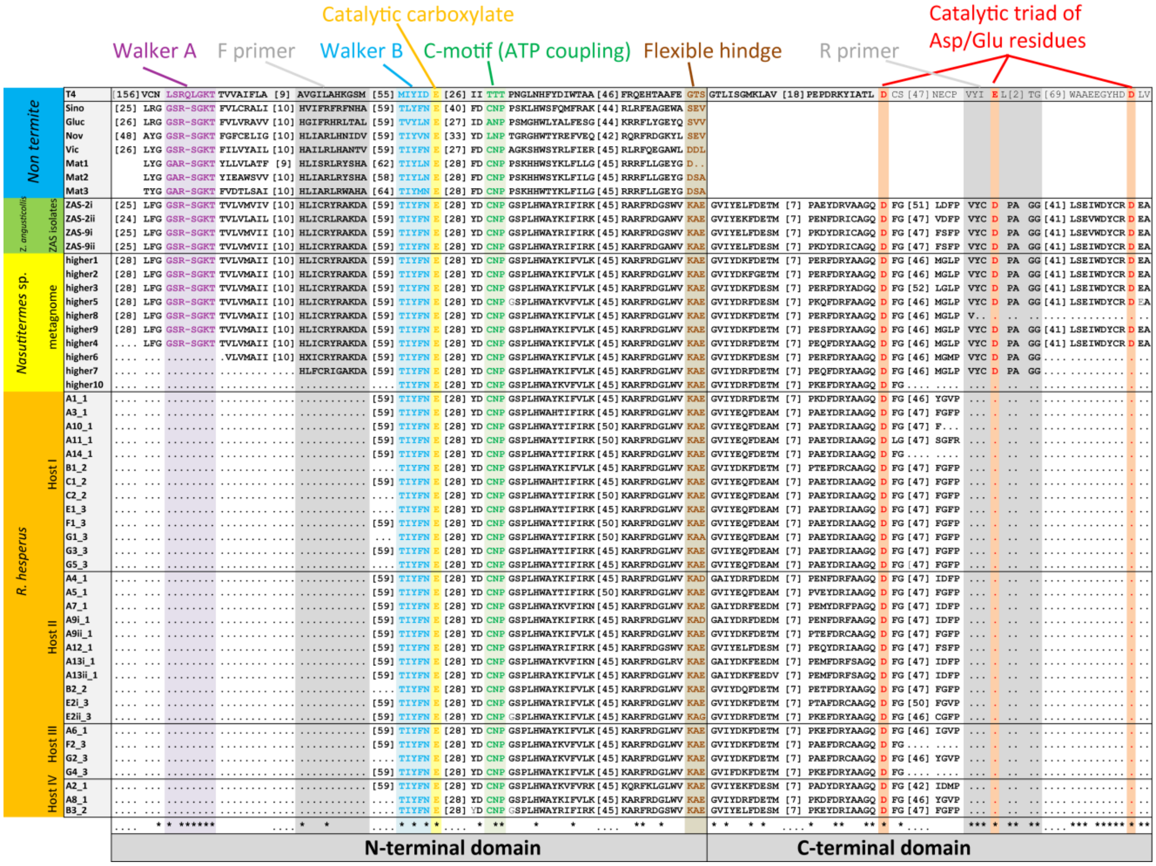


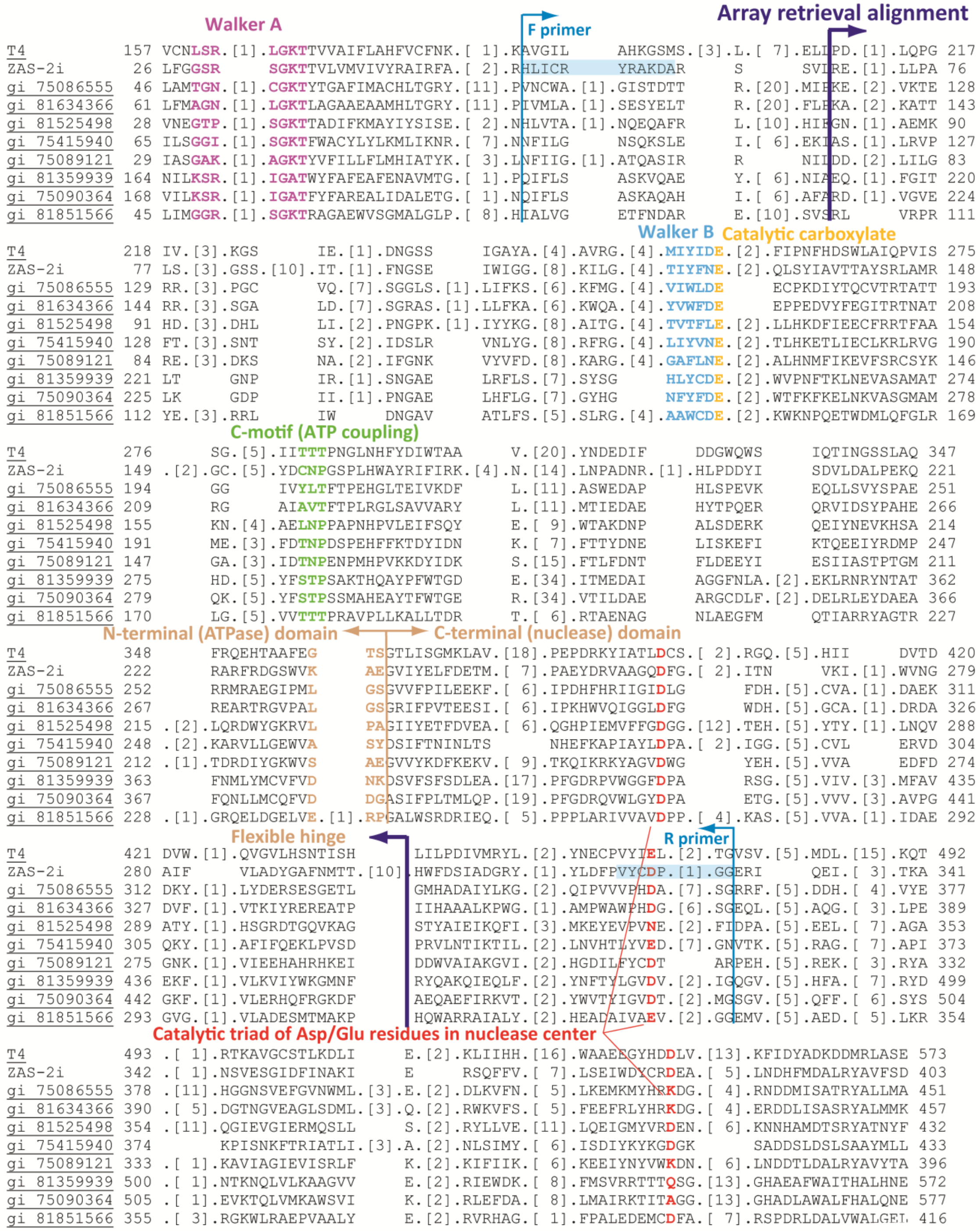
Figure S2. Multiple alignment of termite-related terminase sequences and closest homologs. Here we show a multiple alignment of terminase genes of both termite and non-termite origin highlighting putative functional motifs. Terminase sequences included are (1) terminase sequences retrieved from *R. hesperus* termites using the digital PCR, (2) homologous terminases from the metagenome of a *Nasutitermes* sp. termite, (3) homologous terminases from *Treponema* isolates obtained from a *Z. angusticollis* termite, and (4) homologous terminases from non-termite-related bacteria found in public databases (NCBI's protein RefSeq database and the Joint Genome Institute database). Also highlighted are putative conserved functional motifs for the N-terminal ATPase center and the C-terminal nuclease center (see Fig. S3). When searching for homologs for the ZAS2-i terminase gene in public databases, the N-terminal ATPase domain of this gene (amino acids 1–234—see Fig. S3) appeared to be much more conserved (47% identity) than the entire gene (29% identity). Consistent with this fact, the ATPase domain of the large terminase subunit has been shown to be conserved in a wide variety of dsDNA (*30*) viruses and even shows certain conserved motifs with the putative herpesvirus terminase (*30,87*) suggesting it is an ancient viral domain (*30, 88, 89*). We therefore show here only the N-terminal domain alignment of non-termite homologous terminases.

N-terminal alignment: The boundary of the N-terminal domain for the terminase alleles was determined based on its location in T4 (residue 360) (*90*) by aligning the amino acid sequences of the ZAS2-i terminase and all non-termite-related terminases with RPS-BLAST against pfam03237 (*90*) in the CDD (*91*) (see Fig. S3 for ZAS2-i alignment). The N-terminal domain of other termite-related sequences was then determined by a MUSCLE alignment to the ZAS2-i terminase (*92*). All N-terminal domains were then MUSCLE aligned. **C-terminal alignment:** maximum length termite-related terminases were MUSCLE aligned and then only their C-terminal regions were juxtaposed to the N-terminus alignment found above (the overlap with the N-terminus alignment was identical).

Functional motifs were identified based on an RPS-BLAST alignment of ZAS-2i against pfam03237 (Fig. S3). This figure demonstrates that the termite-related terminase sequences exhibit terminase-like functional motifs. Putative functional motifs include (1) Walker A motif G/A-XXXXGK(T/S) (purple) with a single residue X deletion, (2) Walker B ZZZZD motif with D replaced by N—a relatively common substitution for this residue (blue), (3) catalytic carboxylate group motif —E (orange), (4) putative ATP coupling motif (green), and (5) catalytic Asp/Glu triad motif—here a conserved D (red) (*26,31*). Also highlighted is the putative flexible hinge motif (brown) (*31*) based on the RPS-BLAST alignment. Numbers in brackets correspond to aligned residues not shown. Stars indicate conserved residues excluding T4. Dots indicate end of available sequence. X residues in the higher termite sequences are due to ambiguous base pairs in the nucleotide sequence. The RPS-BLAST ZAS2-i alignment with T4 (Fig. S3) was superimposed to guide the eye and was not part of the MUSCLE alignment. Also shown are the primer binding sites. The degenerate core region of the CODEHOP primers (*17*) that is required to be conserved consists of 4 amino acids at the 3' end of the primer. Out of the 50 ZAS and higher termite gut alleles, 31 alleles included the forward primer motif and 26 alleles included the reverse primer motif. In all cases, the degenerate core region of the primers was strictly conserved. In one additional allele, the sequence began from the center Asp residue in the conserved catalytic Asp/Glu triad motif. This residue was mutated in this allele from an Asp residue to a Gly residue suggesting this partial allele encodes a non-functional terminase. Thus, all functional alleles of the terminase gene exhibited a strictly conserved degenerate core region. Note that the Walker A motif was not chosen for a forward primer binding site due to the high degeneracy involved with this amino acid sequence.

To check what diversity of terminase genes are expected to be amplified, we BLASTed the core region of the forward (ter7F) and reverse (ter5eR) terminase primers against all viral genes in NCBI's viral RefSeq database v37. Only the core region of the primer was used in the BLAST analysis (a more general search) because the primers are CODEHOP primers and therefore while the degenerate core region (11–12 bases in the 3' region of the primer) must base pair with the target, homology of the clamp region is less critical for initial amplification. We then crossed the list of hits for the forward and reverse primers searching for mutual hits present in the same gene within the same bacteriophages, however no such solutions were found. Based on this result we anticipate that the degenerate terminase primers target the unique diversity of terminase genes currently known to exist only in termite and possibly related insect species.

Non-termite-related terminases (Vic, Sino, Gluc, and Nov) are gram negative isolates belonging to the Lentisphaerae and Proteobacteria phyla. These bacteria grow in a variety of habitats (human gut, soil, fresh water, plants, etc.) and can either be free living or symbiotic, anaerobic or aerobic. Mat1, Mat2, and Mat3 were found to be present in the metagenome of a hypersaline microbial mat from Mexico (see Table S10 for accession numbers).

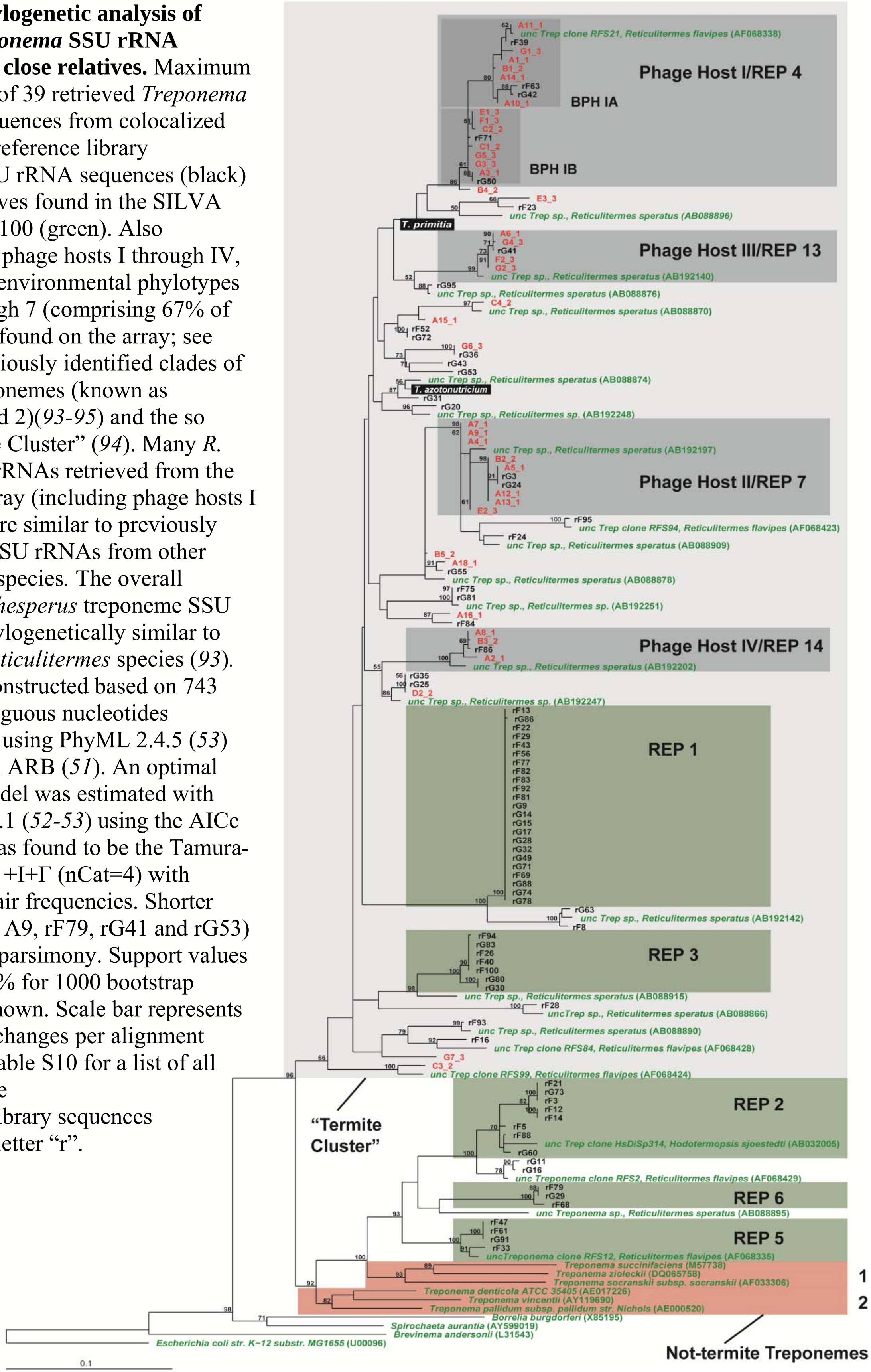

Walker A
F primer
Array retrieval alignment
Walker B
Catalytic carboxylate
C-motif (ATP coupling)
N-terminal (ATPase) domain
C-terminal (nuclease) domain
Flexible hinge
R primer
Catalytic triad of Asp/Glu residues in nuclease center

```
T4             157 VCNLSR.[1].LGKTTVVAIFLAHFVCFNK.[ 1].KAVGIL     AHKGSMS.[3].L.[ 7].ELIPD.[1].LQPG 217
ZAS-2i          26 LFGGSR     SGKTTVLVMVIVYRAIRFA.[ 2].RHLICR     YRAKDAR     S      SVIRE.[1].LLPA 76
gi 75086555     46 LAMTGN.[1].CGKTYTGAFIMACHLTGRY.[11].PVNCWA.[1].GISTDTT     R.[20].MIEKE.[2].VKTE 128
gi 81634366     61 LFMAGN.[1].LGKTLAGAAEAAMHLTGRY.[11].PIVMLA.[1].SESYELT     R.[20].FLEKA.[2].KATT 143
gi 81525498     28 VNEGTP.[1].SGKTTADIFKMAYIYSISE.[ 2].NHLVTA.[1].NQEQAFR     L.[10].HIEGN.[1].AEMK 90
gi 75415940     65 ILSGGI.[1].SGKTFWACYLYLKMLIKNR.[ 7].NNFILG     NSQKSLE     I.[ 6].EKIAS.[1].LRVP 127
gi 75089121     29 IASGAK.[1].AGKTYVFILLFLMHIATYK.[ 3].LNFIIG.[1].ATQASIR     R      NIIDD.[2].LILG 83
gi 81359939    164 NILKSR.[1].IGATWYFAFEAFENAVMTG.[ 1].PQIFLS     ASKVQAE     Y.[ 6].NIAEQ.[1].FGIT 220
gi 75090364    168 VILKSR.[1].IGATFYFAREALIDALETG.[ 1].NQIFLS     ASKAQAH     I.[ 6].AFARD.[1].VGVE 224
gi 81851566     45 LIMGGR.[1].SGKTRAGAEWVSGMALGLP.[ 8].HIALVG     ETFNDAR     E.[10].SVSRL     VRPR 111

T4             218 IV.[3].KGS      IE.[1].DNGSS     IGAYA.[4].AVRG.[4].MIYIDE.[2].FIPNFHDSWLAIQPVIS 275
ZAS-2i          77 LS.[3].GSS.[10].IT.[1].FNGSE     IWIGG.[8].KILG.[4].TIYFNE.[2].QLSYIAVTTAYSRLAMR 148
gi 75086555    129 RR.[3].PGC      VQ.[7].SGGLS.[1].LIFKS.[6].KFMG.[4].VIWLDE     ECPKDIYTQCVTRTATT 193
gi 81634366    144 RR.[3].SGA      LD.[7].SGRAS.[1].LLFKA.[6].KWQA.[4].YVWFDE     EPPEDVYFEGITRTNAT 208
gi 81525498     91 HD.[3].DHL      LI.[2].PNGPK.[1].IYYKG.[8].AITG.[4].TVTFLE.[2].LLHKDFIEECFRRTFAA 154
gi 75415940    128 FT.[3].SNT      SY.[2].IDSLR     VNLYG.[8].RFRG.[4].LIYVNE.[2].TLHKETLIECLKRLRVG 190
gi 75089121     84 RE.[3].DKS      NA.[2].IFGNK     VYVFD.[8].KARG.[4].GAFLNE.[2].ALHNMFIKEVFSRCSYK 146
gi 81359939    221 LT     GNP      IR.[1].SNGAE     LRFLS.[7].SYSG     HLYCDE.[2].WVPNFTKLNEVASAMAT 274
gi 75090364    225 LK     GDP      II.[1].PNGAE     LHFLG.[7].GYHG     NFYFDE.[2].WTFKFKELNKVASGMAM 278
gi 81851566    112 YE.[3].RRL      IW     DNGAV     ATLFS.[5].SLRG.[4].AAWCDE.[2].KWKNPQETWDMLQFGLR 169

T4             276      SG.[5].IITTTPNGLNHFYDIWTAA     V.[20].YNDEDIF     DDGWQWS     IQTINGSSLAQ 347
ZAS-2i         149 .[2].GC.[5].YDCNPGSPLHWAYRIFIRK.[4].N.[14].LNPADNR.[1].HLPDDYI     SDVLDALPEKQ 221
gi 75086555    194      GG     IVYLTFTPEHGLTEIVKDF     L.[11].ASWEDAP     HLSPEVK     EQLLSVYSPAE 251
gi 81634366    209      RG     AIAVTFTPLRGLSAVVARY     L.[11].MTIEDAE     HYTPQER     QRVIDSYPAHE 266
gi 81525498    155      KN.[4].AELNPPAPNHPVLEIFSQY     E.[ 9].WTAKDNP     ALSDERK     QEIYNEVKHSA 214
gi 75415940    191      ME.[3].FDTNPDSPEHFFKTDYIDN     K.[ 7].FTTYDNE     LISKEFI     KTQEEIYRDMP 247
gi 75089121    147      GA.[3].IDTNPENPMHPVKKDYIDK     S.[15].FTLFDNT     FLDEEYI     ESIIASTPTGM 211
gi 81359939    275      HD.[5].YFSTPSAKTHQAYPFWTGD     E.[34].ITMEDAI     AGGFNLA.[2].EKLRNRYNTAT 362
gi 75090364    279      QK.[5].YFSTPSSMAHEAYTFWTGE     R.[34].VTILDAE     ARGCDLF.[2].DELRLEYDAEA 366
gi 81851566    170      LG.[5].VVTTTPRAVPLLKALLTDR     T.[ 6].RTAENAG     NLAEGFM     QTIARRYAGTR 227

T4             348      FRQEHTAAFEG     TSGTLISGMKLAV.[18].PEPDRKYIATLDCS.[ 2].RGQ.[5].HII     DVTD 420
ZAS-2i         222      RARFRDGSWVK     AEGVIYELFDETM.[ 7].PAEYDRVAAGQDFG.[ 2].ITN     VKI.[1].WVNG 279
gi 75086555    252      RRMRAEGIPML     GSGVVFPILEEKF.[ 6].IPDHFHRIIGIDLG      FDH.[5].CVA.[1].DAEK 311
gi 81634366    267      REARTRGVPAL     GSGRIFPVTEESI.[ 6].IPKHWVQIGGLDFG      WDH.[5].GCA.[1].DRDA 326
gi 81525498    215 .[2].LQRDWYGKRVL     PAGIIYETFDVEA.[ 6].QGHPIEMVFFGDGG.[12].TEH.[5].YTY.[1].LNQV 288
gi 75415940    248 .[2].KARVLLGEWVA     SYDSIFTNINLTS      NHEFKAPIAYLDPA.[ 2].IGG.[5].CVL     ERVD 304
gi 75089121    212 .[1].TDRDIYGKWVS     AEGVVYKDFKEKV.[ 9].TKQIKRKYAGVDWG      YEH.[5].VVA     EDFD 274
gi 81359939    363      FNMLYMCVFVD     NKDSVFSFSDLEA.[17].PFGDRPVWGGFDPA      RSG.[5].VIV.[3].MFAV 435
gi 75090364    367      FQNLLMCQFVD     DGASIFPLTMLQP.[19].PFGDRQVWLGYDPA      ETG.[5].VVV.[3].AVPG 441
gi 81851566    228 .[1].GRQELDGELVE.[1].RPGALWSRDRIEQ.[ 5].PPPLARIVVAVDPP.[ 4].KAS.[5].VVA.[1].IDAE 292

T4             421 DVW.[1].QVGVLHSNTISH      LILPDIVMRYL.[2].YNECPVYIEL.[2].TGVSV.[5].MDL.[15].KQT 492
ZAS-2i         280 AIF     VLADYGAFNMTT.[10].HWFDSIADGRY.[1].YLDFPVYCDP.[1].GGERI     QEI.[ 3].TKA 341
gi 75086555    312 DKY.[1].LYDERSESGETL      GMHADAIYLKG.[2].QIPVVVPHDA.[7].SGRRF.[5].DDH.[ 4].VYE 377
gi 81634366    327 DVF.[1].VTKIYREREATP      IIHAAALKPWG.[1].AMPWAWPHDG.[6].SGEQL.[5].AQG.[ 3].LPE 389
gi 81525498    289 ATY.[1].HSGRDTGQVKAG      STYAIEIKQFI.[3].MKEYEVPVNE.[2].FIDPA.[5].EEL.[ 7].AGA 353
gi 75415940    305 QKY.[1].AFIFQEKLPVSD      PRVLNTIKTIL.[2].LNVHTLYVED.[7].GNVTK.[5].RAG.[ 7].API 373
gi 75089121    275 GNK.[1].VIEEHAHRHKEI      DDWVAIAKGVI.[2].HGDILFYCDT     ARPEH.[5].REK.[ 3].RYA 332
gi 81359939    436 EKF.[1].VLKVIYWKGMNF      RYQAKQIEQLF.[2].YNFTYLGVDV.[2].IGQGV.[5].HFA.[ 7].RYD 499
gi 75090364    442 GKF.[1].VLERHQFRGKDF      AEQAEFIRKVT.[2].YWVTYIGVDT.[2].MGSGV.[5].QFF.[ 6].SYS 504
gi 81851566    293 GVG.[1].VLADESMTMAKP      HQWARRAIALY.[2].HEADAIVAEV.[2].GGEMV.[5].AED.[ 5].LKR 354

T4             493 .[ 1].RTKAVGCSTLKDLI     E.[2].KLIIHH.[16].WAAEEGYHDDLV.[13].KFIDYADKDDMRLASE 573
ZAS-2i         342 .[ 1].NSVESGIDFINAKI     E     RSQFFV.[ 7].LSEIWDYCRDEA.[ 5].LNDHFMDALRYAVFSD 403
gi 75086555    378 .[11].HGGNSVEFGVNWML.[3].E.[2].DLKVFN.[ 5].LKEMKMYHRKDG.[ 4].RNDDMISATRYALLMA 451
gi 81634366    390 .[ 5].DGTNGVEAGLSDML.[3].Q.[2].RWKVFS.[ 5].FEEFRLYHRKDG.[ 4].ERDDLISASRYALMMK 457
gi 81525498    354 .[11].QGIEVGIERMQSLL     S.[2].RYLLVE.[11].LQEIGMYVRDEN.[ 6].KNNHAMDTSRYATNYF 432
gi 75415940    374       KPISNKFTRIATLI.[3].A.[2].NLSIMY.[ 6].ISDIYKYKGDGK      SADDSLDSLSAAYMLL 433
gi 75089121    333 .[ 1].KAVIAGIEVISRLF     K.[2].KIFIIK.[ 6].KEEIYNYVWKDN.[ 6].LNDDTLDALRYAVYTA 396
gi 81359939    500 .[ 1].NTKNQLVLKAAGVV     E.[2].RIEWDK.[ 8].FMSVRRTTTQSG.[13].GHAEAFWAITHALHNE 572
gi 75090364    505 .[ 1].EVKTQLVMKAWSVI     K.[2].RLEFDA.[ 8].LMAIRKTITAGG.[13].GHADLAWALFHALQNE 577
gi 81851566    355 .[ 3].RGKWLRAEPVAALY     E.[2].RVRHAG.[ 1].FPALEDEMCDFA.[ 7].RSPDRLDALVWALGEL 416
```

Figure S3. Multiple alignment of pfam03237 with a ZAS-associated terminase. Multiple sequence alignment of pfam03237 (Terminase_6) with the ZAS-2 terminase sequence (ZAS-2i) aligned with RPS-BLAST in the CDD (*91*) (E value 1.2e-19). Conserved functional motifs (*26, 31*) are indicated as well as the boundary between the N-terminal ATPase domain (T4: amino acids 1–360 (*31*)) and C-terminal nuclease domain (T4: amino acids 361–610 (*31*)) based on T4 (*90, 26*). Conserved functional motifs for the N terminal ATPase center include (*26,31*) a Walker A motif G/A-XXXXGK(T/S) (purple), a Walker B motif ZZZZD where Z represents a

hydrophobic amino acid (blue), a catalytic carboxylate group motif (usually) Glu (orange), and an ATPase coupling motif (T/S-G/A-T/S(N)) (green). The functional motif for the C-terminal nuclease center is a catalytic triad of Asp/Glu residues (red) (*26, 31*). The forward primer (upper light blue box) targeted a conserved region between the putative Walker A and Walker B motifs in the ATPase domain and the reverse primer targeted a conserved region that included the central aspartic acid residue in the catalytic triad (lower light blue box). Also indicated is the 235 residue alignment region (without gaps) used for phylogenetic analysis. The alignment shows the 10 most diverse members (out of 43) of the pfam with the T4 large terminase subunit gene gp17 being the representative sequence. Numbers in brackets are unaligned residues. ZA2-2i was chosen for the alignment because this gene was found to be present in the largest (43.5 kb) prophoage-like element of the ZAS genome (see supporting text).

Figure S4. Phylogenetic analysis of retrieved *Treponema* SSU rRNA sequences and close relatives. Maximum likelihood tree of 39 retrieved *Treponema* SSU rRNA sequences from colocalized pairs (red), 78 reference library *Treponema* SSU rRNA sequences (black) and close relatives found in the SILVA (*50*) database v100 (green). Also highlighted are phage hosts I through IV, *Reticulitermes* environmental phylotypes (REPs) 1 through 7 (comprising 67% of all treponemes found on the array; see Table S5), previously identified clades of traditional treponemes (known as subgroups 1 and 2)(*93-95*) and the so called "Termite Cluster" (*94*). Many *R. hesperus* SSU rRNAs retrieved from the microfluidic array (including phage hosts I through IV) were similar to previously characterized SSU rRNAs from other *Reticulitermes* species. The overall diversity of *R. hesperus* treponeme SSU rRNAs was phylogenetically similar to that of other *Reticulitermes* species (*93*). The tree was constructed based on 743 aligned unambiguous nucleotides excluding gaps using PhyML 2.4.5 (*53*) implemented in ARB (*51*). An optimal substitution model was estimated with jModelTest 0.1.1 (*52-53*) using the AICc criterion and was found to be the Tamura-Nei model (*54*) +I+Γ (nCat=4) with unequal base pair frequencies. Shorter sequences (A7, A9, rF79, rG41 and rG53) were added by parsimony. Support values greater than 50% for 1000 bootstrap iterations are shown. Scale bar represents 0.1 nucleotide changes per alignment position. See Table S10 for a list of all sequences. Note
that reference library sequences
begin with the letter "r".


Phage Host I/REP 4
BPH IA
BPH IB
62 A11_1
unc Trep clone RFS21, Reticulitermes flavipes (AF068338)
rF39
G1_3
A1_1
B1_2
80 A14_1
88 rF63
rG42
A10_1
E1_3
51 F1_3
C2_2
rF71
C1_2
G5_3
61 G3_3
88 A3_1
86 rG50
B4_2
66 E3_3
rF23
50 unc Trep sp., Reticulitermes speratus (AB088896)
T. primitia
Phage Host III/REP 13
90 A6_1
71 G4_3
73 rG41
91 F2_3
99 G2_3
52 unc Trep sp., Reticulitermes speratus (AB192140)
88 rG95
unc Trep sp., Reticulitermes speratus (AB088876)
97 C4_2
unc Trep sp., Reticulitermes speratus (AB088870)
A15_1
100 rF52
rG72
100 G6_3
73 rG36
71 rG43
rG53
56 unc Trep sp., Reticulitermes speratus (AB088874)
87 T. azotonutricium
rG31
96 rG20
unc Trep sp., Reticulitermes sp. (AB192248)
Phage Host II/REP 7
98 A7_1
62 A9_1
A4_1
unc Trep sp., Reticulitermes speratus (AB192197)
98 B2_2
A5_1
91 rG3
rG24
A12_1
61 A13_1
E2_3
100 rF95
unc Trep clone RFS94, Reticulitermes flavipes (AF068423)
rF24
unc Trep sp., Reticulitermes speratus (AB088909)
B5_2
91 A18_1
rG55
unc Trep sp., Reticulitermes speratus (AB088878)
97 rF75
100 rG81
unc Trep sp., Reticulitermes sp. (AB192251)
87 A16_1
rF84
Phage Host IV/REP 14
A8_1
69 B3_2
rF86
100 A2_1
55 unc Trep sp., Reticulitermes speratus (AB192202)
56 rG35
100 rG25
86 D2_2
unc Trep sp., Reticulitermes sp. (AB192247)
REP 1
rF13
rG86
rF22
rF29
rF43
rF56
rF77
rF82
rF83
rF92
rF81
rG9
rG14
rG15
rG17
rG28
rG32
rG49
rG71
rF69
rG88
100 rG74
100 rG78
rG63
100 unc Trep sp., Reticulitermes speratus (AB192142)
rF8
REP 3
rF94
rG83
rF26
90 rF40
100 rF100
100 rG80
98 rG30
unc Trep sp., Reticulitermes speratus (AB088915)
100 rF28
uncTrep sp., Reticulitermes speratus (AB088866)
99 rF93
79 unc Trep sp., Reticulitermes speratus (AB088890)
92 rF16
unc Trep clone RFS84, Reticulitermes flavipes (AF068428)
66 G7_3
100 C3_2
96 unc Trep clone RFS99, Reticulitermes flavipes (AF068424)
"Termite Cluster"
REP 2
rF21
100 rG73
82 rF3
100 rF12
rF14
70 rF5
rF88
unc Trep clone HsDiSp314, Hodotermopsis sjoestedti (AB032005)
100 rG60
90 rG11
78 rG16
unc Treponema clone RFS2, Reticulitermes flavipes (AF068429)
REP 6
88 rF79
100 rG29
rF68
unc Treponema sp., Reticulitermes speratus (AB088895)
93
REP 5
rF47
100 rF61
100 rG91
91 rF33
uncTreponema clone RFS12, Reticulitermes flavipes (AF068335)
100
89 Treponema succinifaciens (M57738)
93 Treponema zioleckii (DQ065758)
Treponema socranskii subsp. socranskii (AF033306)
92
Treponema denticola ATCC 35405 (AE017226)
82 Treponema vincentii (AY119690)
Treponema pallidum subsp. pallidum str. Nichols (AE000520)
1
2
98
71 Borrelia burgdorferi (X85195)
Spirochaeta aurantia (AY599019)
Brevinema andersonii (L31543)
Escherichia coli str. K-12 substr. MG1655 (U00096)
Not-termite Treponemes
0.1

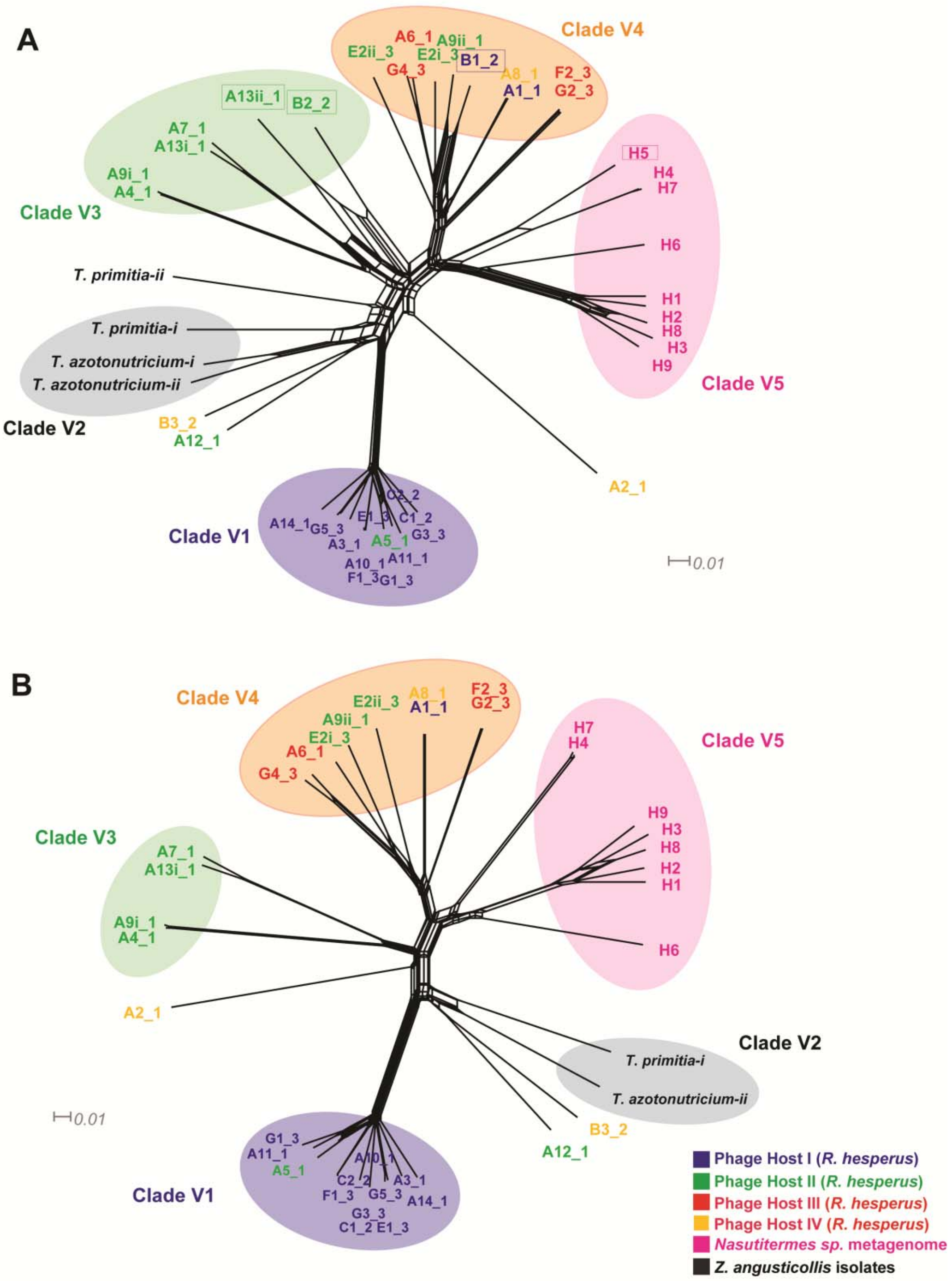


Figure S5. NeighborNet network of termite-related terminase alleles. (A) NeighborNet (*96*) of **(1)** all terminase alleles that were retrieved with phage hosts I through IV, **(2)** terminases genes present in Z. *angusticollis* isolates, *Treponema primitia* (ZAS-2), and *Treponema azotonutricium* (ZAS-9), and **(3)** terminase alleles found in the metagenome of the hindgut of an *Nasutitermes sp.* termite. Boxed sequences are the first four events identified by RDP3 as recombinant (see Methods). **(B)** Same as (A) but excluding **(1)** RDP3 identified recombinant sequences, **(2)** ZAS terminases alleles associated with most likely defunct phage cassettes. ZAS-2 and ZAS-9 both

have two copies of the terminase gene. Each copy resides in a region coding for other viral genes, however only a single one of these copies in each genome appears to be present in a large enough contiguous region of putative viral genes (~36–43 kbp) that could constitute a viable phage and therefore only this copy was included. After removal of recombinant sequences (B1, B2, A13ii, H5) there remains some residual reticulate patterns at the base of the network, however the network largely appears to be tree-like (confirmed by likelihood mapping; see Methods). These sequences were used to generate the terminase tree in Fig. 2. The network structure shown here is consistent with the topology shown in Fig 2. The network was calculated using SplitsTree4 (*67*) on 705 aligned unambiguous nucleotides without gaps using the optimal model found by FindModel (*68*), a K80 substitution model (*97*) +Γ with $\alpha \simeq 0.5$. The LSfit score for networks A and B was 99.97% and 99.94%, respectively. Note that sample B1 associated with host I in (A) was found by RDP3 to be a chimera of A1 (host I) and A9ii (host II), possibly indicating a lateral gene transfer event between these two distinct subpopulations of viruses. Alternatively, since only one such event was observed, it could also be due to an unlikely experimental artifact. Sample notation is as described in Fig. 2.

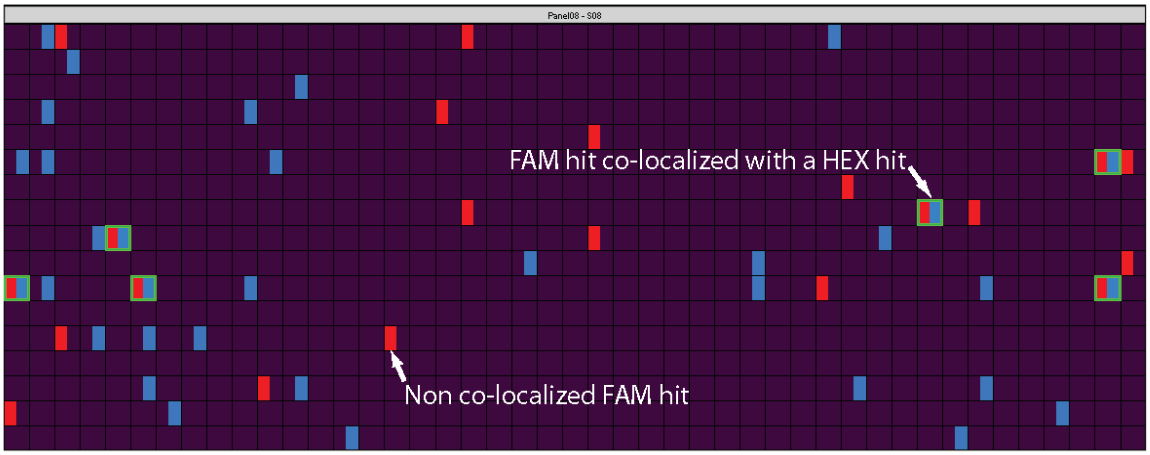


Figure S6. Example of microfluidic array panel readout after thresholding. Blue squares represent hits in the HEX/rRNA channel and red squares represent hits in the FAM/terminase channel. Colocalized hits are highlighted in green. In this example, spurious amplification is expected to account for ~50% of all non-colocalized FAM hits based on the number of FAM hits in the no-template-control panel for this microfluidic array (7 hits).

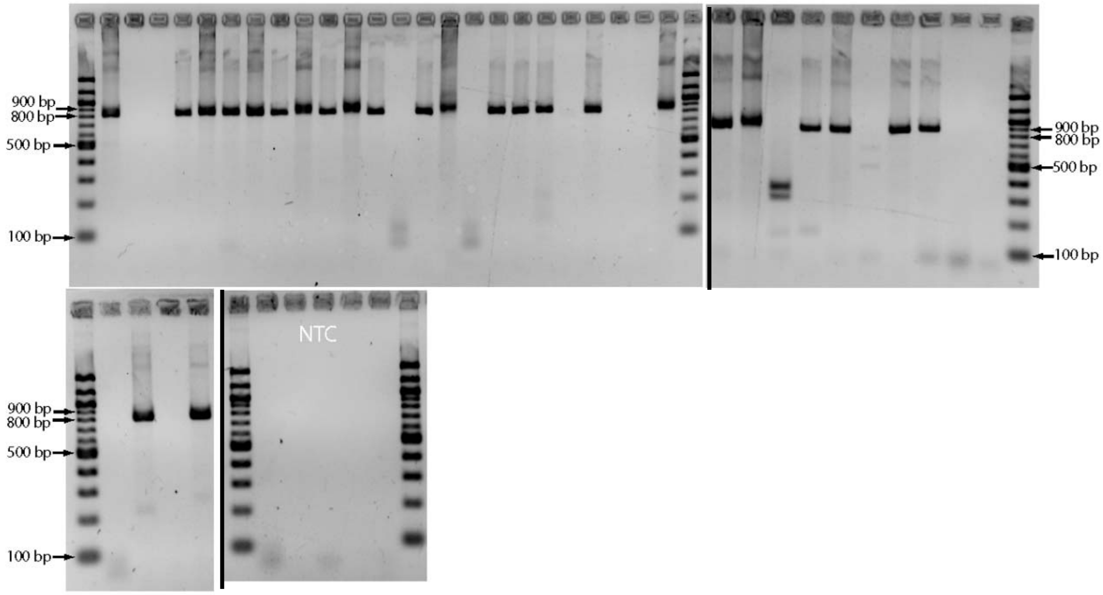


Figure S7. Agarose gel electrophoresis analysis of all FAM hits in a microfluidic array panel. All 38 FAM hits in panel #7 of chip B were post-amplified and analyzed by agarose gel electrophoresis. Also shown are the five no-template-control (NTC) samples for this PCR reaction. The expected amplicon size is ~820 bp (compared to a 100 bp ladder). Out of 38 reactions, 13 were negative for the template. This value is consistent with the number of FAM hits in the no-template-control panel for this microfluidic array, which was 16. The gel image was inverted, brightness was linearly scaled to maximize contrast and size was proportionally scaled to fit the figure. The microfluidic array was analyzed with the BioMark Digital PCR analysis software (Fluidigm, v.2.0.6) using a FAM threshold 0.2 and linear baseline correction.

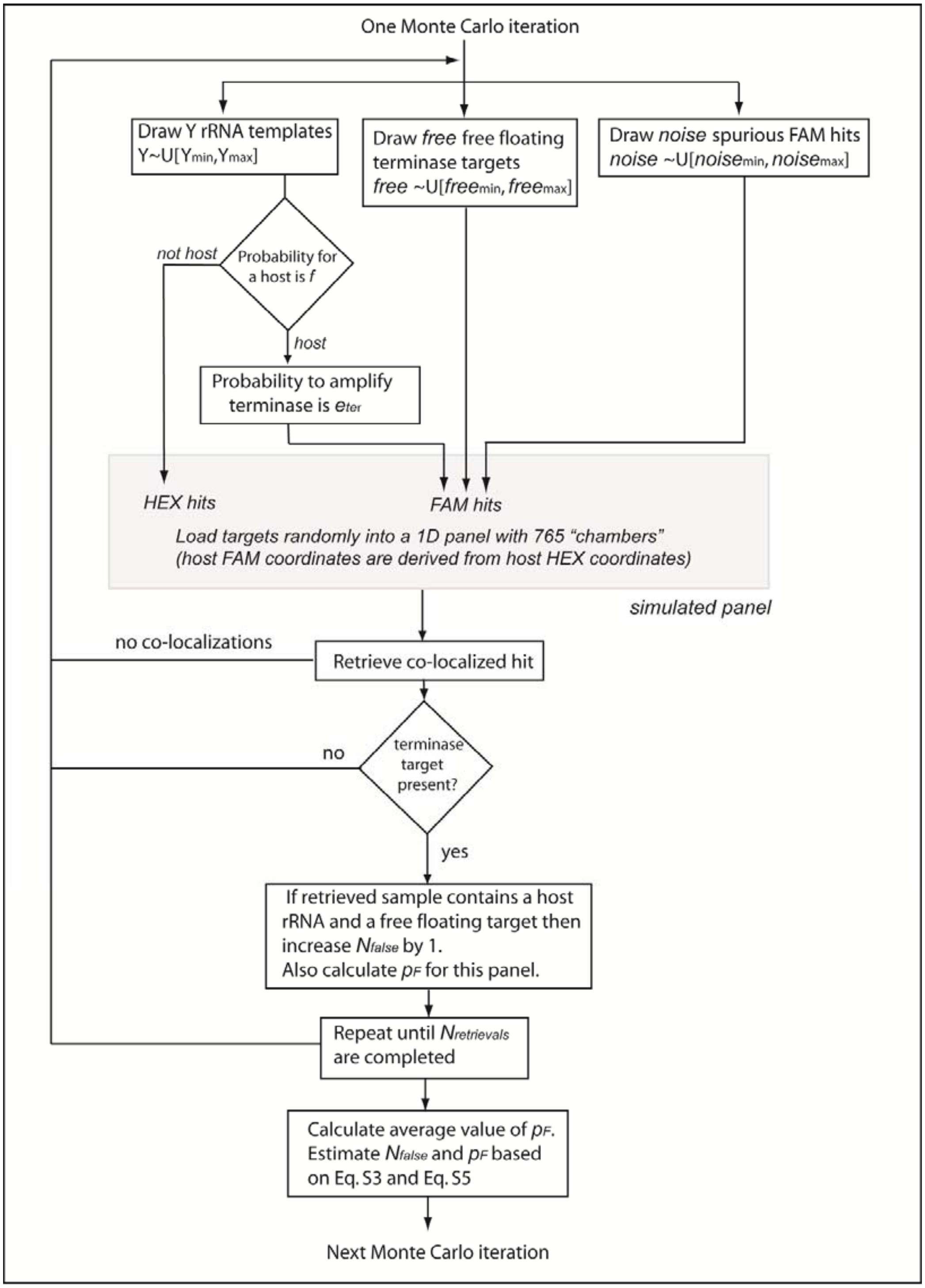

One Monte Carlo iteration
Draw Y rRNA templates $Y \sim U[Y_{min}, Y_{max}]$
Draw *free* free floating terminase targets *free* $\sim U[free_{min}, free_{max}]$
Draw *noise* spurious FAM hits *noise* $\sim U[noise_{min}, noise_{max}]$
not host
Probability for a host is *f*
host
Probability to amplify terminase is e_{ter}
HEX hits
FAM hits
Load targets randomly into a 1D panel with 765 "chambers"
(host FAM coordinates are derived from host HEX coordinates)
simulated panel
no co-localizations
Retrieve co-localized hit
no
terminase target present?
yes
If retrieved sample contains a host rRNA and a free floating target then increase N_{false} by 1.
Also calculate p_F for this panel.
Repeat until $N_{retrievals}$ are completed
Calculate average value of p_F.
Estimate N_{false} and p_F based on Eq. S3 and Eq. S5
Next Monte Carlo iteration

Figure S8. Schematic diagram of a Monte Carlo simulation of microfluidic array loading and sampling. See supporting text for further details.

SUPPORTING TABLES

Table S1. Abundance of homologs of known viral genes in the higher termite metagenome. This table describes the number (or *abundance*, see definition in Materials and methods) of metagenome gene objects in the higher termite metagenome that were homologous to the indicated viral phage genes (E value ≤ 0.001, *abundance* ≥ 10 metagenome gene objects). This list constitutes the most abundant viral-specific genes in the metagenome (i.e., viral genes related to building a virion), using the viral RefSeq database v37 (*32*) as a reference for known viral genes. The two highlighted rows are the portal protein and terminase protein that were found to have homologs in the ZAS prophage-like elements.

Phage	Accession #	Gene function	# of homologous metagenome gene objects
Enterobacteria phage N15	NP_046908.1	major tail protein	56
Lactobacillus phage phig1e	NP_695158.1	minor capsid protein	49
Bacillus phage 0305phi8-36	YP_001429638.1	baseplate hub protein	36
Salmonella phage Fels-1	YP_001700571.1	putative bacteriophage major tail protei	27
Lactobacillus prophage Lj965	NP_958579.1	putative terminase large subunit	25
Burkholderia phage phi644-2	YP_001111083.1	portal protein, HK97 family	23
Streptococcus phage P9	YP_001469206.1	terminase large subunit	22
Burkholderia phage BcepMu	YP_024702.1	putative portal protein	20
Clostridium phage phiC2	YP_001110720.1	terminase large subunit	19
Lactobacillus phage phiJL-1	YP_223885.1	large subunit terminase	18
Yersinia phage PY54	NP_892049.1	capsid protein	16
Bacillus phage B103	NP_690641.1	major head protein	14
Enterobacteria phage WV8	YP_002922822.1	putative tail protein	13
Pseudomonas phage MP22	YP_001469162.1	Mu-like prophage major head subunit	12
Enterobacteria phage Mu	YP_950582.1	major tail subunit	11
Streptococcus phage SMP	NP_050643.1	terminase large subunit	11
Burkholderia phage phiE255	NP_599050.1	putative portal protein	10
Enterobacteria phage SfV	YP_001111202.1	tail protein	10

Table S2. Similarity analysis of the termite-associated terminase gene and portal protein gene with close homologs. The following table describes the result of a BLAST analysis of the large terminase subunit gene (411 aa in length) and the portal protein gene (396 aa in length) found in *T. primitia's* prophage-like element with close homologs. Close homologs were searched for in: (1) the larger prophage-like element present in the genome of *T. azotonutricium*, (2) the metagenome of the hindgut of a *Nasutitermes sp*. termite, and (3) the viral RefSeq database v37 (*32*). The table demonstrates that the alleles of the termite-associated phage genes were very similar to each other and highly divergent from their closest homologs found among all currently known viral genomes. Alignments were performed on the amino acid sequences.

Large terminase subunit gene	**% identity***	**% similarity***	**Gaps***	**E value**
	T. azotonutricium			
	363/411 (89%)	385/411 (94%)	4/411 (0%)	0
	Higher termite metagenome			
	317/407 (78%)	359/407 (89%)	5/407 (1%)	0
	Viral RefSeq database *(Lactobacillus johnsonii* prophage Lj771)			
	107/415 (25%)	177/415 (42%)	64/415 (15%)	4.00E-19
Portal protein	**% identity**	**% similarity**	**Gaps**	**E value**
	T. azotonutricium			
	309/382 (81%)	348/382 (92%)	3/382 (0%)	0
	Higher termite metagenome			
	273/392 (70%)	324/392 (83%)	11/392 (2%)	1.00E-167
	Viral RefSeq database *(Streptomyces* phage mu1/6)			
	99/382 (25%)	156/382 (40%)	52/382 (13%)	6.00E-17

*Numbers divided by a forward slash correspond to the number of amino acids in each pair-wise alignment ("identity/total", "similarity/total", and "gaps/total", depending on the column).

Table S3. Sample collection and analysis information. Collection dates, collection sites, and dPCR execution dates for the *R. hesperus* specimens. The different colonies were on average 120 meters apart. The microfluidic array and colony labels noted here were used to label the samples throughout this report.

Chip ID	Chip designation in trees	Termite collection date	Date of chip execution	Colony	GPS coordinates
1151065015	A	11/13/2008	11/25/2008	1	34 19' 25.6"N/ 118 0' 17.9"W
1151065011	B	5/27/2009	5/29/2009	2	34 19' 31"N/118 00' 20.8"W
1151065010	C	5/27/2009	6/6/2009	2	"
1151065012	D	5/27/2009	6/7/2009	2	"
1151065017	E	5/27/2009	6/21/2009	3	34 19' 28"N/118 00' 17.5"W
1151065018	F	5/27/2009	6/22/2009	3	"
1151065019	G	5/27/2009	6/24/2009	3	"

Table S4. Estimated evolutionary distance between bacterial host SSU rRNA phylotypes. The number of base substitutions per site from averaging over all sequence pairs within and between host groups is shown. With the exception of samples A7 and A9 (that were composed of 784 and 810 nucleotides, respectively) the SILVA (*50*)-based alignment contained 898 unambiguous nucleotides. Distances were calculated using the Jukes-Cantor (*98*) nucleotide substitution model in MEGA4 (*58*). The number of repetitions appearing in Table 1 are based on an Operational Taxonomical Unit (OTU) cutoff of 2% assigned by DOTUR (*41*) with the furthest neighbor sequence assignment method. The next significant OTU cutoff was 2.5% adding a more divergent member (B4) to host I, however due to the larger divergence and single instance of this event it cannot be statistically validated and therefore it was not included in this analysis. The distance matrix used by DOTUR was based on the above alignment and calculated in ARB (*51*) using the Jukes-Cantor substitution model. Each bacterial host was less than 0.9% divergent on average. The maximum divergence was observed between host III and ZAS-9 where the corrected evolutionary distance across their deduced rRNAs was measured to be 9.3%.

	Host I *(n=13)*	**Host II** *(n=8)*	**Host III** *(n=4)*	**Host IV** *(n=3)*	**ZAS-2** *(n=1)*	**ZAS-9** *(n=1)*
Host I	0.0084					
Host II	0.0822	0.0083				
Host III	0.0685	0.0544	0.005			
Host IV	0.0817	0.0841	0.087	0.0075		
ZAS-2	0.0396	0.0678	0.06	0.0712	-	
ZAS-9	0.073	0.086	0.0933	0.0865	0.0603	-

Table S5. Retrieved *Treponema* phylotypes from the microfluidic arrays

OTU (3.1%)	# species (ref lib)	Reference library sequences	Co-localization sequences	# species (co-loc)
REP1	23	16S_F13,16S_F22,16S_F29,16S_F43,16S_F56,16S_F77,16S_F82,16S_F83,16S_F92,16S_F81,16S_G9,16S_G14,16S_G15,16S_G17,16S_G28,16S_G32,16S_G49,16S_G71,16S_G74,16S_G78,16S_F69,16S_G86,16S_G88	-	-
REP2	8	16S_F3,16S_F5,16S_F12,16S_F14,16S_F21,16S_F88,16S_G60,16S_G73	-	-
REP3	7	16S_F26,16S_F40,16S_F94,16S_F100,16S_G80,16S_G83,16S_G30	-	-
REP4	5	16S_F39,16S_F63,16S_F71,16S_G42,16S_G50	A1_1,A3_1,A10_1,A11_1,A14_1,B1_2,B4_2,C1_2,C2_2,G1_3,E1_3,F1_3,G3_3,G5_3	14
REP5	4	16S_F33,16S_F47,16S_F61,16S_G91	-	-
REP6	3	16S_F68,16S_F79,16S_G29	-	-
REP7	2	16S_G3,16S_G24	A4_1,A5_1,A7_1,A9_1,A12_1,A13_1,B2_2,E2_3	8
REP8	2	16S_F8,16S_G63	-	-
REP9	2	16S_F52,16S_G72	A15_1	1
REP10	2	16S_F75,16S_G81	-	-
REP11	2	16S_G16,16S_G11	-	-
REP12	2	16S_G25,16S_G35	D2_2	1
REP13	1	16S_G41	A6_1,F2_3,G2_3,G4_3	4
REP14	1	16S_F86	A2_1,A8_1,B3_2	3
REP15	1	16S_F16	-	-
REP16	1	16S_F24	-	-
REP17	1	16S_F28	-	-
REP18	1	16S_F84	A16_1	1
REP19	1	16S_F93	-	-
REP20	1	16S_F95	-	-
REP21	1	16S_F23	E3_3	1
REP22	1	16S_G20	-	-
REP23	1	16S_G31	-	-
REP24	1	16S_G43	-	-
REP25	1	16S_G53	-	-
REP26	1	16S_G55	A18_1,B5_2	2
REP27	1	16S_G95		
REP28	1	16S_G36	G6_3	1
REP29	-	-	C3_2	1
REP30	-	-	C4_2	1
REP31	-	-	G7_3	1
-	-	-	ZAS2	1
-	-	-	ZAS9	1
total	***78***			***39***

All reference library sequences (*n*=118; 876 ± 71 bp SD) were initially classified with RDB (*99*) and *Treponema* phylotypes (66.1%, *n*=78 with 99–100% confidence) were subsequently aligned by the SILVA incremental aligner SINA (*50*). A distance matrix was calculated in ARB (*51*) for the 78 reference library *Treponema* species, the 39 colocalized *Treponema* species, and ZAS-2 and ZAS-9 (*n*=119). Note that REP4 was colocalized 14 times, however one of these colocalizations, B4, was more divergent than the other ribotypes of this group (see Table S4) and was therefore not regarded as a repeated colocalization of host I in Table 1 and Table S4. The distance matrix was calculated based on 780 unambiguous nucleotides (with the exception of A7, A9, rF79, rG41, rG53 that were in the range of 624–767 nucleotides) using the Jukes-Cantor (*98*) method. Operational taxonomical units (OTUs) were then determined by DOTUR (*41*) based on the furthest neighbor sequence assignment method using an OTU cutoff of 3.1%. This cutoff is slightly higher than the OTU cutoff used to identify the repeated colocalizations (2%) in Fig. 2 in order to make the statistical test for repeated colocalization more stringent. REPs corresponding to putative bacterial hosts are highlighted in gray. All *Treponema* sequences were also screened with Bellerophon v3 (*48*) on Greengenes (*49*) for chimeras and were found to be negative. The remaining phyla indentified by RDB to be present in the reference library were Proteobacteria (13.6%, 100% confidence), Firmicutes (6.8%, Clostridia 53–100% confidence), Tenericutes (5.9%, Mycoplasmataceae with 77–90% confidence), Bacteroidetes (3.4%, 100% confidence), Actinobacteria (3.4%, 100% confidence) and Planctomycetes (0.8%, 100% confidence). All these phyla have been observed previously in SSU rRNA libraries of *Reticulitermes speratus* (*21*). However, from the number of rRNA targets observe in the no-template-control panels we anticipate that background amplification (see Materials and methods) should contribute to 34.1 ± 18.4% SD of the reference library sequences due to sparse loading of the panels (increasing the fraction of background amplification products). Based on retrieval of rRNA sequences from the no-template-control panel (not shown) we expect the major contributor to this fraction to be bacteria from the Proteobacteria phylum. The finding that free living prokaryotes in the termite hindgut are dominated by spirochetes is consistent with electron microscope observations showing that spirochetes can account for over 50% of the gut microbes in some termites (*100*). The absence of bacteria belonging to the TG-1 phylum (*101*) is an indication that large flagellates were successfully filtered out by the 5 μm pre-filter and did not lyse in this process (see Methods).

Table S6. Selection pressure analysis of the terminase gene. Codon-based test of purifying (negative) selection for hosts I through IV excluding suspected recombinant sequences (B1, B2, and A13ii). d_S and d_N are the number of synonymous and nonsynonymous substitutions per number of synonymous and nonsynonymous sites respectively obtained from averaging over all sequence pairs within a given group. d_S and d_N were calculated by various methods: **NG86**—Nei-Gojobori method (*102*) with the Jukes-Cantor (*98*) nucleotide substitution model, **Modified NG86**—Modified Nei-Gojobori method (*103*) with the Jukes-Cantor nucleotide substitution model, **LWL85**—Li-Wu-Luo method (*104*), **PBL85**—Pamilo-Bianchi-Li method (*105*), and **Kumar**—Kumar method (*106*). For the modified NG86 method, the ratio of transitional to transversional distances per site (R) was calculated by averaging over all sequence pairs within each group using the 3rd codon position based on the Kimura 2-parameter method (*97*). All results are based on the pairwise analysis of 235 unambiguous codon positions without gaps. Standard error estimates were obtained by a bootstrap procedure with 1000 replicates. The distribution of the test statistic (*D*) is approximated to be normal since the number of nucleotides contributing to d_S and d_N were sufficiently large (>10), allowing to test the null hypothesis using a one-tailed (Z > 0) Z test (*106*). The P value (one-tailed Z test) for observing Z > 0 ($d_S > d_N$) by chance is shown in the table. Z is shown to be greater than zero in a statistically significant manner ($P < 10^{-7}$ for hosts I–III and $P < 0.025$ for host IV) indicating negative selection was statistically significant. n/c denotes cases in which it was not possible to estimate evolutionary distances. All analyses were carried out with MEGA4 (*58*).

Host	Method	d_S (± S.E.)	d_N (± S.E.)	d_N/d_S	$D = d_S - d_N$ (± S.E.)	Z = D/std(D)	P value
I	NG86 (R=0.5)	0.57 ± 0.08	0.04 ± 0.01	0.08	0.53 ± 0.07	7.58	1.7E-14
(*n* =12)	Modified NG86 (R=2.02)	0.33 ± 0.03	0.05 ± 0.01	0.15	0.28 ± 0.03	8.94	0.0E+00
	LWL85	0.49 ± 0.06	0.04 ± 0.01	0.09	0.45 ± 0.06	7.52	2.8E-14
	PBL93	0.44 ± 0.05	0.04 ± 0.01	0.10	0.40 ± 0.05	7.42	5.7E-14
	Kumar	0.37 ± 0.04	0.04 ± 0.01	0.12	0.32 ± 0.04	8.28	0.0E+00
II	NG86 (R=0.5)	1.50 ± 0.12	0.17 ± 0.02	0.11	1.34 ± 0.12	11.18	0.0E+00
(*n* =9)	Modified NG86 (R=1.44)	0.99 ± 0.07	0.18 ± 0.02	0.18	0.81 ± 0.07	11.32	0.0E+00
	LWL85	1.48 ± 0.11	0.17 ± 0.02	0.11	1.31 ± 0.11	11.69	0.0E+00
	PBL93	1.49 ± 0.10	0.17 ± 0.02	0.11	1.32 ± 0.10	13.33	0.0E+00
	Kumar	1.14 ± 0.08	0.16 ± 0.02	0.14	0.97 ± 0.08	12.25	0.0E+00
III	NG86 (R=0.5)	0.72 ± 0.13	0.06 ± 0.01	0.08	0.66 ± 0.12	5.35	4.3E-08
(*n* =4)	Modified NG86 (R=1.80)	0.50 ± 0.06	0.06 ± 0.01	0.12	0.44 ± 0.06	6.92	2.2E-12
	LWL85	0.70 ± 0.09	0.05 ± 0.01	0.08	0.64 ± 0.10	6.75	7.2E-12
	PBL93	0.62 ± 0.09	0.06 ± 0.01	0.09	0.56 ± 0.09	6.26	1.9E-10
	Kumar	0.55 ± 0.07	0.05 ± 0.01	0.10	0.50 ± 0.08	6.63	1.7E-11
IV	NG86 (R=0.5)	n/c ± n/c	0.19 ± 0.02	n/c	n/c ± n/c	n/c	n/c
(*n* =3)	Modified NG86 (R=1.97)	1.53 ± 0.20	0.21 ± 0.03	0.14	1.32 ± 0.19	6.76	6.8E-12
	LWL85	2.30 ± 1.06	0.20 ± 0.09	0.09	2.10 ± 0.99	2.11	1.7E-02
	PBL93	1.65 ± 0.82	0.20 ± 0.09	0.12	1.45 ± 0.73	1.98	2.4E-02
	Kumar	1.94 ± 0.62	0.17 ± 0.07	0.09	1.76 ± 0.57	3.09	9.9E-04

Table S7. Similar terminase sequences associated with different bacterial hosts. Terminase alleles associated with different bacterial hosts having less than 10% difference between their nucleotide sequences.

Sequence 1	Sequence 2	% p-distance (705 bp)
A1_1 (host I)	A8_1 (host IV)	0
G1_3 (host I)	A5_1 (host II)	3
B1_1* (host I)	A9ii_1 (host II)	6.5

*Identified by RDP3 as a recombination between A9ii_1 (host II) and A1_1 (host I). See also Fig. S5.

Table S8. P values for the P Test comparing terminase alleles by bacterial host. The P Test (*34*) estimates the similarity between communities as the number of parsimony changes that would be required to explain the distribution of sequences between the different samples in the tree (samples here were grouped by bacterial host). The P value is the fraction of trials in which the true tree requires fewer changes than trees in which the sample assignments have been randomized (*123*). The P test was implemented in Fast UniFrac (*35*) selecting the "P Test Significance" option, comparing "Each pair of samples" using *n*=1000 random permutations. The analysis was performed on the phylogenetic tree in Fig. 2 applying midpoint rooting. P values shown have been corrected for multiple comparisons using the Bonferroni correction.

	Host II	Host III	Host IV
Host I	≤0.001	≤0.001	0.024
Host II	-	0.018	1
Host III	-	-	0.204

Table S9. P values for the P Test comparing terminase alleles by colonies. Samples here were grouped by termite colony. P values shown have been corrected for multiple comparisons using the Bonferroni correction. *n*=1000 random permutations were used to calculate P Values. See Table S8 for further details.

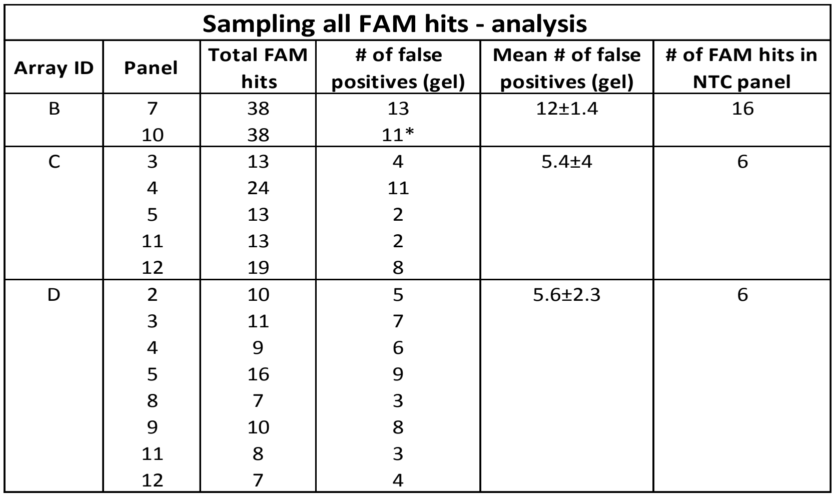
	Colony 2	Colony 3
Colony 1	0.399	0.927
Colony 2	-	0.537

Table S10. Sequences analyzed in this study. Accession numbers of the uncultured treponemes associated with phage host I through IV in Fig. 2 were AF068338, AB192197, AB192140, and AB192202, respectively.

Clone ID	Termite/bacterium species	Location/Source	Method	Accession (NCBI/JGI)	Figure	Reference
Terminase gene – isolates						
ZAS2i	*Z. angusticollis /T. primitia*	California	Isolate		2,S2,S3,S5	this study
ZAS2ii	*Z. angusticollis /T. primitia*	California	Isolate		S2,S5	this study
ZAS9i	*Z. angusticollis /T. azotonutricium*	California	Isolate		S2,S5	this study
ZAS9ii	*Z. angusticollis /T. azotonutricium*	California	Isolate		2,S2,S5	this study
Terminase gene – colocalization						
A1_1	*Reticulitermes hesperus*	California	Digital PCR	HQ202808	2,S2,S5	this study
A3_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187752	2,S2,S5	this study
A10_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187760	2,S2,S5	this study
A11_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187761	2,S2,S5	this study
A14_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187765	2,S2,S5	this study
B1_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187766	S2,S5	this study
C1_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187769	2,S2,S5	this study
C2_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187770	2,S2,S5	this study
E1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187771	2,S2,S5	this study
F1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187774	2,S2,S5	this study
G1_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187776	2,S2,S5	this study
G3_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187778	2,S2,S5	this study
G5_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187780	2,S2,S5	this study
A4_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187753	2,S2,S5	this study
A5_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187754	2,S2,S5	this study
A7_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187756	2,S2,S5	this study
A9i_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187758	2,S2,S5	this study
A9ii_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187759	2,S2,S5	this study
A12_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187762	2,S2,S5	this study
A13i_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187763	2,S2,S5	this study
A13ii_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187764	S2,S5	this study
B2_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187767	S2,S5	this study
E2i_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187772	2,S2,S5	this study
E2ii_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187773	2,S2,S5	this study
A6_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187755	2,S2,S5	this study
F2_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187775	2,S2,S5	this study
G2_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187777	2,S2,S5	this study
G4_3	*Reticulitermes hesperus*	California	Digital PCR	HQ187779	2,S2,S5	this study
A2_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187751	S2,S5	this study
A8_1	*Reticulitermes hesperus*	California	Digital PCR	HQ187757	S2,S5	this study
B3_2	*Reticulitermes hesperus*	California	Digital PCR	HQ187768	S2,S5	this study
Terminase gene – close relatives						
H1	*Nasutitermes sp.*	Costa Rica	Metagenome	2004118157	2,S2,S5	(*22*)
H2	*Nasutitermes sp.*	Costa Rica	Metagenome	2004126816	2,S2,S5	(*22*)
H3	*Nasutitermes sp.*	Costa Rica	Metagenome	2004144277	2,S2,S5	(*22*)
H4	*Nasutitermes sp.*	Costa Rica	Metagenome	2004144007	2,S2,S5	(*22*)
H5	*Nasutitermes sp.*	Costa Rica	Metagenome	2004132071	S2,S5	(*22*)
H6	*Nasutitermes sp.*	Costa Rica	Metagenome	2004107522	2,S2,S5	(*22*)
H7	*Nasutitermes sp.*	Costa Rica	Metagenome	2004111244	2,S2,S5	(*22*)
H8	*Nasutitermes sp.*	Costa Rica	Metagenome	2004124547	2,S2,S5	(*22*)
H9	*Nasutitermes sp.*	Costa Rica	Metagenome	2004134785	2,S2,S5	(*22*)
H10	*Nasutitermes sp.*	Costa Rica	Metagenome	2004136622	S2,S5	(*22*)
Terminase gene – non-termite-related						

unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088874	S4	(*21*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088890	S4	(*21*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088878	S4	(*21*)
unc Trep sp.	*Reticulitermes speratus*	Asia	PCR	AB088909	S4	(*21*)
unc Trep clone HsDiSp314	*Hodotermopsis sjoestedti*	Asia	PCR	AB032005	S4	(*112*)
SSU rRNA gene – non-termite-related						
Treponema vincentii (D2A-2)		Oral cavity	isolate	AY119690	S4	(*113*)
Treponema denticola (ATCC 35405)		Oral cavity	isolate	AE017226	S4	(*114*)
Treponema pallidum (Nichols)		Human genital tract	isolate	AE000520	S4	(*115*)
Treponema zioleckii (kT)		Sheep rumen	isolate	DQ065758	S4	(*116*)
Treponema socranskii (socranskii)		Oral cavity	isolate	AF033306	S4	(*117*)
Treponema succinifaciens		Pig colon	isolate	M57738	S4	(*118*)
Brevinema andersonii		Shrews and mice	isolate	L31543	S4	(*119*)
Borrelia burgdorferi (DK7)		Ticks, deer and humans	isolate	X85195	S4	(*120*)
Spirochaeta aurantia (M1)		Fresh water	isolate	AY599019	S4	(*121*)
Escherichia coli K-12 MG1655		-	isolate	U00096	S4	(*122*)

Table S11. Analysis of all FAM hits for a number of microfluidic array panels. For several microfluidic array panels, all chambers exhibiting amplification in the FAM fluorescence channel were retrieved, post-amplified and analyzed by agarose gel electrophoresis. In this table we show the total number of chambers that exhibited FAM fluorescence on the given panel (“Total FAM hits”), the number of false positives based on analysis by agarose gel electrophoresis (“# of false positives”), the mean number of false positives per array (“Mean # of false positives”), and the average number of chambers that exhibited FAM fluorescence in the no-template-control panel on the same array (“# of FAM hits in NTC panel”). The mean number of false positive hits agrees well with the number of hits in the corresponding no-template-control panel indicating the latter is a good predictor of the former. See supporting text for further details.



Sampling all FAM hits - analysis					
Array ID	**Panel**	**Total FAM hits**	**# of false positives (gel)**	**Mean # of false positives (gel)**	**# of FAM hits in NTC panel**
B	7	38	13	12±1.4	16
	10	38	11*		
C	3	13	4	5.4±4	6
	4	24	11		
	5	13	2		
	11	13	2		
	12	19	8		
D	2	10	5	5.6±2.3	6
	3	11	7		
	4	9	6		
	5	16	9		
	8	7	3		
	9	10	8		
	11	8	3		
	12	7	4		

* 3 retrievals were not tested due to an experimental problem

Table S12. Definition of variables used in the microfluidic array statistical model. See supporting text for further details.

Variable	Definition	Estimation method
X	Number of FAM hits per panel	Measured
Y	Number of HEX hits per panel	Measured
I	Number of wells per panel with both a FAM hit and a HEX hit (i.e., colocalization)	Measured
noise	Number of FAM hits that are due to spurious amplification	Measured
f_S	Frequency of ribotype ***S*** on the chip	Measured
$\varepsilon_{ter/16S}$	Terminase/16S primer efficiency	Measured
X_f	Number of non-colocalized FAM events	$X_f = X - I$
p_{ter}	The probability that a given well will contain a free floating terminase target	Eq. S1
I_{S}	Average number of free floating terminase targets to colocalize with a *particular* 16S rRNA ribotype ***S*** on a panel	Eq. S2a
$I_{\text{all 16S-ter}}$	Average number of any terminase target to colocalize with any 16S rRNA target on a panel	Eq. S2b
p_F	Probability that a successful retrieval from a panel contains a particular ribotype ***S*** and any terminase gene by chance	Eq. S3
X_T	Sum of the total number of free floating terminase targets and spurious targets	Eq. S4
N_{false}	Expected number of false colocalizations in the dataset	Eq. S5
p_S	Probability that a successful retrieval will contain host ***S***	Eq. S6

Table S13. Statistics for all sampled panels. This table lists for each ribotype in Fig. 2 the panel from which the ribotype was retrieved, the number of FAM hits *X* on that panel, the number of HEX hits *Y* on that panel, their intersection *I*, the number of FAM hits found in the no-target-control-panel for the microfluidic array containing the given panel (*noise*), the frequency of this host in the reference rRNA library, f_s (based on Table S5), the estimated probability for false colocalization p_F (Eq. S3), and the P value (one-tailed test, *n*=41) for each host for obtaining at least the number of observed colocalizations by chance (based on the data in Table 1). The statistical test to determine the P value is explained in the supporting text. Chip analysis was performed using the Fluidigm Digital PCR Analysis software v.2.1.1 with the linear baseline correction. See supporting text for further details.

#	Retrieval ID (*n* =41)	Host	chip	panel	X (FAM)	Y (HEX)	I	*noise* (FAM)	X_T-*noise*	p_{ter}	$I_{all16S\text{-}ter}$	f_S	p_F	P value (*n* =41)
1	A1_1	I	A	3	22	38	2	15	6.0	7.9E-03	1.3	4.2%	7.77E-03	5.45E-18
2	A3_1	I	A	5	33	66	8	15	12.4	1.6E-02	6.7			
3	A10_1	I	A	8	40	59	12	15	15.3	2.0E-02	10.8			
4	A11_1	I	A	9	34	46	9	15	11.6	1.5E-02	8.1			
5	A14_1	I	A	10	30	46	11	15	5.2	6.8E-03	10.1			
6	B1_2	I	B	10	42	52	5	20	19.7	2.6E-02	3.6			
7	C1_2	I	C	11	13	55	3	6	4.8	6.2E-03	2.6			
8	C2_2	I	C	5	13	69	4	6	3.9	5.1E-03	3.5			
9	E1_3	I	E	2	14	21	2	5	7.3	9.6E-03	1.9			
10	F1_3	I	F	3	22	32	2	7	13.9	1.8E-02	1.7			
11	G1_3	I	G	3	12	51	4	6	2.6	3.4E-03	3.6			
12	G3_3	I	G	8	17	33	2	6	9.7	1.3E-02	1.7			
13	G5_3	I	G	11	14	26	1	6	7.5	9.7E-03	0.8			
14	A4_1	II	A	6	54	79	10	15	34.1	4.5E-02	8.5	1.7%	3.11E-03	7.63E-13
15	A5_1	II	A	6	54	79	10	15	34.1	4.5E-02	8.5			
16	A7_1	II	A	8	40	59	12	15	15.3	2.0E-02	10.8			
17	A9_1	II	A	8	40	59	12	15	15.3	2.0E-02	10.8			
18	A12_1	II	A	10	30	46	11	15	5.2	6.8E-03	10.1			
19	A13_1	II	A	10	30	46	11	15	5.2	6.8E-03	10.1			
20	B2_2	II	B	10	42	52	5	20	19.7	2.6E-02	3.6			
21	E2_3	II	E	2	14	21	2	5	7.3	9.6E-03	1.9			
22	A6_1	III	A	7	40	66	8	15	20.0	2.6E-02	6.7	0.9%	1.55E-03	5.65E-07
23	F2_3	III	F	8	21	34	6	7	8.7	1.1E-02	5.7			
24	G2_3	III	G	4	20	53	3	6	12.3	1.6E-02	2.6			
25	G4_3	III	G	10	19	36	2	6	11.8	1.5E-02	1.7			
26	A2_1	IV	A	4	46	129	17	15	19.9	2.6E-02	14.5	0.9%	1.55E-03	3.83E-05
27	A8_1	IV	A	8	40	59	12	15	15.3	2.0E-02	10.8			
28	B3_2	IV	B	10	42	52	5	20	19.7	2.6E-02	3.6			
29	A15_1	-	A	4	46	129	17	15	19.9	2.6E-02	14.5	-	-	-
30	A16_1	-	A	5	33	66	8	15	12.4	1.6E-02	6.7	-	-	-
31	A17_1	-	A	10	30	46	11	15	5.2	6.8E-03	10.1	-	-	-
32	A18_1	-	A	11	27	84	7	15	7.5	9.8E-03	5.4	-	-	-
33	B5_2	-	B	7	46	53	11	20	17.6	2.3E-02	9.6	-	-	-
34	B4_2	-	B	7	46	53	11	20	17.6	2.3E-02	9.6	-	-	-
35	C3_2	-	C	11	13	55	3	6	4.8	6.2E-03	2.6	-	-	-
36	C4_2	-	C	11	13	55	3	6	4.8	6.2E-03	2.6	-	-	-
37	D1_2	-	D	4	9	24	1	8	0.3	3.4E-04	0.7	-	-	-
38	D2_2	-	D	3	11	26	1	8	2.4	3.1E-03	0.7	-	-	-
39	E3_3	-	E	11	12	24	2	5	5.3	7.0E-03	1.8	-	-	-
40	G6_3	-	G	4	20	53	3	6	12.3	1.6E-02	2.6	-	-	-
41	G7_3	-	G	4	20	53	3	6	12.3	1.6E-02	2.6	-	-	-

REFERENCES

1. C. A. Suttle, *Nat. Rev. Microbiol.* **5**, 801 (2007).
2. M. B. Sullivan *et al.*, *Environ. Microbiol.* **12**, 3035 (2010).
3. D. Lindell *et al.*, *Nature* **449**, 83 (2007).
4. F. E. Angly *et al.*, *PLoS Biol* **4**, e368 (2006).
5. S. Williamson *et al.*, *PLoS ONE* **3**, e1456 (2008).
6. P. Hugenholtz, *Genome Biol.* **3**, S0003 (2002).
7. R. A. Edwards, F. Rohwer, *Nat. Rev. Microbiol.* **3**, 504 (2005).
8. E. A. Dinsdale *et al.*, *Nature* **452**, 629 (2008).
9. D. Kristensen, A. Mushegian, V. Dolja, E. Koonin, *Trends Microbiol.* **18**, 11 (2009).
10. A. F. Andersson, J. F. Banfield, *Science* **320**, 1047 (2008).
11. R. N. Zare, S. Kim, *Annu. Rev. Biomed. Eng.* **12**, 187 (2010).
12. E. A. Ottesen, J. W. Hong, S. R. Quake, J. R. Leadbetter, *Science* **314**, 1464 (2006).
13. Y. Marcy *et al.*, *Proc. Natl. Acad. Sci. USA* **104**, 11889 (2007).
14. L. Warren, D. Bryder, I. L. Weissman, S. R. Quake, *Proc. Natl. Acad. Sci. USA* **103**, 17807 (2006).
15. S. Dube, J. Qin, R. Ramakrishnan, *PLoS ONE* **3**, e2876 (2008).
16. F. Rohwer, R. Edwards, *J. Bacteriol.* **184**, 4529 (2002).
17. T. M. Rose *et al.*, *Nucleic Acids Res.* **26**, 1628 (1998).
18. Materials and methods are available as supporting material on *Science* Online.
19. Supporting text is available as supporting material on Science Online.
20. A. Tholen, B. Schink, A. Brune, *FEMS Microbiol. Ecol.* **24**, 137 (1997).
21. Y. Hongoh, M. Ohkuma, T. Kudo, *FEMS Microbiol. Ecol.* **44**, 231 (2003).
22. F. Warnecke *et al.*, *Nature* **450**, 560 (2007).
23. J. R. Leadbetter, T. M. Schmidt, J. R. Graber, J. A. Breznak, *Science* **283**, 686 (1999).
24. T. G. Lilburn *et al.*, *Science* **292**, 2495 (2001).
25. S. D. Moore, P. E. Prevelige Jr, *Curr. Biol.* **12**, R96 (2002).
26. V. B. Rao, M. Feiss, *Annu. Rev. Genet.* **42**, 647 (2008).
27. S. Chai *et al.*, *J. Mol. Biol.* **224**, 87 (1992).
28. K. Eppler, E. Wyckoff, J. Goates, R. Parr, S. Casjens, *Virology* **183**, 519 (1991).
29. S. Casjens, *Mol. Microbiol.* **49**, 277 (2003).
30. M. S. Mitchell, S. Matsuzaki, S. Imai, V. B. Rao, *Nucleic Acids Res.* **30**, 4009 (2002).
31. S. Sun *et al.*, *Cell* **135**, 1251 (2008).
32. K. D. Pruitt, T. Tatusova, D. R. Maglott, *Nucleic Acids Res.* **33**, D501 (2005).
33. Percent identity was measured across 235 unambiguous aligned amino acids.
34. A. P. Martin, *Appl. Environ. Microbiol.* **68**, 3673 (2002).
35. M. Hamady, C. Lozupone, R. Knight, *The ISME journal* **4**, 17 (2009).
36. S. R. Casjens *et al.*, *J. Bacteriol.* **187**, 1091 (2005).
37. N. Wolfe *et al.*, *Global Change & Human Health* **1**, 10 (2000).
38. L. Moore, G. Rocap, S. Chisholm, *Nature* **393**, 465 (1998).
39. J. R. Thompson *et al.*, *Appl. Environ. Microbiol.* **70**, 4103 (2004).
40. S. G. Acinas, R. Sarma-Rupavtarm, V. Klepac-Ceraj, M. F. Polz, *Appl. Environ. Microbiol.* **71**, 8966 (2005).
41. P. D. Schloss, J. Handelsman, *Appl. Environ. Microbiol.* **71**, 1501 (2005).
42. E. A. Ottesen, PhD thesis, California Institute of Technology (2008).

43. J. Austin, A. Szalanski, B. Cabrera, *Ann. Entomol. Soc. Am.* **97**, 548 (2004).
44. M. Ohkuma *et al.*, *Mol. Phylogenet. Evol.* **31**, 701 (2004).
45. K. Maekawa, N. Lo, O. Kitade, T. Miura, T. Matsumoto, *Mol. Phylogenet. Evol.* **13**, 360 (1999).
46. H. Liu, A. T. Beckenbach, *Mol. Phylogenet. Evol.* **1**, 41 (1992).
47. K. E. Ashelford, N. A. Chuzhanova, J. C. Fry, A. J. Jones, A. J. Weightman, *Appl. Environ. Microbiol.* **71**, 7724 (2005).
48. T. Huber, G. Faulkner, P. Hugenholtz, *Bioinformatics* **20**, 2317 (2004).
49. T. Z. DeSantis *et al.*, *Appl. Environ. Microbiol.* **72**, 5069 (2006).
50. E. Pruesse *et al.*, *Nucleic Acids Res.* **35**, 7188 (2007).
51. W. Ludwig, O. Strunk, R. Westram, L. Richter, H. Meier, *Nucleic Acids Res.* **32**, 1363 (2004).
52. D. Posada, *Mol. Biol. Evol.* **25**, 1253 (2008).
53. S. Guindon, O. Gascuel, *Syst. Biol.* **52**, 696 (2003).
54. K. Tamura, M. Nei, *Mol. Biol. Evol.* **10**, 512 (1993).
55. J. Felsenstein, *Phylogeny Inference Package (PHYLIP), version 3.6 a3* (2002).
56. W. M. Fitch, E. Margoliash, *Science* **155**, 279 (1967).
57. J. D. Thompson, D. G. Higgins, T. J. Gibson, *Nucleic Acids Res.* **22**, 4673 (1994).
58. K. Tamura, J. Dudley, M. Nei, S. Kumar, *Mol. Biol. Evol.* **24**, 1596 (2007).
59. L. Heath, E. van der Walt, A. Varsani, D. P. Martin, *J. Virol.* **80**, 11827 (2006).
60. M. Padidam, S. Sawyer, C. M. Fauquet, *Virology* **265**, 218 (1999).
61. J. M. Smith, *J. Mol. Evol.* **34**, 126 (1992).
62. D. Martin, E. Rybicki, *Bioinformatics* **16**, 562 (2000).
63. D. Posada, *Mol. Biol. Evol.* **19**, 708 (2002).
64. D. Posada, K. A. Crandall, *Proc. Natl. Acad. Sci. USA* **98**, 13757 (2001).
65. M. O. Salminen, J. K. Carr, D. S. Burke, F. E. McCutchan, *AIDS Res. Hum. Retroviruses* **11**, 1423 (1995).
66. M. Salminen, D. Marin, in *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing,* P. Lemey, M. Salemi, A. Vandamme, Eds. (Cambridge Univ. Press, Cambridge, ed. 2, 2010), pp. 519–548.
67. D. H. Huson, D. Bryant, *Mol. Biol. Evol.* **23**, 254 (2006).
68. N. Tao *et al.*, thesis, University of New Mexico (2005).
69. K. Strimmer, A. Von Haeseler, *Proc. Natl. Acad. Sci. USA* **94**, 6815 (1997).
70. H. Schmidt, A. von Haeseler, in *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing,* P. Lemey, M. Salemi, A. Vandamme, Eds. (Cambridge Univ. Press, Cambridge, ed. 2, 2010), pp. 181–209.
71. Y. Zhang *et al.*, *Nucleic Acids Res.* **31**, e123 (2003).
72. C. E. Corless *et al.*, *J. Clin. Microbiol.* **38**, 1747 (2000).
73. S. L. K. Pond, S. D. W. Frost, S. V. Muse, Bioinformatics **21,** 676 (2005)
74. S. L. K. Pond, S. V. Muse, in *The Phylogenetic Handbook: A Practical Approach to Phylogenetic Analysis and Hypothesis Testing,* P. Lemey, M. Salemi, A. Vandamme, Eds. (Cambridge Univ. Press, Cambridge, ed. 2, 2010), pp. 419–490.
75. S. V. Muse, B. S. Gaut, *Mol. Biol. Evol.* **11**, 715 (1994).
76. S. F. Altschul *et al.*, *Nucleic Acids Res.* **25**, 3389 (1997).
77. L. W. Black, *Bioessays* **17**, 1025 (1995).
78. H. Ackermann, *Adv. Virus Res.* **51**, 135 (1998).
79. S. R. Casjens, *Res. Microbiol.* **159**, 340 (2008).

80. J. G. Lawrence, G. F. Hatfull, R. W. Hendrix, *J. Bacteriol.* **184**, 4891 (2002).
81. M. A. Daw, F. R. Falkiner, *Micron* **27**, 467 (1996).
82. K. Nakayama *et al.*, *Mol. Microbiol.* **38**, 213 (2000).
83. Y. Michel-Briand, C. Baysse, *Biochimie* **84**, 499 (2002).
84. A. S. Lang, J. T. Beatty, *Proc. Natl. Acad. Sci. USA* **97**, 859 (2000).
85. A. S. Lang, J. T. Beatty, *Arch. Microbiol.* **175**, 241 (2001).
86. A. S. Lang, J. T. Beatty, *Trends Microbiol.* **15**, 54 (2007).
87. A. J. Davison, *Virology* **186**, 9 (1992).
88. E. V. Koonin, T. G. Senkevich, V. V. Dolja, *Biol. Direct* **1**, 29 (2006).
89. M. L. Baker, W. Jiang, F. J. Rixon, W. Chiu, *J. Virol.* **79**, 14967 (2005).
90. S. Kanamaru, K. Kondabagil, M. G. Rossmann, V. B. Rao, *J. Biol. Chem.* **279**, 40795 (2004).
91. A. Marchler-Bauer *et al.*, *Nucleic Acids Res.* **33**, D192 (2005).
92. R. C. Edgar, *BMC bioinformatics* **5**, 113 (2004).
93. T. G. Lilburn, T. M. Schmidt, J. A. Breznak, *Environ. Microbiol.* **1**, 331 (1999).
94. J. Breznak, J. Leadbetter, *Prokaryotes* **7**, 318 (2006).
95. B. J. Paster *et al.*, *J. Bacteriol.* **173**, 6101 (1991).
96. D. Bryant, V. Moulton, *Mol. Biol. Evol.* **21**, 255 (2004).
97. M. Kimura, *J. Mol. Evol.* **16**, 111 (1980).
98. T. Jukes, C. Cantor, *Mammalian protein metabolism* **3**, 21 (1969).
99. J. R. Cole *et al.*, *Nucleic Acids Res.* **37**, D141 (2009).
100. B. J. Paster *et al.*, *Appl. Environ. Microbiol.* **62**, 347 (1996).
101. W. Ikeda-Ohtsubo, M. Desai, U. Stingl, A. Brune, *Microbiology* **153**, 3458 (2007).
102. M. Nei, T. Gojobori, *Mol. Biol. Evol.* **3**, 418 (1986).
103. J. Zhang, H. F. Rosenberg, M. Nei, *Proc. Natl. Acad. Sci. USA* **95**, 3708 (1998).
104. W. H. Li, C. I. Wu, C. C. Luo, *Mol. Biol. Evol.* **2**, 150 (1985).
105. P. Pamilo, N. O. Bianchi, *Mol. Biol. Evol.* **10**, 271 (1993).
106. M. Nei, S. Kumar, *Molecular evolution and phylogenetics* (Oxford Univ. Press, New York, 2000).
107. E. S. Miller *et al.*, *Microbiol. Mol. Biol. Rev.* **67**, 86 (2003).
108. W. Reeve *et al.*, *Standards in Genomic Sciences* **2**, 77 (2010).
109. C. Prust *et al.*, *Nat. Biotechnol.* **23**, 195 (2005).
110. V. Kunin *et al.*, *Mol. syst. biol.* **4**, 198 (2008).
111. Y. Hongoh *et al.*, *Appl. Environ. Microbiol.* **71**, 6590 (2005).
112. T. Iida, M. Ohkuma, K. Ohtoko, T. Kudo, *FEMS Microbiol. Ecol.* **34**, 17 (2000).
113. A. M. Edwards, D. Dymock, M. J. Woodward, H. F. Jenkinson, *Microbiology* **149**, 1083 (2003).
114. R. Seshadri *et al.*, *Proc. Natl. Acad. Sci. USA* **101**, 5646 (2004).
115. C. M. Fraser *et al.*, *Science* **281**, 375 (1998).
116. M. Piknova *et al.*, *FEMS Microbiol. Lett.* **289**, 166 (2008).
117. B. J. Paster, F. E. Dewhirst, B. C. Coleman, C. N. Lau, R. L. Ericson, *Int. J. Syst. Bacteriol.* **48**, 713 (1998).
118. T. B. Stanton *et al.*, *Int. J. Syst. Bacteriol.* **41**, 50 (1991).
119. D. L. Defosse, R. C. Johnson, B. J. Paster, F. E. Dewhirst, G. J. Fraser, *Int. J. Syst. Bacteriol.* **45**, 78 (1995).
120. M. Theisen *et al.*, *J. Bacteriol.* **177**, 3036 (1995).

121. R. McLaughlin, D. M. Secko, C. J. Paul, A. M. Kropinski, *Can. J. Microbiol.* **50**, 967 (2004).
122. W. K. Kang, T. Icho, S. Isono, M. Kitakawa, K. Isono, *Mol. Gen. Genet.* **217**, 281 (1989).
123. C. Lozupone, M. Hamady, R. Knight, *BMC bioinformatics* **7**, 371 (2006).