Supplementary Materials for

Title: Synthesis of methylphosphonic acid by marine microbes: a source for methane in the aerobic ocean

Authors: William W. Metcalf[1,3]*, Benjamin M. Griffin[1†], Robert M. Cicchillo[1,2#], Jiangtao Gao[1,2], Sarath Chandra Janga[1], Heather A. Cooke[1,2], Benjamin T. Circello[1,3], Bradley S. Evans[1], Willm Martens-Habbena[4], David A. Stahl[4], and Wilfred A. van der Donk[1,2]*

*correspondence to: E-mail: metcalf@uiuc.edu, vddonk@illinois.edu

This PDF file includes:

Materials and Methods

Identification of a *N. maritimus* gene cluster for synthesis of Mpn.

The GenBank nonredundant protein database was searched using BLASTP with the PEP mutase (Ppm) sequence from *S. viridochromogenes* as the query (Accession # AAU00071). Among the significant hits was a protein from *Nitrosopumilus maritimus* (Accession # YP001581494). Examination of the chromosomal region surrounding the *N. maritimus ppm* gene revealed a number of genes encoding proteins similar to known phosphonate and polysaccharide biosynthetic enzymes (Fig. S1A, Table S1). Among these are proteins with significant homology to Ppm, a two-subunit phosphonopyruvate decarboxylase (PpdAB) and phosphonoacetaldehyde dehydrogenase (Pdh). Based on the experimentally validated functions of homologous enzymes (*10, 16, 17*), it is very likely that *N. maritimus* has the capacity to synthesize 2-hydroxyethylphosphonate (HEP) (Fig. 1B). Immediately adjacent to the putative HEP biosynthetic genes is an operon encoding a putative oxidoreductase, two putative sulfatases and a protein of the cupin superfamily, annotated as a DNA binding protein, that we designated this cupin-family protein MpnS.

The cupin superfamily is very diverse and includes numerous enzymes, in addition to DNA binding transcriptional regulators (*25*). Among these are two phosphonate biosynthetic enzymes: hydroxypropylphosphonate epoxidase (HppE) and hydroxyethylphosphonate dioxygenase (HepD), which are involved in synthesis of the antibiotic fosfomycin (*26*), and the herbicide phosphinothricin tripeptide (*27*), respectively. HepD and HppE catalyze unusual O_2-dependent transformations of similar phosphonate substrates (Fig. 1B). The two enzymes also share common secondary and tertiary structures, with HppE having a single cupin domain and HepD having two (*27, 28*). Despite these similarities, the enzymes carry out very different reactions. HppE catalyzes conversion of 2-hydroxypropylphosphonate (HPP) to the corresponding epoxide in a reaction that requires an exogenous electron donor, whereas HepD catalyzes the cleavage of the unactivated carbon-carbon bond of HEP to produce formic acid and hydroxymethylphosphonate without input of exogenous electrons. The fate of the molecular oxygen used in the two reactions also differs. In the HppE reaction, O_2 is reduced to water, whereas HepD incorporates one oxygen atom into each of the two cleavage products (*29*). Significantly, *N. maritimus* MpnS has weak homology to both HppE and HepD. HepD and MpnS align over their entirety suggesting that, like HepD, MpnS has two cupin folds. Alignment of the HppE, HepD and MpnS sequences also reveals conservation of several key residues involved in catalysis, including both histidines of the iron-binding 2-His-1-carboxylate facial triad and residues shown to be required for phosphonate recognition (Fig S2). Based on these similarities we suspected that MpnS would carry out an enzymatic transformation of a similar phosphonate substrate.

Construction of the MpnS expression vector

The *mpnS* gene was amplified by PCR using template DNA from *N. maritimus* cells lysed by heat. The forward primer included an *NdeI* restriction site and the first 22 bp of the *mpnS* gene. The reverse primer contained a *BamHI* site and the last 21 bp of the *mpnS* gene. The amplified gene product was treated with *NdeI*/*BamHI* and was subsequently ligated into pET-15b digested with the same enzymes. The construct was

verified by digestion and sequence analysis. The vector was subsequently used to transform *E. coli* Rosetta 2(DE3) cells for overexpression.

Overexpression and Purification of MpnS

A single colony was used to inoculate 50 mL of LB media containing chloramphenicol (25 μg/mL) and kanamycin (50 μg/mL). The starter culture was grown overnight at 37 ºC and was used to inoculate 3 L of the same media. This culture was grown at 37 ºC to an optical density of 0.6 at which time the flask was placed in a water bath at room temperature and allowed to cool for 15 min. The flask was then placed in a 30 ºC incubator and the culture was induced by the addition of 0.5 mM IPTG. After 5 hours, the cells were cooled in an ice bath for 10 min and subsequently harvested by centrifugation at 10,000*g*. Harvested cells were resuspended in lysis buffer (50 mM HEPES-KOH pH 7.5, 300 mM NaCl, 20 mM imidazole, 10% glycerol) and incubated with lysozyme (1 mg mL^{-1}) for 30 min at 4 ºC. Cells were lysed by two passages through a French pressure cell at a setting of 20,000 psi. The cell debris was removed by centrifugation at 35000*g* for 45 min. Recombinant MpnS was purified by immobilized metal affinity chromatography (IMAC) using a Ni-NTA matrix (Qiagen) and was buffer exchanged into 50 mM HEPES-KOH pH 7.5, 300 mM NaCl, 10% glycerol by gel filtration as previously described for HEPD (*27*). The enzyme was frozen in 200 μL aliquots and stored at -80 °C until ready for use. The protein concentration was determined from a theoretical extinction coefficient of 64,540 M^{-1} cm^{-1}.

In vitro reconstitution and assay of MPnS Activity

Anaerobic activation of MpnS was achieved using the protocol established for HEPD (*27*). Briefly, in an anaerobic chamber, MPnS (200 μM) was treated with one molar equivalent of $Fe(II)(NH_4)_2(SO_4)_2$ for 20 min in a final volume of 620 μL. The solution was subsequently brought outside the glove box and combined with the remaining assay components. The MPnS reaction was performed in a final volume of 600 μL and contained 50 mM HEPES-KOH pH 7.5, 2 mM 1-^{13}C-HEP, and 100 μM MPnS. Additionally, the reaction buffer was purged with $O_{2(g)}$ on ice to increase the amount of dissolved oxygen. The reaction was initiated by the addition of 1-^{13}C-HEP and allowed to proceed at room temperature for 1 hour at which time the protein was removed by passing through an Amicon Ultra concentrator. The flow-through was analyzed by LC–MS as previously described for HEPD (*27*) (Fig S3). APCI ionization was performed in positive mode and chromatographic separations were achieved using a 150 x 4.6 mm Synergi C18 Fusion-RP column with a 4 μm particle size (Phenomenex). MPn has a retention time of approximately 4.7 min at a flow rate of 0.5 mL min^{-1} using 0.1% formic acid as the mobile phase.

NMR analysis of enzyme reaction products

^{1}H-decoupled ^{31}P-NMR analyses were performed as previously described (*27*) on a Varian Inova 600 spectrometer equipped with a 5-mm Varian 600DB AutoX probe tuned for phosphorus at 242.79 MHz (Fig. 1). Phosphorus chemical shifts are reported relative to an external standard of 85% phosphoric acid ($\delta = 0$). ^{13}C spectra were recorded on a Varian Unity 500 spectrometer, where carbon chemical shifts were referenced to an external standard of 0.1% tetramethylsilane in $CDCl_3$. For ^{13}C NMR analysis of the carbon containing products of MpnS catalysis, $O_{2(g)}$ purged water and MpnS (10 μM) were placed in an airtight vial sealed with a septum. Through the septum, 1 mM of both

2-[1-^{13}C]-HEP and 2-[2-^{13}C]-HEP were added to initiate the assay. After two hours, the 600 μL assay was quenched with 25 μL 1 M KOH and incubated for one hour to allow for mass transfer of CO_2 from the headspace to the aqueous phase and to equilibrate as bicarbonate. The protein was removed from the assay mixture by centrifuging through a pre-washed Amicon Ultra concentrator. The sample was analyzed by ^{13}C NMR spectroscopy with 20% D_2O (150.828 MHz).

Preparation and NMR characterization of synthetic standards methyl methylphosphonate and ethyl ethylphosphonate

A series of authentic standards were synthesized and analyzed by both ^{31}P-^{1}H Heteronuclear Multiple Bond Correlation (HMBC) and by regular ^{1}H and ^{31}P NMR spectroscopy to identify spectral acquisition and processing parameters for which the multiplicities of the peaks extracted from the ^{1}H and ^{31}P dimensions of the HMBC corresponded to the multiplicities observed in regular ^{1}H and ^{31}P spectra. The standards used for this purpose were methyl methylphosphonate (**1**) and ethyl ethylphosphonate (**2**) (Fig S5). Methyl methylphosphonate was generated as a standard to mimic the expected esterified methylphosphonate in the *N. maritimus* sample. Ethyl ethylphosphonate was generated to provide direct experimental evidence that the spectroscopic data of the material with a peak at 28.7 ppm in the ^{31}P NMR spectrum of the *N. maritimus* sample cannot be accounted for by phosphonates other than methylphosphonate.

Methyl methylphosphonate **1** was generated by removing a single methyl ester group from dimethyl methylphosphonate as previously described (*30, 31*). Dimethyl methyphosphonate (DMMPn) was purchased from Aldrich and 2 mL (18.4 mmol) was dissolved in a solution of 10% aqueous NaOH (5 mL). The solution was stirred at room temperature overnight. The pH was adjusted to 2 and unreacted starting material was extracted with dichloromethane. The product was concentrated under reduced pressure. Prior to analysis by NMR spectroscopy, the solution was adjusted to pH 7.5. ^{31}P NMR, ^{1}H decoupled (202 MHz, D_2O) δ: 28.7 (s). ^{31}P NMR, ^{1}H coupled (202 MHz, D_2O) δ: 28.2 (m). ^{1}H NMR, (500 MHz, D_2O) δ: 3.52 (-OC**H$_3$**, d, J=10.6 Hz), 1.26 (-PC**H$_3$**, d, J=16.3 Hz).

The HMBC spectrum of methyl methylphosphonate **1** shows correlations of the peak at 28.2 ppm in the ^{31}P dimension with ^{1}H resonances at 3.52 ppm (methyl ester) and 1.26 ppm (methyl group attached to phosphorus) (Fig S6, panel A). The correlation with the protons of the methyl group attached to phosphorus is stronger leading to a more intense crosspeak (see integration in Fig S6, panel B). The correlations, multiplicities, and coupling constants in the ^{1}H dimension extracted from the HMBC spectrum (Fig S6, panel B) agree very well with those observed in a regular ^{1}H spectrum of **1** (Fig S6, panel C).

Next ethyl ethylphosphonate **2** was prepared. This compound was chosen to illustrate that the crosspeak at 1.40 ppm in the phosphonate produced by *N. maritimus* is not associated with a methylene (CH_2) group. The compound was made by removing a single ethyl ester group from diethyl ethylphosphonate. Diethyl ethyphosphonate was purchased from Aldrich and 2 mL (12.3 mmol) was dissolved in a solution of 10% aqueous NaOH (5 mL). The solution was stirred at room temperature overnight. The pH was adjusted to 2 and unreacted starting material was extracted with dichloromethane. The product was concentrated under reduced pressure. Prior to analysis by NMR spectroscopy, the solution was adjusted to pH 7.5. ^{31}P NMR, ^{1}H decoupled (202 MHz,

D_2O) δ: 30.5 (s). ^{31}P NMR, ^{1}H coupled (202 MHz, D_2O) δ: 30.5 (m). ^{1}H NMR, (500 MHz, D_2O) δ: 3.75 (-OC**H**$_2$CH$_3$, m), 1.43 (-PC**H**$_2$CH$_3$, m), 1.11 (-OCH$_2$C**H**$_3$, t, J=7.1 Hz), 0.90 (-PCH$_2$C**H**$_3$, m) (Fig S7).

The NMR spectra of ethyl ethylphosphonate show a number of important differences from the spectra of methyl methylphosphonate (Figs S6 & S7). First, the resonance derived from the CH_2 group attached to phosphorus in ethyl ethylphosphonate shows a much more complex splitting pattern than the methyl group attached to phosphorus in methyl methylphosphonate. This increased complexity is the result of couplings to the β-methyl group in addition to the coupling to the ^{31}P nucleus. Hence, the splitting pattern of the α-protons (protons attached to the carbon that is directly attached to phosphorus) is very diagnostic: if a doublet is observed it shows that the α-proton only couples to the ^{31}P nucleus and that no β-protons are present. This conclusion is further corroborated by the observation that the β-protons show up in the HMBC spectrum of **2** (Fig S7, panels A and B), but that such β-protons would not be observed in any analogs of methylphosphonate (Fig S6).

NMR Spectroscopy of synthetic standards

All NMR spectra were recorded at 25 ºC on a Varian Inova 600 MHz spectrometer (600 MHz for ^{1}H and 243 MHz for ^{31}P) equipped with an AutoX Dual Broadband probe with a ProTune accessory and a Z-gradient capability. Proton chemical shifts are reported in δ values relative to an external standard of 0.1% tetramethylsilane in D_2O and phosphorus shifts are reported in δ values relative to an external standard of 85% phosphoric acid. The ^{1}H-^{31}P gHMBC (gradient Heteronuclear Multiple-Bond Correlation) spectra were collected after optimization of long-range coupling of proton-phosphorus at 18 Hz. An acquisition time of 1 s was used to obtain better spectral resolution. A total of 200 complex data points was collected in the F1 dimension. The spectrum was processed with a 90 degree shifted sine-bell square window function in MestReNova 7 software.

Cultivation of *N. maritimus*

Cultures of *Candidatus* "*Nitrosopumilus maritimus*" strain SCM1 were maintained in artificial seawater medium at 30ºC as described (*32*). To obtain sufficient cell material for NMR analysis, SCM1 was cultured in slowly stirred batches of 5 l volume (a total of 40 l was processed). Cells were harvested by filtration on 0.2 μm polycarbonate membrane filters (Whatman Inc.) and immediately frozen at –80ºC.

Preparation of *N. maritimus* cell extracts

After thawing, the cells were resuspended from the membranes with a spatula and by pipetting 1 mL aliquots of water over the filters until clean (~10 mL). The membranes were further washed with 7 mL water and then were placed in a sonication bath to remove any remaining cells. The resuspended cell fractions were combined, then lysed by sonication (1 x 8 min, 10 W). The samples were then frozen and water was removed by lyophilization. The dried cell debris was resuspended in D_2O and incubated in a 60 ºC water bath for one hour. Any remaining cell debris was removed via centrifugation. The supernatant was concentrated via lyophilization to 200 μL and EDTA was added to a final concentration of 20 mM. The sample was transferred to a Shigemi tube and analyzed by NMR spectroscopy (sample 1). In a second sample (sample 2), the cells

debris from the centrifugation step above was resuspended and placed in a water bath at 60 °C overnight. Again, the debris was removed via centrifugation and the supernatant was transferred to a Shigemi tube for NMR analysis after addition of EDTA to 20 mM final concentration. The final cell debris pellet was resuspended in water without removal of undissolved solids (sample 3). Each sample was analyzed by 31P NMR as described below (Fig. S4).

NMR analysis of phosphonates produced by *Nitrosopumilus maritimus*

All NMR spectra of *N. maritimus* samples were obtained on a Varian Unity Inova 600 MHz NMR spectrometer. ^{31}P NMR spectra were referenced against an external phosphoric acid standard (0 ppm). Sample 2 containing the cell extract was analyzed by ^{1}H-decoupled ^{31}P NMR spectroscopy, demonstrating two main compounds with chemical shifts of 28.7 ppm and 16.6 ppm (Fig 2 & Fig. S4B). Authentic methylphosphonate was detected at 25 ppm when spiked into the sample, demonstrating that neither peak is methylphosphonate. On the other hand, the methyl ester of methylphosphonate had a chemical shift (28.2 ppm) that was very close to the peak at 28.7 ppm (see above). Both the ^{31}P and ^{1}H chemical shifts of the standard are supporting the assignment of the peak at 28.7 ppm in the ^{31}P NMR of the *N. maritimus* extract as an esterified methylphosphonate, which was confirmed by the experiments described below. Similar spectra were obtained for samples 1 and 3, except that the relative intensities of the peaks at 28.7 and 16.6 ppm were different (Fig. S4). In the discussion below, we only focus on the peak at 28.7 ppm in the *N. maritimus* sample, which will be shown to correspond to esterified methylphosphonate.

Next, the sample was analyzed using a HMBC experiment (Fig. 2). The peak at 28.7 ppm showed a correlation with a proton resonance at 1.40 ppm. This chemical shift in the ^{1}H dimension is in the range expected for a methyl group. Furthermore, the peak at 1.40 ppm in the ^{1}H dimension is a doublet, showing only the expected coupling with the ^{31}P nucleus and no additional couplings. This observation shows that no other protons are coupled and is again consistent with the peak at 28.7 ppm in the ^{31}P NMR belonging to esterfied methylphosphonate. A weak crosspeak at 3.92 ppm in the ^{1}H dimension is also observed. This crosspeak agrees well with a proton of an ester group attached to the methylphosphonate, similar to the crosspeak observed in Fig S6A and 6B. The peak in Figure 2 does not show clear splitting and is weak, which suggests that it is further coupled to other protons, but we cannot make any definitive assignment. The weak intensity of this crosspeak is likely the consequence of one or more of the following factors. First, the *N. maritimus* sample is very dilute because of the very low cell densities that can be obtained for this organism. The spectrum in Figure 3 is from 40 L of cell culture and the sample was run for 5 days because of the long acquisition time (1 s) required to extract accurate splitting patterns with the standards shown above. Second, as shown in Figure S6, the intensity of the crosspeak with protons on the ester group is weaker than those to the α-protons. Third, the crosspeak in Figure S6 arises from 3 protons whereas the crosspeak in Figure 2 very likely arises from a CH_2 or a CH group. Although we cannot make any conclusions as to what compound methylphosphonate is attached to in the product of *N. maritimis*, we can conclude with high confidence that the phosphonate portion of the peak at 28.7 is indeed methylphosphonate. Only methylphosphonate can account for the chemical shift (1.4 ppm), the doublet splitting,

and the absence of any crosspeaks to β-protons. This assignment is further corroborated by the mass spectrometry data presented below.

Mass spectrometric detection of methylphosphonate

To provide further support for the assignment based on the NMR data, the cell debris of sample 3 was hydrolyzed in an attempt to release methylphosphonate from the ester to which it is attached to (possibly a macromolecule or membrane component). The insoluble matter of sample 3 was resuspended in 4 mL of 6 M HCl and refluxed for 3 h. The sample was concentrated via rotovap and lyophilization. The resulting material was taken up in 0.6 mL of H_2O, and the insoluble material was removed by centrifugation (13,000 rpm, 10 min). The pH of the crude hydrosylate (approximately 1 mL) was adjusted to 10 with ammonium hydroxide and then treated sequentially with 15 μL each of 1 M calcium acetate and ammonium bicarbonate to precipitate excess phosphate and calcium, respectively. Insoluble material was removed by centrifugation and the supernatant was dried under vacuum and subsequently reconstituted in 1 mL of 0.1% acetic acid. The reconstituted hydrosylate was applied to a homemade spin column loaded with 0.1 mL of Fe^{3+}-charged Hypercel IMAC resin (Pall Life Sciences). Unbound material was removed by washing with 10 column volumes each of 0.1% acetic acid and 0.1% acetic acid in 20% acetonitrile. The bound material was eluted with 200 μL of 0.5 M ammonium bicarbonate and the eluent was then dried under reduced pressure. The sample was reconstituted in 1 mL of HPLC grade water and treated with 100 mg Chelex 100 mesh beads (Biorad) to remove excess iron. The sample was dried under reduced pressure and dissolved in 200 μL of 10 mM ammonium bicarbonate in 90% acetonitrile. LC-MS analysis was performed on a custom 11 tesla hybrid linear ion-trap Fourier-transform mass spectrometer (LTQ-FT, Thermo-Fisher Scientific) equipped with a 1200 HPLC system (Agilent). An aliquot of 100 μL was injected onto a 2.1 mm × 150 mm Zic pHILIC column (Sequant) equilibrated with 100% B (10 mM ammonium bicarbonate in 90% acetonitrile), 0% A (10 mM ammonium bicarbonate). After 5 min of washing the column at initial conditions to remove unbound material, a linear gradient from 100% B to 40% B was developed over 10 min followed by a 5 min linear gradient back to initial conditions and a 15 min re-equilibration period prior to subsequent injections. Mass spectral data acquisition was comprised of three scan events detected in negative ion mode: a full scan detected in the FTMS, a collision-induced dissociation (CID) MS^2 scan of m/z 94.99 detected in the linear ion trap and a pseudo MS^2 scan utilizing source-fragmentation detected in the linear ion-trap. Data were analyzed manually using the Qualbrowser application of Xcalibur (Thermo-Fisher Scientific). Methylphosphonate was observed by extracted ion chromatography with a tolerated mass difference of 2 ppm. The methylphosphonate thus observed had an observed mass of 94.99024 showing excellent accuracy (-1.2 ppm; theor mass 94.99035) (Figs. 3C & S8D). The assignment was further corroborated by analysis of the ions in the ion trap using collision-induced dissociation (CID) mass spectrometry (Fig. S8E) and by source fragmentation (Fig. S8F) showing identical fragmentation patterns as observed by analysis of authentic methyl phosphonate using the same methods (Figs S8A-C). Hence, these mass spectrometric data fully agree with the conclusions from the NMR data and collectively, the production of methylphosphonate by *N. pumilis* is unambiguously shown.

Identification and multiple sequence alignment of MpnS, HepD and HppE homologs from metagenomic data

To identify the homologs of Mpns, HepD and HppE we used a set of biochemically validated proteins to search the non-redundant and environmental databases (*33*) at NCBI using BLASTP. These reference sequences used were *Nitrosopumilus maritimus* SCM1 MpnS (YP_001581491.1), *Streptomyces viridochromogenes* HepD (CAJ14042.1) and HppE from *Streptomyces fradiae* (ACG70827.1) and *Pseudomonas syringae* (BAA94418.1). This resulted in the identification of 46 homologous proteins for MpnS family members at an E-value < 10^{-10}, all of which were derived from the Global Ocean Survey (GOS) dataset (*21*). Queries with HepD and HppE families resulted in 20 and 0 homologs from the environmental database at the same E-value threshold. Two HepD homologs and one MpnS homolog were recovered from the non-redundant database. Thus, a total of 73 sequences were used for multiple sequence alignment and subsequent phylogenetic analysis. To align these sequences we used COBALT, a constraint-based multiple alignment tool (*33*) available from NCBI using the default parameters.

Inference and construction of the maximum likelihood based phylogenetic tree

Phylogenetic analysis of the identified protein sequences from the three families was performed using the RAXML package (*34*). We first generated 100 Maximum Likelihood (ML) trees using the COBALT aligned sequences with the amino acid substitution model set to BLOSUM62 and GAMMA rate variation along with other parameters set to default. The consensus tree representing this set of ML trees was then generated using the bipartition algorithm (option -f b) included in package. The fraction of ML trees that had each branch in the best ML tree is reported to support the topology of this best tree (bootstrap support). The final consensus tree was represented using figtree (http://tree.bio.ed.ac.uk/software/figtree/), a graphical viewer of phylogenetic trees, with the HppE clade as the outgroup. Using this approach, The MpnS, HppE and HepD homologs identified by BlastP form distinct clades with robust bootstrap support (Figs. 3A & S9).

Clustering of MpnS, HepD and HppE homologs from metagenomic data based on pairwise global sequence alignment

To provide independent approach for functional assignment of the MpnS, HepD and HppE families we performed a pairwise global alignment of all 73 homologs using the Needle algorithm available as part of the Emboss suite of tools (*35*). Needle creates an optimal global alignment of two sequences using the Needleman-Wunsch algorithm and provides a raw score reflecting the extent of similarity between the two sequences being compared. We employed this raw score for every pair of sequences to construct a matrix of sequence similarity scores, which allowed us to cluster the sequences based on this metric. Clustering of sequences was performed using the open source software Cluster (*36*). This analysis involves hierarhical clustering of the sequences using uncentered correlation as the distance/similarity metric and complete linkage as the clustering method. The resulting clustered heatmap was visualized using Java treeview package (*37*). The intensity of the red color in the heatmap corresponds to the extent of similarity between pairs of sequences. As with the BlastP and phylogentic tree methods described above, the hierarchical clustering reproducibly forms discrete groups for the produced HepD, MpnS and HppE (Fig S10).

Calculation of the abundance of MpnS and Ppm genome equivalents in the marine metagenome

To estimate the number of genomes which contain the *mpnS* and *ppm* genes, we compared the relative abundance of reads homologous to these genes to the number of reads homologous to six single-copy essential genes (*recA*, *tufA*, *atpD*, *dnaK*, *gyrB* and *rpoB*) as previously described (*23*). This was achieved by performing a local analysis against the complete set of reads reported in the GOS dataset using TBLASTN. Homologous genes were recovered at two e-value thresholds (10^{-10} and 10^{-20}). Because some of the reads in the GOS dataset represent paired-end reads of the same clone (essentially counting the same clone twice) the number of hits were reduced by the number of paired end reads. Because larger genes are expected to be cloned at a higher frequency, the counts were then normalized to the size of the *recA* gene. The final corrected values of the six single-copy genes were then averaged to provide a robust estimate of the number of genome equivalents in the dataset and compared to the corrected counts of mpnS to provide an estimate for the fraction of genomes sampled in the GOS dataset that encode this gene (Table S4).

Phylogenetic classification of the 23sRNA gene on GOS scaffold EN800522.

GOS scaffold EN800522 encodes a 1267 bp fragment from the 3' end of a 23s rRNA gene. To determine the phylogenetic placement of this gene, we performed a BLASTN search of the reference genomic sequences at NCBI using the extracted 23s rRNA region as a query. A phylogenetic tree was then generated using the Fast Minimum Evolution method (*38*) and visualized using figtree as described above (Fig. S11).

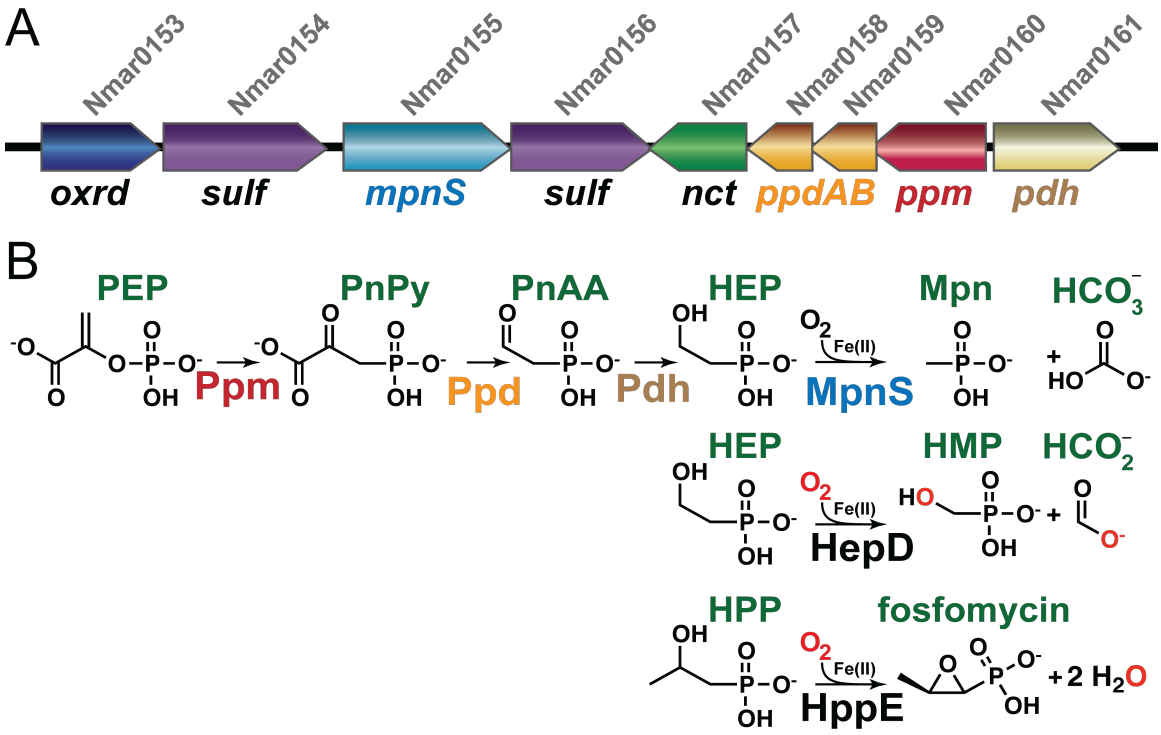


Fig. 1. **A biosynthetic pathway for MPn synthesis in *N. maritimus*.** (A) A cluster of genes found on the *N. maritimus* chromosome has significant homology to known phosphonate biosynthetic genes including phosphonopyruvate mutase (*ppm*), phosphonopyruvate decarboxylase (*ppdAB*) and phosphonoacetaldehyde dehydrogenase (*pdh*). Additional nearby genes have homology to oxidoreductases (*oxrd*) and sulfatases (*sulf*). Data presented in the text demonstrate that *mpnS* encodes a methylphosphonate synthase. The *N. maritimus* locus tags are shown above each gene. (B) The reactions catalyzed by enzymes of the MPn biosynthetic pathway are shown with the enzymes catalyzing each reaction shown below the arrow. Two related phosphonate biosynthetic reactions catalyzed hydroxyethylphosphonate dioxygenase (HepD) and hydroxypropylphosphonate epoxidase (HppE) are also shown. The fate of O_2 consumed in these reactions is shown in red. Abbreviations used: PEP, phosphoenolpyruvate; PnPy, phosphonopyruvate; PnAA, phosphonoacetaldehyde; HEP, hydroxyethylphosphonate; MPn, methylphosphonate; HMP, hydroxymethylphosphonate; HPP, hydroxypropylphosphonate, HCO_3^-, bicarbonate; HCO_2^-, formate.

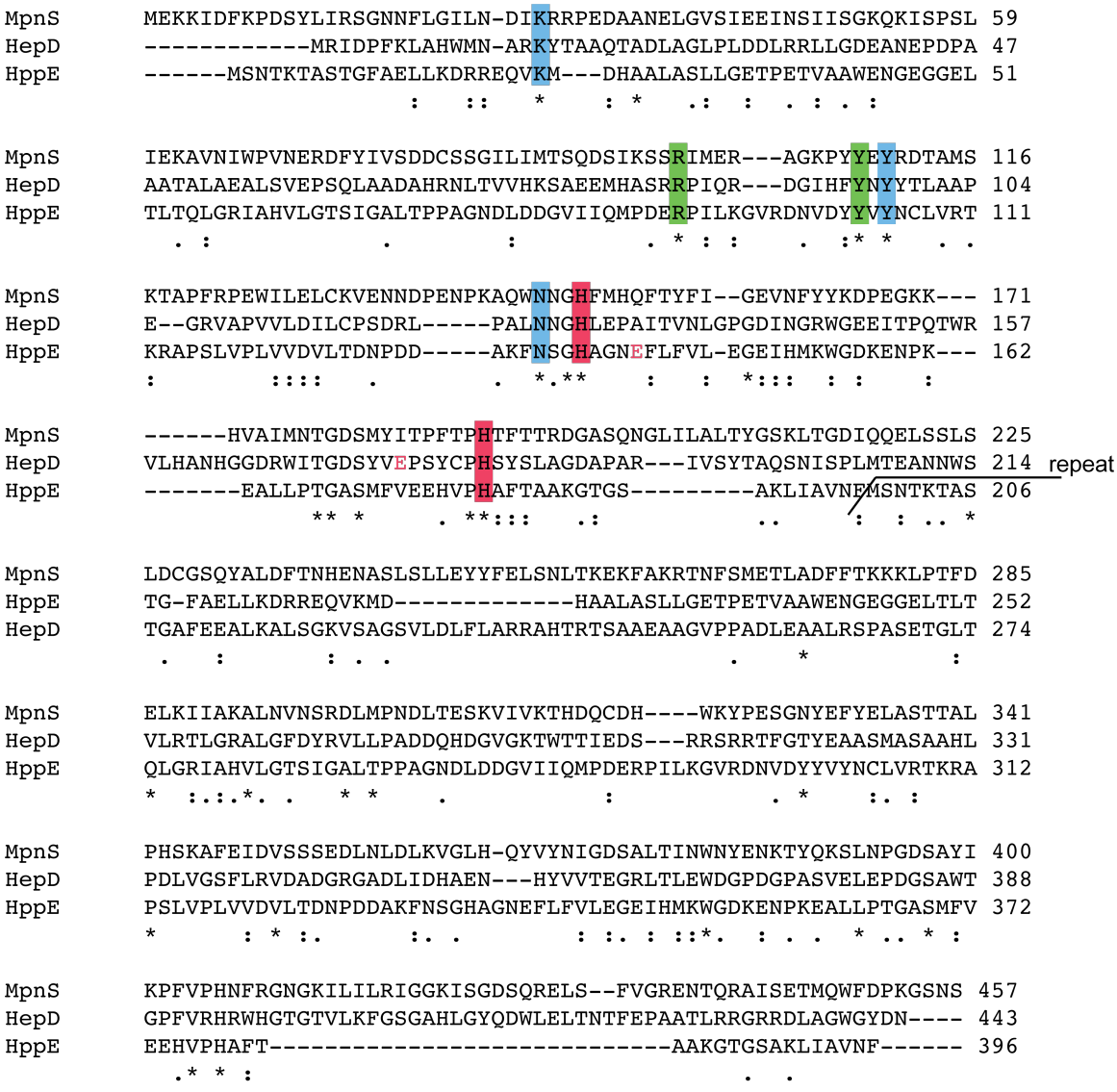


Fig. S2.

Conserved features in cupin family dioxygenases involved in phosphonate biosynthesis. MpnS, HepD, and HppE were aligned using Clustal W, then modified by hand to improve the alignment. A back-to-back duplication of the HppE sequence was used to maintain the correct alignment because HppE has only a single cupin domain, whereas HepD and MpnS have two. Residues highlighted in blue have been shown to be directly hydrogen-bonded to the phosphonate in HepD, residues highlighted in green are hydrogen-bonded to the phosphonate through water (*27*). These amino acids are conserved among all three enzymes. Residues highlighted in red represent the conserved Fe-binding ligands of the 2-His-1-carboxylate facial triad.

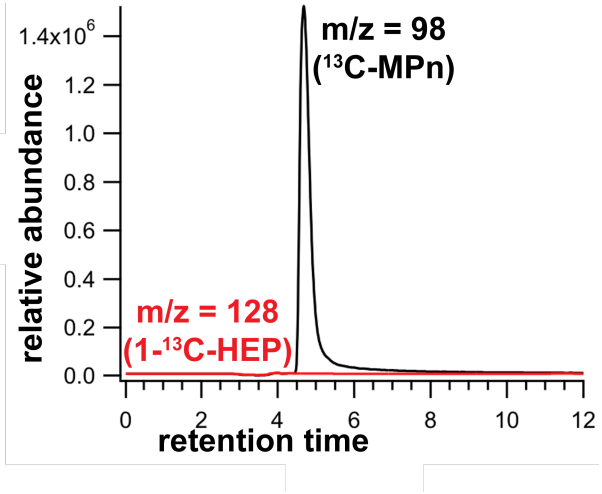


Fig. S3

Liquid chromatography-mass spectrometric (LC-MS) assay of the MpnS reaction. The reaction shown in Fig1A was analyzed by LC-MS using single ion monitoring (SIM) of the substrate 1-^{13}C-HEP at *m/z* 128 (red line) and product ^{13}C-MPn at *m/z* 98 (black line).

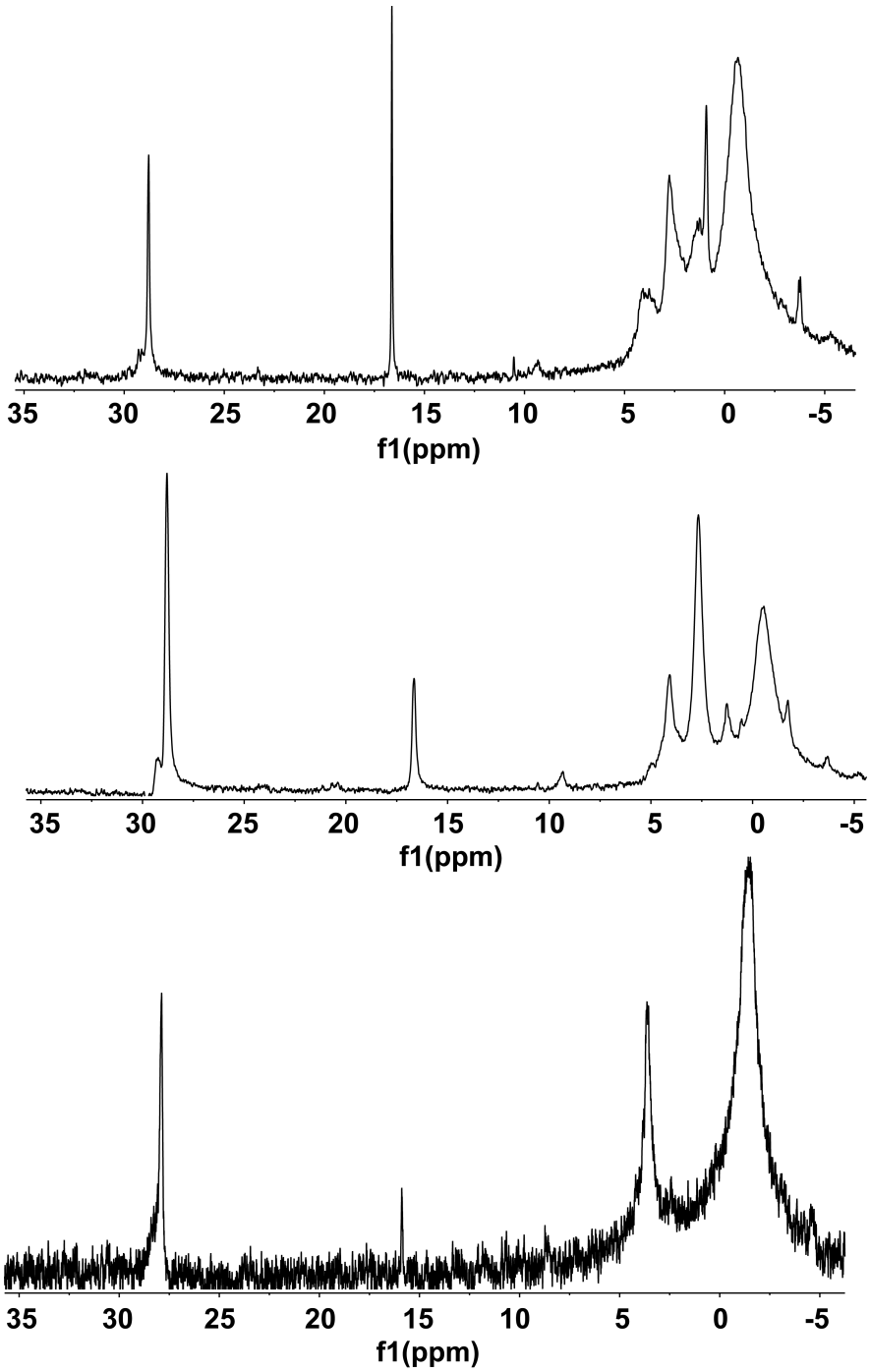
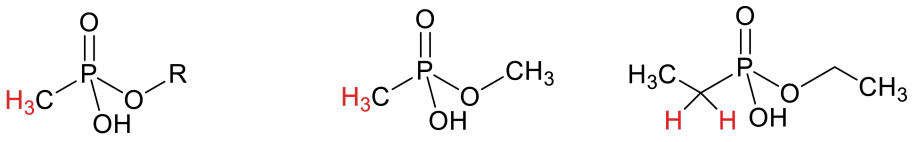


Fig. S4

Variation in the ^{31}P NMR spectra of *N. maritimus* cell extracts. Samples were prepared as described in the accompanying text. (top) soluble cell extract (sample 1), (middle) 60 ^{o}C aqueous extract of cell debris (sample 2), (bottom) resuspended cell debris (sample 3).



Structure of peak at 28.7 ppm R is unknown	**1**	**2**
^{31}P: 28.7 ppm, m	^{31}P: 28.2 ppm, m	^{31}P: 30.5 ppm, m
^{1}H: 1.40 ppm, d	^{1}H: 1.28 ppm, d	^{1}H: 1.43 ppm, m

Fig. S5

Structures of the compounds analyzed in this study and their associated chemical shifts in ^{1}H and ^{31}P NMR. ^{1}H chemical shift values are shown for protons colored in red.

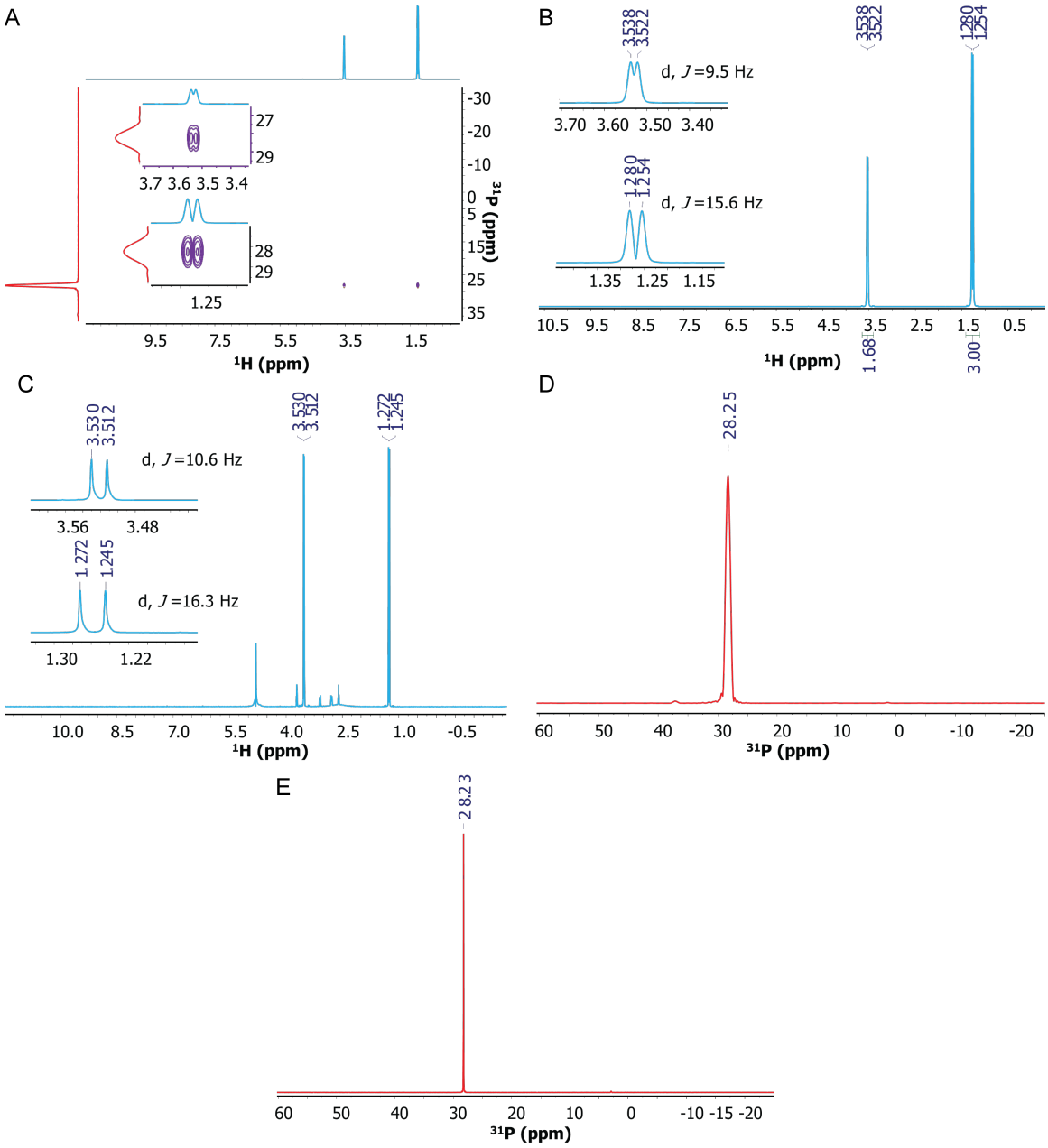


Fig. S6

NMR analysis of synthetic methyl methylphosphonate. (A) HMBC spectrum. (B) ^{1}H dimension extracted from panel A. (C) ^{1}H NMR spectrum of methyl methylphosphonate. (D) ^{31}P dimension extracted from panel A. (E) ^{31}P NMR spectrum of methyl methylphosphonate. The doublet of the proton coupled to the phosphorus peak at 28 ppm is diagnostic for a methyl group bonded directly to phosphorus, *i.e.* a methylphosphonate moiety. See accompanying text for detailed discussion.

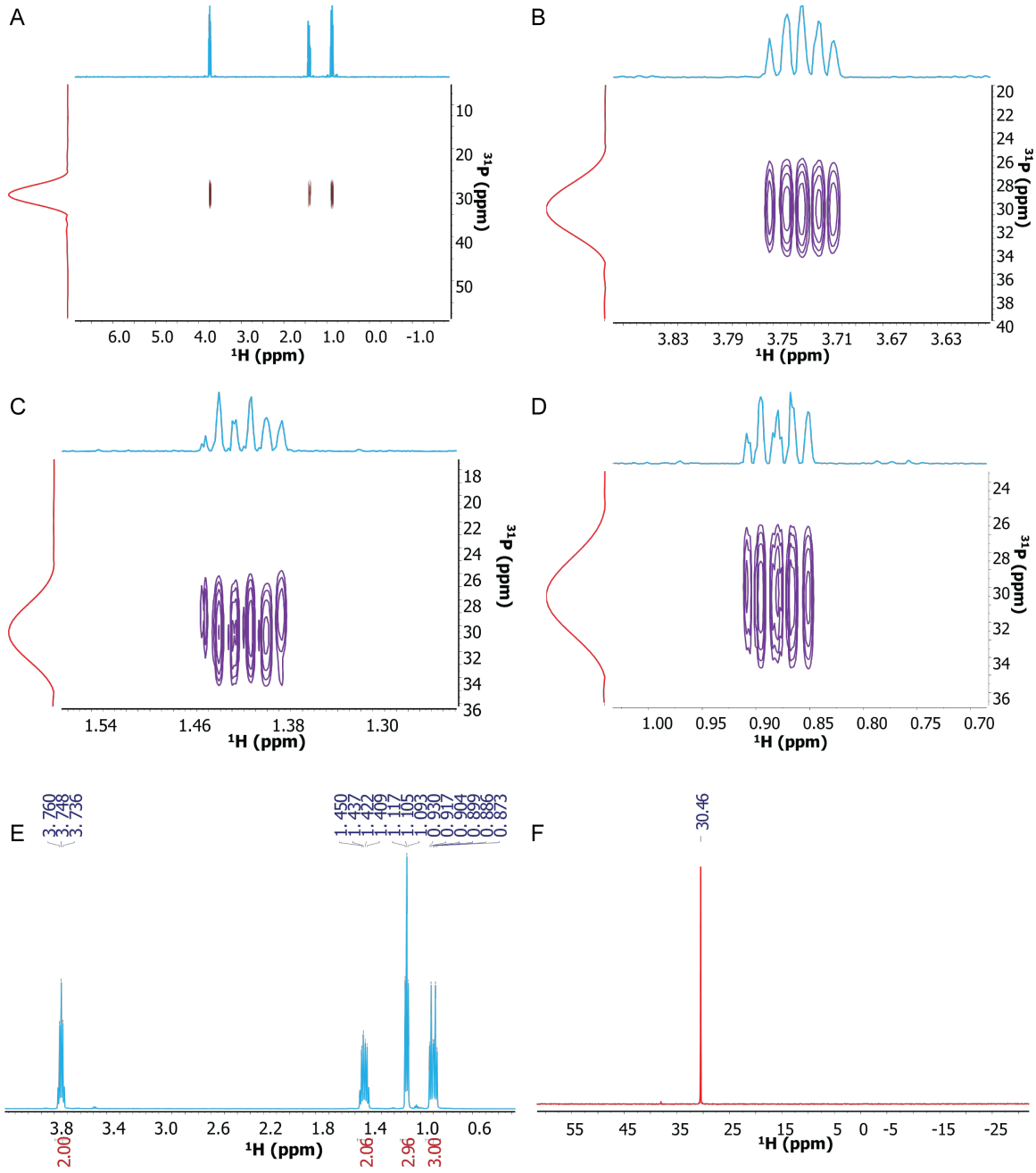


Fig. S7

NMR analysis of ethyl ethylphosphonate . (A) HMBC spectrum. (B) HMBC showing the region of the spectrum of the methylene (CH_2) of the ethyl ester group. (C) HMBC showing the region of the spectrum of the methylene (CH_2) of the ethyl group attached to phosphorus. (D) HMBC showing the region of the spectrum of the two methyl groups. (E) ^{1}H NMR spectrum of ethyl ethylphosphonate. (F) ^{31}P NMR spectrum of ethyl ethylphosphonate. Note that the complex splitting patterns of the protons coupled to phosphorus rule out the presence of a methyl group bonded directly to phosphorus, *i.e.* this compound does not contain a methylphosphonate moiety.

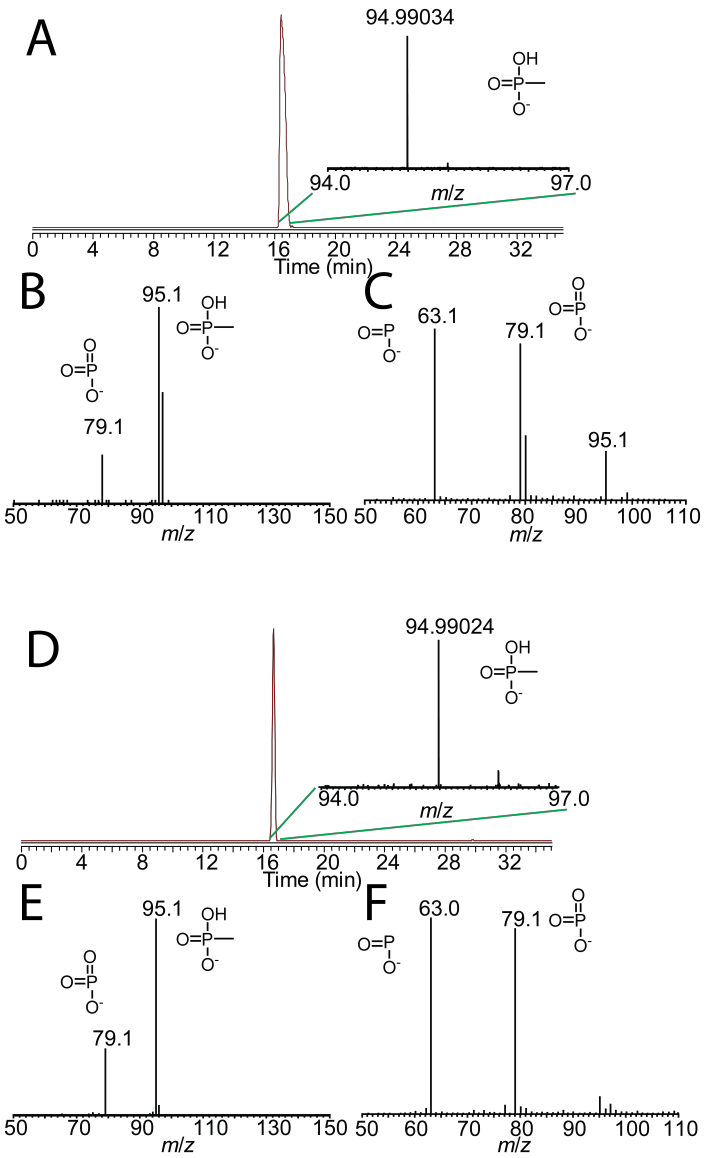


Fig. S8

High resolution LC-MS analysis of authentic methylphosphonate (panels A-C) and free methylphosphonate released from *N. maritimus* cell debris by strong acid hydrolysis (panels D-F). (A and D) Extracted ion chromatogram at 2 ppm centered around *m/z* 94.99035 (exact monoisotopic mass of methylphosphonate $[M-H]^-$) with Fourier-transform mass spectrum and ion structure (*insets*). (B and E) Ion-trap tandem mass spectrum of the methylphosphonate ion shown in panels A and C with precursor and fragment ion structures. (C and F) Pseudo-tandem mass spectrum derived from the source fragmentation scan with fragment ion structures. Note that data from the *N. maritimus* sample is identical to that from the standard.

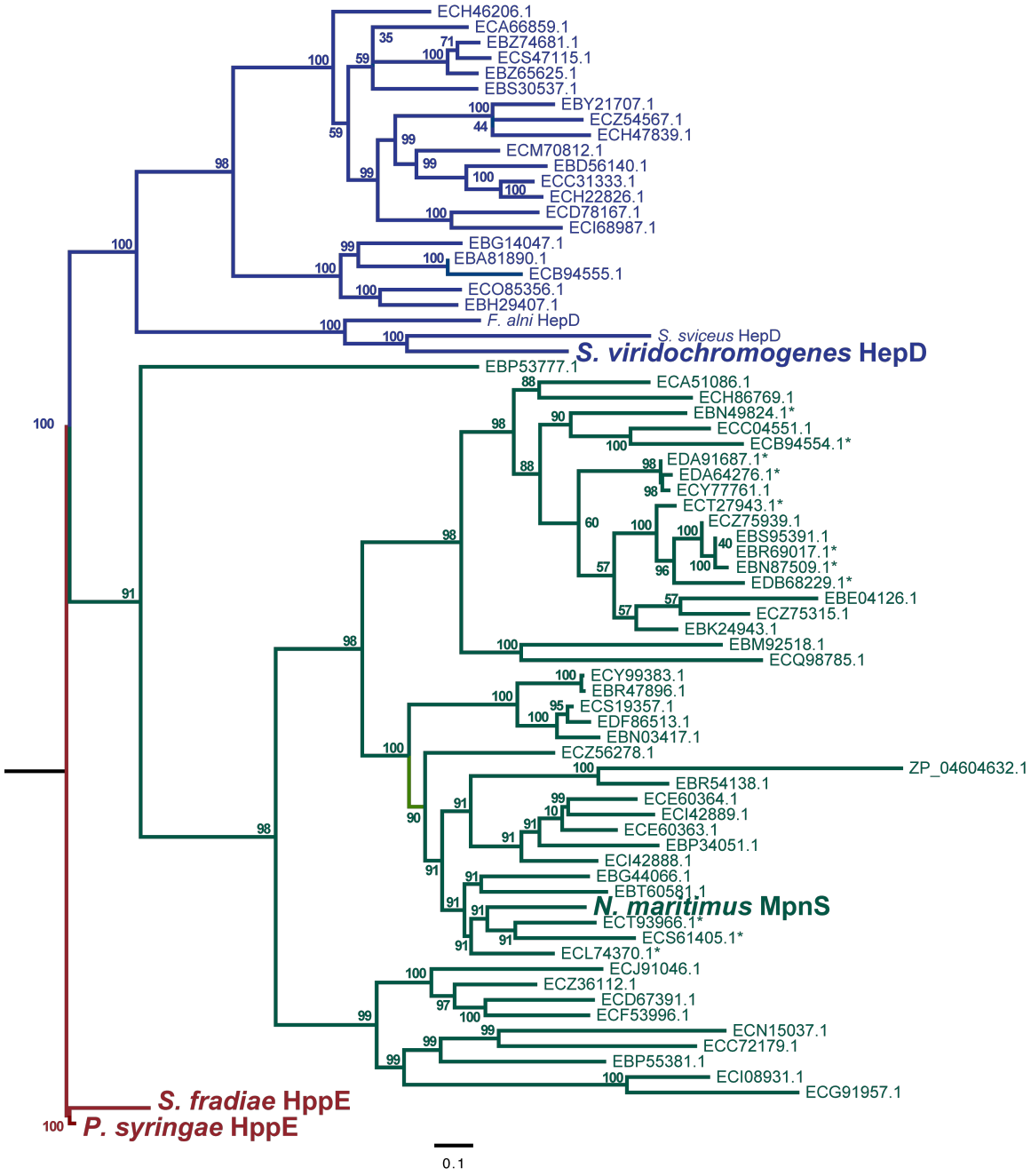


Fig. S9

Evolutionary relationships of biochemically characterized MpnS, HepD and HppE proteins (shown in bold) and homologs recovered from Genbank and the GOS metagenomic dataset. The tree was inferred using maximum likelihood analysis as described. Bootstrap values from 100 replicates are shown at the nodes. A simplified tree with several of the nodes collapsed into triangles is presented in Fig 3.

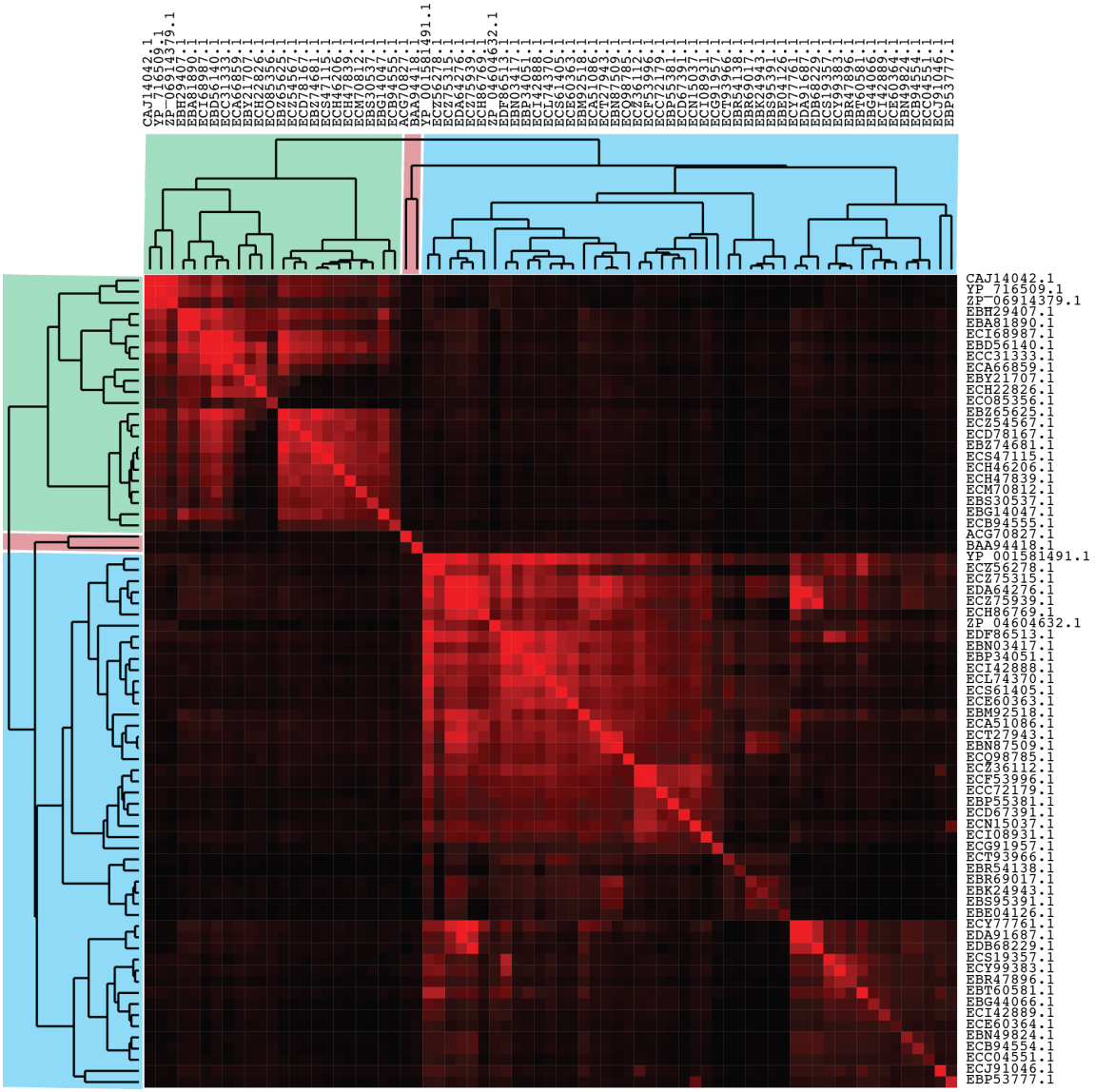


Fig. S10

Hierarchical clustering of biochemically characterized MpnS, HepD and HppE proteins. A pairwise global alignment of all 73 homologs using the Needle algorithm (*35*) was used to construct a matrix of sequence similarity scores, which was used for hierarchical clustering as described. The resulting clustered heatmap was visualized using Java treeview package (*37*). The intensity of the red color in the heatmap corresponds to the extent of similarity between pairs of sequences. As with the BlastP and phylogentic tree methods described above, the hierarchical clustering reproducibly forms discrete groups for the produced HepD, MpnS and HppE

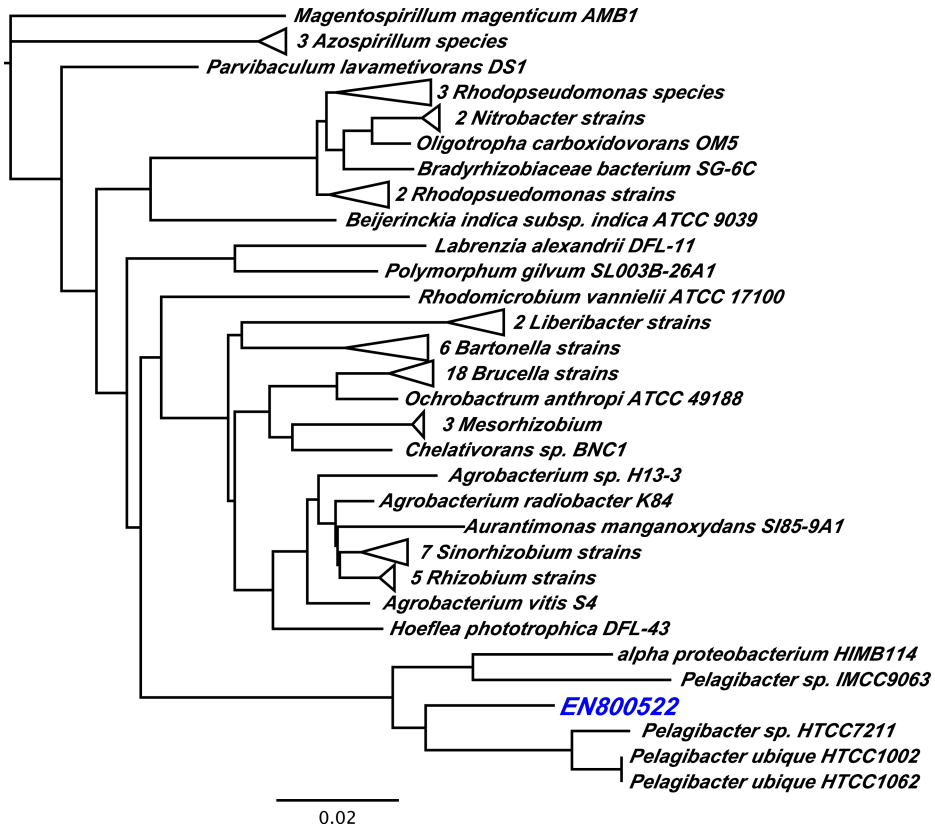


Fig S11

Hierarchical clustering of biochemically characterized MpnS, HepD and HppE proteins. A pairwise global alignment of all 73 homologs using the Needle algorithm (*35*) was used to construct a matrix of sequence similarity scores, which was used for hierarchical clustering as described. The resulting clustered heatmap was visualized using Java treeview package (*37*). The intensity of the red color in the heatmap corresponds to the extent of similarity between pairs of sequences. As with the BlastP and phylogentic tree methods described above, the hierarchical clustering reproducibly forms discrete groups for the produced HepD, MpnS and HppE

Table S1.

Genes found nearby the *N. maritimus* MpnS locus. Gene annotations and putative functions were extracted from the extensive bioinformatics analysis of the *N. maritimus* genome sequence available through the Integrated Microbial Genomes system (*39*). Genes of the MPn biosynthetic locus shown in Fig1 are shaded in yellow.

Locus Tag	GenBank Gene Description	putative function
Nmar_0145	hypothetical protein	polysaccharide biosynthesis
Nmar_0146	polysaccharide biosynthesis protein	polysaccharide biosynthesis
Nmar_0147	hypothetical protein	polysaccharide biosynthesis
Nmar_0148	class III aminotransferase	heme biosynthesis
Nmar_0149	acylneuraminate cytidylyltransferase	polysaccharide biosynthesis
Nmar_0150	epimerase/dehydratase	polysaccharide biosynthesis
Nmar_0151	pseudogene (N-acylneuraminate-9-phosphate synthase)	polysaccharide biosynthesis
Nmar_0152	pseudogene (N-acylneuraminate-9-phosphate synthase)	polysaccharide biosynthesis
Nmar_0153	oxidoreductase domain protein	dehydrogenase
Nmar_0154	sulfatase	
Nmar_0155	cupin family protein	methylphosphonate synthase
Nmar_0156	sulfatase	
Nmar_0157	sugar nucleotidyltransferase protein	polysaccharide biosynthesis
Nmar_0158	phosphonopyruvate decarboxylase	phosphonate biosynthesis
Nmar_0159	phosphonopyruvate decarboxylase	phosphonate biosynthesis
Nmar_0160	phosphoenolpyruvate phosphomutase	phosphonate biosynthesis
Nmar_0161	iron-containing alcohol dehydrogenase	HEP biosynthesis
Nmar_0162	Rieske (2Fe-2S) domain protein	redox mediator
Nmar_0163	hypothetical protein	Rieske 2Fe-2S family protein
Nmar_0164	signal transduction protein	transcriptional regulator
Nmar_0165	glycosyl transferase family protein	polysaccharide biosynthesis
Nmar_0166	sulfotransferase	
Nmar_0167	transcriptional regulator/ polysaccharide biosynthesis	polysaccharide biosynthesis
Nmar_0168	NAD-dependent epimerase/dehydratase	polysaccharide biosynthesis

Table S2.

BLASTP was used to search the environmental samples database at NCBI on January 12, 2012 using *N. maritimus* MpnS (YP_001581491), *S. viridochromogenes* HepD (AAU00079), *S. fradiae* HppE (ACG70827) and *P. syringae* HppE (BAA94418.1) as query sequences. Proteins with BLAST expect values below 1 x 10^{-10} are shown below. No significant hits were obtained using either HppE query sequence. Proteins shaded yellow map to the scaffolds depicted in Figure 4.

MpnS Homologs		**HepD Homologs**	
GOS Protein ID	Blast E value	GOS Protein ID	Blast E value
EDF86513.1	1.00E-156	EBH29407.1	1.00E-89
ECZ56278.1	6.00E-138	EBD56140.1	2.00E-81
EBP34051.1	8.00E-132	EBA81890.1	3.00E-70
ECL74370.1	1.00E-128	EBZ65625.1	5.00E-63
EDA64276.1	3.00E-119	ECI68987.1	9.00E-62
ECS61405.1	5.00E-109	ECZ54567.1	4.00E-52
EBN03417.1	5.00E-104	EBG14047.1	1.00E-50
ECI42888.1	2.00E-100	EBZ74681.1	6.00E-49
ECZ75315.1	2.00E-99	ECD78167.1	9.00E-49
ECE60363.1	3.00E-98	ECC31333.1	5.00E-47
EBM92518.1	6.00E-95	ECA66859.1	6.00E-44
EBT60581.1	1.00E-88	ECH46206.1	3.00E-43
ECZ75939.1	6.00E-87	ECS47115.1	2.00E-39
ECA51086.1	4.00E-72	ECH47839.1	3.00E-34
ECZ36112.1	4.00E-69	ECM70812.1	7.00E-29
ECT27943.1	1.00E-66	EBY21707.1	3.00E-27
ECH86769.1	4.00E-65	ECH22826.1	5.00E-27
ECQ98785.1	5.00E-62	EBS30537.1	2.00E-26
EBN87509.1	5.00E-62	ECB94555.1	4.00E-20
ECS19357.1	3.00E-58	ECO85356.1	6.00E-19
ECY77761.1	8.00E-58		
ECF53996.1	3.00E-55		
ECY99383.1	4.00E-55		
ECC72179.1	2.00E-51		
ECN15037.1	3.00E-51		
EBG44066.1	1.00E-43		
ECT93966.1	3.00E-42		
EBR47896.1	1.00E-39		
EBP55381.1	6.00E-39		
ECD67391.1	2.00E-37		
ECI42889.1	3.00E-36		
EDA91687.1	8.00E-32		
ECI08931.1	2.00E-30		
EDB68229.1	6.00E-29		
EBR69017.1	2.00E-27		
EBK24943.1	8.00E-24		

ECJ91046.1	1.00E-23
EBN49824.1	1.00E-23
ECE60364.1	1.00E-22
ECB94554.1	4.00E-21
EBR54138.1	9.00E-21
ECG91957.1	4.00E-18
EBE04126.1	3.00E-15
ECC04551.1	3.00E-15
EBS95391.1	6.00E-14
EBT60581.1	2.00E-11
EBP53777.1	2.00E-11

Table S3. The proteins encoded by the GOS scaffolds containing MpnS homologs were used as query sequences in BLASTP searches against the NCBI non-redundant database on 4-1-2012. The top hit for each search is shown. In cases where the best match was to an uncultivated organism, the best match to a cultivated organism is also shown.

GOS Scaffold ID	GOS Protein ID of query	Putative function	Accession # of best hit	organism	query coverage (%)	E-value	Max Identity (%)
EM010993	not annotated	sulfatase	YP_001581490.1	*Nitrosopumilus maritimus* SCM1	92	5.00E-112	55
EM010993	ECS61405.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	94	2.00E-98	70
EM348795	ECL74369.1	sulfatase	ABZ07909.1	uncultured marine microorganism	96	1.00E-59	47
EM348795	ECL74369.1	sulfatase	ZP_08258183.1	*Nitrosoarchaeum limnia* SFB1	96	3.00E-26	34
EM348795	ECL74369.1	sulfatase	YP_001581490.1	*Nitrosopumilus maritimus* SCM1	96	8.00E-24	35
EM348795	ECL74370.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	99	6.00E-117	66
EM690505	ECT93967.1	sulfatase (C-terminus)	YP_001581490.1	*Nitrosopumilus maritimus* SCM1	100	2.00E-32	37
EM690505	ECT93968.1	sulfatase (N-terminus)	YP_001581490.1	*Nitrosopumilus maritimus* SCM1	91	6.00E-27	54
EM690505	ECT93966.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	92	4.00E-41	67
EP889348	EDA64274.1	Pdh	YP_003514386.1	*Stackebrandtia nassauensis*	96	7.00E-52	35
EP889348	EDA64278.1	Ppd	YP_003993893.1	*Halanaerobium hydrogeniformans*	89	1.00E-15	40
EP889348	EDA64275.1	Nct	ZP_03462318.1	*Bacteroides pectinophilus*	97	2.00E-12	49
EP889348	EDA64276.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	98	2.00E-109	43
EP889348	EDA64277.1	ManC	ZP_05068884.1	*Pelagibacter sp.* HTCC7211	99	1.00E-133	77
EP880318	EDA91686.1	ManC	ZP_05068884.1	*Pelagibacter sp.* HTCC7211	99	3.00E-152	77
EP880318	EDA91687.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	99	1.00E-33	35
EP816386	EDB68229.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	97	4.00E-28	34
EP816386	EDB68230.1	ManC	ZP_05068884.1	*Pelagibacter sp.* HTCC7211	96	2.00E-165	72
EN550784	EBR69017.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	89	7.00E-27	55
EN550784	EBR69018.1	Nct	ZP_03275556.1	*Arthrospira maxima* CS-328	79	2.00E-14	49
EN550784	not annotated	Ppm	YP_004168410.1	*Nitratifractor salsuginis*	95	5.00E-05	48
EN783203	EBN87510.1	Ppm	BAI87872.1	*Arthrospira platensis* NIES-39	97	3.00E-94	54
EN783203	EBN87511.1	Nct	BAI87872.1	*Arthrospira platensis* NIES-39	81	3.00E-22	40
EN783203	EBN87509.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	87	1.00E-57	58
EN959135	EBK24941.1	Ppm	ZP_03729787.1	*Dethiobacter alkaliphilus* AHT 1	95	3.00E-42	51
EN959135	EBK24942.1	Nct	BAI87872.1	*Arthrospira platensis* NIES-39	87	5.00E-34	43
EN959135	EBK24943.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM2	94	2.00E-23	53
EM830888	ECB94553.1	Nct	ZP_01288329.1	*delta proteobacterium* MLMS-1	82	1.00E-20	44
EM830888	ECB94554.2	MpnS (C-terminus)	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	97	1.00E-20	41
EM830888	ECB94555.3	MpnS (N-terminus)	YP_716509.1	*Frankia alni* ACN14a	97	7.00E-22	36
EM830888	ECB94551.4	sulfatase	ZP_09383840.1	*Flavonifractor plautii* ATCC 29863	53	4.00E-04	43
EM830888	ECB94552.5	sulfatase	NP_908273.1	*Wolinella succinogenes* DSM 1740	98	2.00E-09	38
EN800522	EBN49824.1	MpnS	YP_001581491.1	*Nitrosopumilus maritimus* SCM1	99	4.00E-23	44
EN800522	not annotated	23s rRNA	AY458637.2	Uncultured marine bacterium	100	0.00	95
EN800522	not annotated	23s rRNA	CP000084.1	*Pelagibacter ubique* HTCC1062	100	0.00	94

Table S4

The relative abundance of the *mpnS* and *ppm* genes in marine microorganisms was estimated as described in (*23*). The number of reads in the Global Ocean Survey (GOS) dataset that encode proteins homologous to six single copy genes was first determined using TBLASTN with expect value cutoff scores (E) of 10^{-10} or 10^{-20}. These numbers were then corrected by subtracting the number of paired end reads (which correspond to mutliple reads of the same clone) and by normalizing the counts by the size of the query gene (relative to the *recA* gene). The final values represent an estimate of the number of genome equivalents in the dataset, which was then compared to the number of *mpnS* and *ppm* reads quantified in the same manner.

	# of GOS reads		**# of paired end reads removed**		**normalized to *recA***	
	E < 10^{-10}	**E < 10^{-20}**	**E < 10^{-10}**	**E < 10^{-20}**	**E < 10^{-10}**	**E < 10^{-20}**
recA	8845	7730	1257	1036	7588	6694
tufA	11642	10313	1910	1684	8722	7733
atpD	15233	11539	2317	1775	9918	7498
dnaK	17524	15733	3014	2673	8038	7235
gyrB	21236	17252	4289	3354	7452	6112
rpoB	26241	24334	6293	5959	5258	4843
			Average genome equivalents		7829	6686
	# of GOS reads		**# of paired end reads removed**		**normalized to *recA***	
	E < 10^{-10}	**E < 10^{-20}**	**E < 10^{-10}**	**E < 10^{-20}**	**E < 10^{-10}**	**E < 10^{-20}**
mpnS	71	55	8	6	55	43
			Fraction of genomes with *mpnS* gene		0.0070	0.0064
	# of GOS reads		**# of paired end reads removed**		**normalized to *recA***	
	E < 10^{-10}	**E < 10^{-20}**	**E < 10^{-10}**	**E < 10^{-20}**	**E < 10^{-10}**	**E < 10^{-20}**
ppm	2672	970	181	51	2940	1085
			Fraction of genomes with *ppm* gene		0.3755	0.1622