Diversity and abundance of phosphonate biosynthetic genes in nature

Xiaomin Yu[a,b], James R. Doroghazi[b], Sarath C. Janga[b,1], Jun Kai Zhang[a], Benjamin Circello[a,b], Benjamin M. Griffin[b], David P. Labeda[c], and William W. Metcalf[a,b,2]

[a]Department of Microbiology and [b]Institute for Genomic Biology, University of Illinois at Urbana–Champaign, Urbana, IL 61801; and [c]US Department of Agriculture Agricultural Research Service, National Center for Agricultural Utilization Research, Peoria, IL 61604

Edited by Edward F. DeLong, Massachusetts Institute of Technology, Cambridge, MA, and approved November 6, 2013 (received for review August 14, 2013)

Phosphonates, molecules containing direct carbon–phosphorus bonds, compose a structurally diverse class of natural products with interesting and useful biological properties. Although their synthesis in protozoa was discovered more than 50 y ago, the extent and diversity of phosphonate production in nature remains poorly characterized. The rearrangement of phosphoenolpyruvate (PEP) to phosphonopyruvate, catalyzed by the enzyme PEP mutase (PepM), is shared by the vast majority of known phosphonate biosynthetic pathways. Thus, the *pepM* gene can be used as a molecular marker to examine the occurrence and abundance of phosphonate-producing organisms. Based on the presence of this gene, phosphonate biosynthesis is common in microbes, with ∼5% of sequenced bacterial genomes and 7% of genome equivalents in metagenomic datasets carrying *pepM* homologs. Similarly, we detected the *pepM* gene in ∼5% of random actinomycete isolates. The *pepM*-containing gene neighborhoods from 25 of these isolates were cloned, sequenced, and compared with those found in sequenced genomes. PEP mutase sequence conservation is strongly correlated with conservation of other nearby genes, suggesting that the diversity of phosphonate biosynthetic pathways can be predicted by examining PEP mutase diversity. We used this approach to estimate the range of phosphonate biosynthetic pathways in nature, revealing dozens of discrete groups in *pepM* amplicons from local soils, whereas hundreds were observed in metagenomic datasets. Collectively, our analyses show that phosphonate biosynthesis is both diverse and relatively common in nature, suggesting that the role of phosphonate molecules in the biosphere may be more important than is often recognized.

Phosphorus is an essential nutrient for all living organisms, required for the synthesis of nucleic acids, phospholipids, phosphorylated exopolysaccharides, and numerous metabolites. In most organisms, the preferred source of phosphorus is inorganic phosphate. However, because the majority of phosphate salts are highly insoluble, this ion is rarely available in concentrations that support unbridled growth. Thus, despite the fact that phosphorus is the 11th most abundant element in the Earth's crust, it is a limiting nutrient in many ecosystems (1, 2). As a result, nature has evolved highly efficient phosphate transport systems, as well as systems for acquisition of phosphorus from essentially all known biomolecules containing this element (3, 4). These biomolecules are composed mostly of assorted phosphate esters and anhydrides secreted by other organisms during growth or released by decomposition after their death. An impressive body of research has characterized the processes by which these compounds are made and subsequently used. However, recent reports suggest that other, less-well studied phosphorus compounds may also be important in the biosphere (5). Among these are the phosphonic and phosphinic acids, compounds characterized by the presence of highly stable carbon–phosphorus (C–P) bonds in place of the labile carbon-oxygen-phosphorus linkages found in more familiar phosphorus-containing biomolecules.

Synthetic phosphonates have been known for more than a century, but their role in biology was not suspected until 1959 when 2-aminoethylphosphonic acid (AEP) was found in the acid hydrolysate of rumen protozoa (6). Since then, C–P compounds have been identified in a variety of bacteria, archaea, and eukaryotes (7, 8). In many of these organisms, phosphonates are found as constituents of extracytoplasmic macromolecules, including phosphonolipids and phosphonoglycans (5). It is generally believed that the inert nature of the C–P bonds in these structural molecules imparts resistance to hydrolytic enzymes, thus providing an advantage in phosphorus-limited environments (9). Other organisms produce a diversity of low-molecular-weight, bioactive phosphonates, whose inhibitory properties stem from the structural similarity of stable phosphonates to labile phosphate esters and carboxylates (5). The range of biological activities exhibited by these molecules is impressive, and examples with antibacterial, antiviral, antiparasitic, and herbicidal properties are known (5, 10). Thus, it seems likely that phosphonate biosynthesis has evolved in these organisms to provide a competitive advantage through chemical warfare.

Although the number of proven phosphonate producers is relatively small, it is increasingly clear that these molecules compose a significant fraction of the bioavailable phosphorus in many ecosystems (11). ^{31}P NMR analyses conducted in the 1990s suggested that phosphonates can account for nearly 25% of the high-molecular-weight dissolved organic P in marine environments (12, 13), whereas newer methods show that 5–10% of the

Significance

Phosphonic acids are organophosphorus molecules containing direct carbon–phosphorus bonds that are often perceived as biological rarities. The data presented here show that the ability to synthesize diverse phosphonates is both widespread and relatively common among microbes. These findings are consistent with recent evidence suggesting that phosphonates are important intermediates in the global phosphorus cycle. Moreover, they support the idea that these molecules play a significant role in the biology of producing organisms, including a mechanism to sequester scarce phosphorus resources and to compete via chemical warfare using toxic phosphonate mimics of common metabolic intermediates.

Author contributions: X.Y., J.R.D., B.M.G., and W.W.M. designed research; X.Y., J.R.D., S.C.J., J.K.Z., and B.C. performed research; D.P.L. contributed new reagents/analytic tools; X.Y., J.R.D., S.C.J., B.M.G., and W.W.M. analyzed data; and X.Y. and W.W.M. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

Data deposition: The sequences reported in this paper have been deposited in the GenBank database (accession nos. KF386859–KF387474).

[1]Present address: School of Informatics and Computing, Indiana University-Purdue University Indianapolis, Indianapolis, IN 46202.

[2]To whom correspondence should be addressed. E-mail: metcalf@illinois.edu.

This article contains supporting information online at www.pnas.org/lookup/suppl/doi:10.1073/pnas.1315107110/-/DCSupplemental.

MICROBIOLOGY

phosphorus pool across the size spectrum is composed of C–P compounds (14). Similar studies show that phosphonates occur in a variety of soils, as well as in freshwater lakes and streams (15–18). The abundance of phosphonate catabolism genes in genomic and metagenomic datasets shows that the trait is widespread and common among microbes, with 30–40% of strains encoding one or more of these pathways (19, 20). Other studies revealed that these catabolic genes are often highly expressed in situ (21–23). Taken together, these data indicate that phosphonates are likely to be an important source of phosphorus in many environments.

In contrast to data supporting a significant role for C–P compounds in the biological phosphorus cycle, evidence regarding the source of phosphonates in nature remains scarce. In agricultural settings, these compounds may arise from xenobiotic sources, as synthetic phosphonates are widely used as herbicides and pesticides (e.g., glyphosate and phosphinothricin); however, in more pristine settings, these compounds must have a biogenic source. The characterized producers described above are not always abundant where phosphonates are detected. Thus, the principal sources of phosphonates in the biosphere remain to be identified.

Early studies addressing the source of environmental phosphonates relied on chemical analysis of whole organisms or tissue samples, which are limited in scope simply due to the inability to sample a significant fraction of the organisms present in any given environment. In most cases, the presence of phosphonates was assessed by ^{31}P NMR using the distinctive chemical shift of C–P molecules, which is observed in the +5 to +45 ppm range, as opposed to the +5 to −20 ppm range observed for phosphate and its esters and anhydrides (24). However, ^{31}P NMR is relatively insensitive and cannot detect phosphorus compounds that are present in low abundance, as is often the case in phosphonate producers. Furthermore, a variety of phosphate esters have chemical shifts in the phosphonate range, including many known biological products, such as glycerol 1′,2′-cyclic phosphate and 2′,3′-cGMP (25–27). Hence, additional confirmatory chemical tests for the presence of phosphonic acids (e.g., 2D NMR, mass spectrometry, chemical and enzymatic reactivity) are highly desirable, but rarely provided.

The unique properties of phosphonic acids have evoked considerable interest in their biosynthesis, and the genes and enzymes involved in the synthesis of several bioactive phosphonates, as well as those needed for the synthesis of certain phosphonolipids and phosphonoglycans, have now been characterized (5, 28–30). With a single exception, all known phosphonates are derived from phosphoenolpyruvate (PEP) by isomerization to phosphonopyruvate in a reaction catalyzed by the enzyme PEP mutase (PepM) (5). Because this initial reaction is thermodynamically unfavorable, net synthesis of phosphonates requires a thermodynamically favorable step immediately following PepM. Initial studies identified phosphonopyruvate decarboxylase (Ppd) as the enzyme providing this driving force; however, recent data show that other driving reactions exist (31, 32). Subsequent biosynthetic steps diverge to create a wide range of C–P biomolecules. The genes encoding these downstream enzymes are usually clustered together with the *pepM* gene, a trait that has greatly simplified genetic and biochemical studies of phosphonate biosynthesis and that allows prediction of the types of C–P molecules produced by the metabolic pathways encoded by these linked genes (29, 33, 34).

Detailed knowledge regarding C–P compound biosynthesis opened the door to gene-based methods for assessing the abundance and identity of biological phosphonate producers. Villarreal-Chiu et al. quantified the occurrence of Ppd homologs to estimate the prevalence of phosphonate metabolism in microbes, finding the gene in ∼10% of sequenced geneomes and genome equivalents in metagenomic datasets (20). PepM homologs—commonly located adjacent to genes involved in the synthesis of methylphosphonic acid, a putative precursor of methane in the aerobic ocean (8)—were also shown to be abundant (∼16% of genome equivalents) in the Global Ocean Survey (GOS) data (35, 36). Although both studies suggest that phosphonate production is relatively common, the methods used have the potential to overestimate the occurrence of phosphonate production due to the difficulties of assigning protein function on the basis of homology alone. PepM is a member of the large isocitrate lyase superfamily (37), whereas Ppd is a member of a large family of thiamine-pyrophosphate–using enzymes (38). Accordingly, homology-based searches (such as BLASTP) using Ppd and PepM as queries can identify family members that are not involved in phosphonate biosynthesis, even when fairly stringent cutoff values are used.

In this report we revisit the use of PepM to estimate the distribution and abundance of phosphonate production in nature. By using a stringent filtering strategy that requires all counted hits to include catalytically important amino acid residues that distinguish PepM from other isocitrate lyase superfamily members, we avoid ambiguities caused by the evolutionary ancestry of the enzyme. Moreover, by analyzing the *pepM* gene neighborhoods, we provide insight into the nature of the phosphonate molecules being produced. Our results reveal a surprising abundance and diversity of phosphonate biosynthesis in nature, suggesting that their role in the biosphere has yet to be fully appreciated.

Results

Ability to Synthesize Phosphonates Is Relatively Common and Widespread in Nature. We searched for *pepM* homologs in the Integrated Microbial Genomes (IMG) database (39), GOS marine metagenomes (35, 36), and the Integrated Microbial Genomes with Microbiome Samples (IMG/M) database (40), using the sequences of biochemically validated PepM sequences as queries (31, 33, 34). All potential hits were screened for the highly conserved catalytic motif EDKXXXXXNS, which distinguishes PEP mutase from other members of the isocitrate lyase superfamily (41).

Two hundred and forty-seven bacterial genomes (5.7%) in the IMG genomes database encode *pepM* homologs. PEP mutase-encoding bacteria include 9 of 33 bacterial phyla represented, predominantly from *Proteobacteria*, *Bacteroidetes*, *Firmicutes*, *Actinobacteria*, and *Spirochaetes*, but the ability to produce phosphonates is not phylogenetically coherent except among *Burkholderia*, *Selenomonas*, and *Nitrosococcus* (*SI Appendix*, Fig. S1 and Dataset S1). Seventy-seven of 82 sequenced *Burkholderia* genomes (accounting for 31% of the *pepM*-positive genomes) encode at least one *pepM* gene, with some strains carrying as many as four, whereas eight of nine sequenced *Selenomonas* genomes and all four sequenced *Nitrosococcus* genomes contain *pepM*. Two archaeal genomes (1.1%), *Nitrosoarchaeum limnia* SFB1 and *Nitrosopumilus maritimus* SCM1, encode *pepM*. Seven *pepM* homologs were discovered in six eukaryotic genomes (3.2%), including sea snail (*Lottia gigantean*), sea anemone (*Nematostella vectensis*), and several protozoa (*Paramecium tetraurelia*, *Perkinsus marinus*, *Tetrahymena thermophile*, and *Trypanosoma cruzi*).

Analysis of the GOS metagenomes, which encompass near-surface marine environments from around the world, revealed *pepM* homologs from 59 of 79 sampling sites (*SI Appendix*, Table S1 and Dataset S2). The IMG/M microbiomes (40), which include metagenomic data from a variety of ecosystems, contain *pepM* homologs in 558 of the 1,281 datasets in the collection (*SI Appendix* and Dataset S3). To quantify the relative occurrence of phosphonate biosynthetic genes in these datasets, we compared the abundance of *pepM* to that of typical single-copy genes as

described (42). By this measure, *pepM* genes occur in ~7.6% of microbial genomes in the GOS and IMG/M metagenomes (assuming only one copy of *pepM* per genome) (*SI Appendix*; Dataset S2 and Dataset S3). The abundance of *pepM* is essentially the same in all GOS sites (between 4.4 and 8.7%), with a mangrove site and a hypersaline site (both in Ecuador) having the highest (20.7%) and the lowest (0.3%) abundance, respectively (Fig. 1*A*). In contrast, there are significant differences among the various IMG/M metagenomes based on the large set of metagenomic data analyzed, with mammalian- and molluscan-associated microbiomes having the highest median *pepM* abundances (17.4 and 29.9%, respectively) (Fig. 1*B*).

To gain insight into which taxa are associated with phosphonate biosynthesis in the metagenomic data, we used PhymmBL, a bioinformatics tool that allows strong inference regarding the phylogeny of organisms from which individual sequence reads are derived (43). This analysis predicts that *Proteobacteria*, followed by *Firmicutes*, *Cyanobacteria*, and *Spirochaetes*, are the most abundant *pepM*-associated phyla in the marine environments represented in the GOS data, with 10 other phyla represented at lower abundances (Fig. 1*C*; *SI Appendix*; Dataset S2). At the genus level, 40% of *pepM* reads seem to be derived from *Prochlorococcus* (17.9%), *Candidatus Pelagibacter* (11.8%), and *Borrelia* (10.8%) whereas remaining reads were predicted to belong to the other 62 genera (*SI Appendix*; Dataset S2). In the IMG/M metagenomes, at least 28 different prokaryotic groups were predicted at the phylum level, primarily from *Firmicutes*, *Proteobacteria*, and *Bacteroidetes* (Fig. 1*D*; *SI Appendix*; Dataset S3), with the top five *pepM*-containing genera being *Bacteroides* (7.0%), *Clostridium* (6.8%), *Bacillus* (6.6%), *Nitrosopumilus* (6.5%), and *Streptococcus* (4.3%) (*SI Appendix* and Dataset S3). PhymBL analysis of a *pepM* PCR clone library from local soils (see below) revealed a similar distribution with 97% of the reads being assigned to *Proteobacteria* and only minor fractions assigned to *Actinobacteria* (1.5%), *Nitrospirae* (1.2%), and *Acidobacteria* (0.2%). The two most abundant subgroups in this library were the proteobacterial orders *Burkholderiales* (73.6%) and *Rhizobiales* (19.2%).

Differences in abundances observed between these various datasets are almost certainly due to the substantial differences between the sampled environments for metagenomic data and the extreme sampling bias toward pathogenic microbes for the genome sequence data.

Finally, we assessed the prevalence of phosphonate biosynthesis in cultivable bacteria by screening using degenerate PCR primers designed from conserved PepM amino acids motifs. Due to our interest in bioactive phosphonic acids, we focused our efforts on actinomycetes, which are known to be prolific producers of diverse natural products. We examined 1,649 strains isolated from local soils using protocols that are selective for spore-forming actinobacteria and an additional 973 strains from the US Department of Agriculture Agricultural Research Service actinobacteria collection: of these, 120 strains (4.6%) gave PCR products that were verified to be authentic *pepM* amplicons by DNA sequencing.

***pepM* Gene Neighborhoods Suggest Considerable Diversity in Phosphonate Biosynthetic Pathways.** Because the genes encoding phosphonate biosynthetic pathways are typically clustered together with the *pepM* gene (29, 33, 34), examination of the *pepM* gene neighborhood provides insight into the diversity of their phosphonic acid products. Accordingly, we examined the *pepM* gene neighborhoods in all known phosphonate producers, as well as those encoded in sequenced microbial genomes and in 25 of the *pepM*-positive actinobacteria, which were cloned and sequenced as described (*SI Appendix, Materials and Methods*). Putative gene

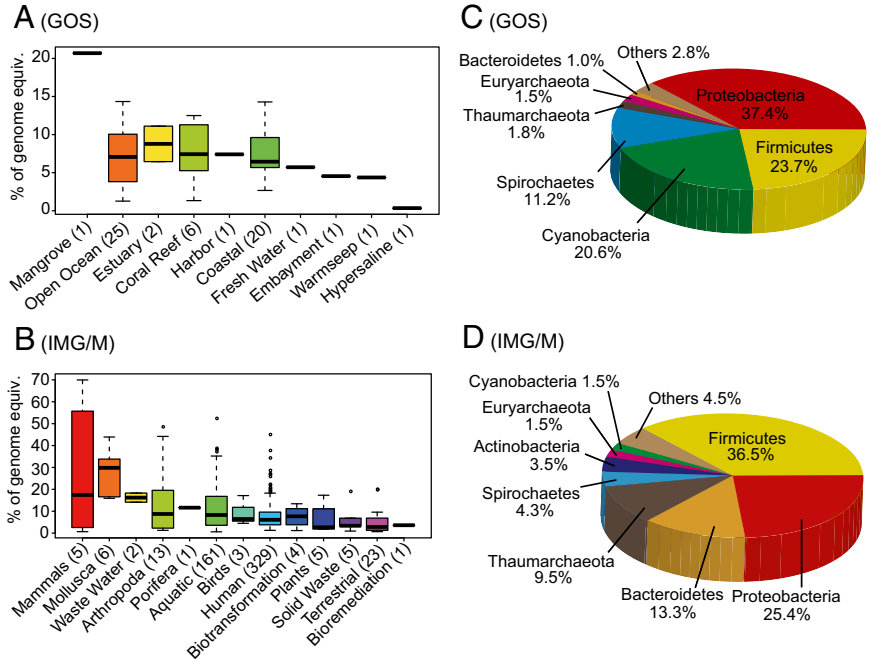

Fig. 1. *pepM* gene abundance in GOS metagenomes and IMG/M microbiomes. (*A*) Boxplot of prokaryotic genome equivalents for *pepM* occurrence by habitat type in GOS. Single black lines represent the median value for environments that were sampled only once. No significant difference was found in the relative *pepM* abundance across GOS habitats ($P = 0.9328$, Kruskal Wallis test applied to habitats with more than one sampling site). (*B*) Boxplot of percentage of prokaryotic genome equivalents for *pepM* occurrence by ecosystem type in IMG/M microbiomes. Relative *pepM* abundance across various ecosystems differed significantly ($P < 3.3 \times 10^{-5}$, Kruskal Wallis test applied to categories with more than one sampling site). In *A* and *B*, the number of sampling sites for each type is shown in parentheses. (*C*) Distribution of predicted prokaryotic phyla for *pepM* homologs identified in GOS metagenomes. (*D*) Distribution of predicted prokaryotic phyla for *pepM* homologs identified in IMG/M microbiomes. Phyla that account for <1% are grouped in "Others."

functions were assigned on the basis of genome annotations and homology of the protein products to enzymes of known function. Annotated diagrams of all gene clusters are shown in *SI Appendix*, Figs. S2–S26 and at http://file-server.igb.illinois.edu/~xyu9/Dataset_S4._Phosphonate_gene_clusters.html.

Several large groups of nearly identical gene clusters were observed that appear to be involved in synthesis of phosphonolipids or phosphonoglycans (Fig. 2*A* and *SI Appendix*, Fig. S27). The largest of these, composed of 93 gene clusters found mostly in *Burkholderia*, consists of a seven-gene putative operon that we propose may be responsible for synthesis of phosphonolipids with 1-hydroxy-2-aminoethylphosphonate as the head group (Fig. 2*A*, group 1; *SI Appendix*, Fig. S27), similar to the lipids previously characterized in *Bdellovibrio stolpii* (44). A second large group is found solely in *Burkholderia* species, most of which also encode the group 1 gene cluster (Fig. 2*A*, group 2). We predict that these genes are responsible for synthesis of 2-hydroxy-phosphonoacetate, which may be used as an alternative phosphonolipid headgroup in these strains (*SI Appendix*, Fig. S27). A third group including 27 gene clusters, mainly from *Bacteroides* and *Treponema*, includes a locus previously implicated in the synthesis of capsular polysaccharide B by *Bacteroides fragilis*, which contains AEP in an ester linkage to a hydroxyl group of the carbohydrate (45) (Fig. 2*A*, group 3). Interestingly, the nearly identical cluster found in *Bacteroides eggerthii* DSM 20697 has replaced the gene needed to make AEP from phosphonoacetaldehyde with a gene that produces hydroxyethylphosphonate (HEP) from the same substrate (46). Thus, we conclude that some *Bacteroides* strains produce a similar capsular polysaccharide decorated with HEP instead of AEP. Group 4 and several additional groups with a few members each also contain genes related to lipid biosynthesis and, thus, are probably involved in the synthesis of phosphonolipids (Fig. 2*A* and *SI Appendix*, Fig. S27). Collectively, these groups account for approximately half of the identified *pepM* gene clusters. The remaining *pepM* gene neighborhoods are highly diverse and include the genes for synthesis of the known bioactive phosphonates, as well as those from all of the actinomycetes sequenced in this study (Fig. 2*A*, group 9 and group 10; *SI Appendix*, Fig. S27). We expect that the majority of these gene clusters will direct synthesis of low-molecular-weight C–P compounds. Interestingly, 10% of sequenced genomes analyzed in our dataset lack the gene encoding Ppd; thus Ppd-independent phosphonate biosynthesis is common in nature.

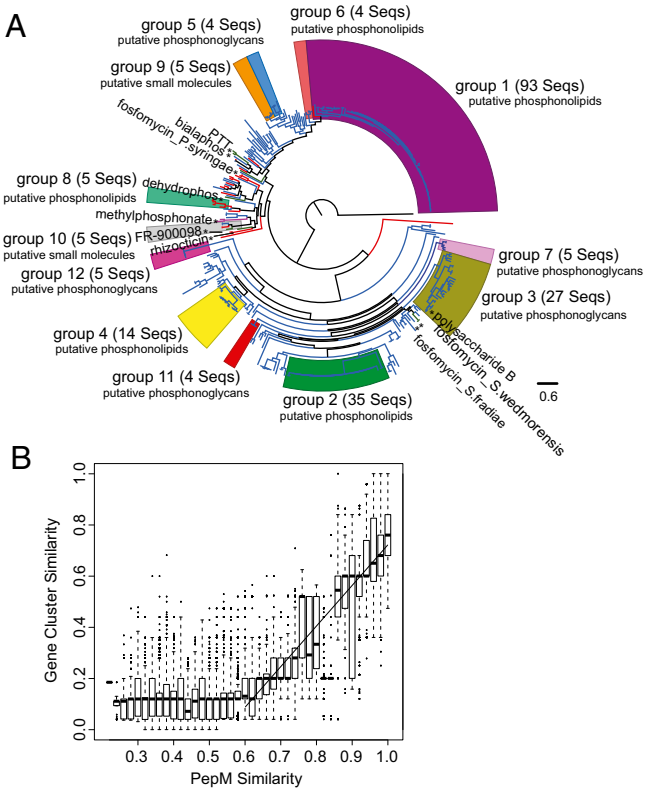

Fig. 2. Analysis of PepM and phosphonate gene clusters from microbial genomes. (*A*) Maximum-likelihood tree of PEP mutase sequences from NCBI and 25 actinomycete strains from this study. The tree was calculated with the FastTree program (53) with default settings. The tree is rooted with 2-methylisocitrate lyase sequence (NP_286072). The branch is colored based on the source of *pepM* sequences: 25 actinomycete isolates from this study (red), NCBI archaeal genomes (purple), NCBI bacterial genomes (blue), and NCBI genomic fragments (green). Selected *pepM* groups are highlighted by color shading, with the number of sequences within a group shown in parentheses. Known phosphonate compounds are indicated and marked with an asterisk. The phosphonate biosynthetic loci for 25 actinomycete strains from this study are shown in *SI Appendix*, Fig. S2–S26. The phosphonate gene cluster for each organism is listed at http://file-server.igb.illinois.edu/~xyu9/Dataset_S4._Phosphonate_gene_clusters.html. and is shown in the same order as in the tree. (*B*) Phosphonate gene cluster similarity as a function of PepM identity. Phosphonate gene cluster similarity was calculated using the fraction of homologous genes shared by two gene clusters. PepM identity was calculated using pairwise deletion of missing sites across the entire PepM alignment. Gene-cluster similarity measures were binned by PepM identity at intervals of 0.02 and plotted with a standard boxplot. The line shown is a linear regression over the PepM identity range 0.6–1.0, with equation $\hat{y} = -0.86 + 1.6x$; R-square = 0.74. The correlation has a *P* value of 2.2×10^{-16}. The full dataset is plotted in *SI Appendix*, Fig. S28.

PEP Mutase Sequence Similarity Is Strongly Correlated with Similarity of *PepM* Gene Neighborhoods. Due to the strongly conserved gene neighborhoods seen in many closely related strains, we were interested in whether PepM phylogeny was a reflection of the organismal phylogeny. To test this idea, we constructed a PepM phylogenetic tree using the sequences obtained from sequenced microbial genomes and from the 25 actinomycete gene clusters sequenced in this study (Fig. 2*A*). For this analysis, neighborhoods defined to encompass 6 genes upstream and downstream of *pepM* (13 genes in total). The results show scant correlation between the PEP mutase phylogeny and organismal phylogenies. This is strikingly illustrated by the group 1 and group 2 gene neighborhoods, which encode distantly related PEP mutase sequences, but which are often found in the same organism. Many other related PepM proteins are also found in unrelated organisms (e.g., the group 3 gene cluster PepM found in *Treponema* and *Bacteroides* and the group 1 gene cluster PepM found in *Burkholderia* and *Cupriavidus*).

In contrast, there is a strong correlation between PEP mutase phylogeny and the *pepM* gene neighborhood. Including those discussed above, we found 12 groups of nearly identical gene neighborhoods found in at least three different organisms (Fig. 2*A*). All of the PEP mutase sequences from each group are monophyletic, suggesting that similarity of PEP mutase is a good predictor of the other genes in the phosphonate biosynthetic pathway. To test this idea in a more rigorous fashion, we plotted the similarity of 342 *pepM* gene neighborhoods against the similarity of PEP mutase amino acid sequences in all pairwise combinations (58,311 comparisons). The data reveal a highly significant linear correlation for PEP mutase pairs having greater than 60% identity, with essentially no similarity in the *pepM* gene neighborhood at lower values (Fig. 2*B* and *SI Appendix*, Fig. S28).

Numerous Undiscovered Phosphonate Natural Products Are Likely to Exist. The preceding analysis shows that, when two organisms have similar PepM sequences, they are much more likely to

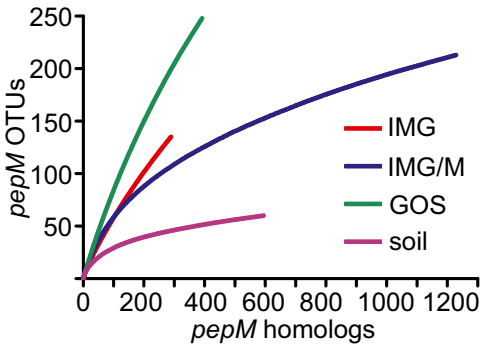

Fig. 3. Rarefaction analysis of amino acid sequences of *pepM* identified from IMG microbial genomes, IMG/M microbiomes, GOS metagenomes, and soil *pepM* clone libraries. Rarefaction curves are shown for OTUs with differences not exceeding 16%.

encode similar (if not identical) phosphonate metabolic pathways and therefore produce similar phosphonate molecules. The converse is also likely to be true. Therefore, we can estimate the diversity of phosphonate biosynthetic pathways in nature by examining the number of different PepM sequences in various datasets. To do this, we estimated the number of different PepM sequences in the microbial genomes, GOS metagenomes, and IMG/M metagenomes using rarefaction analysis (47, 48). We also constructed and sequenced a large library of *pepM* PCR amplicons using DNA isolated directly from local soils to provide a greater depth of coverage specific to the *pepM* gene.

Based on the resulting collector's curves, it is clear that the diversity of PepM sequences in nature is very large (Fig. 3). Among the datasets examined, only the soil PCR library was sampled at sufficient depth to approach saturation. Using an 84% PepM identity level, which represents conservation of roughly half of the genes in the cluster, we project that there are at least 135, 248, 213, and 60 PepM groups in the microbial genomes, GOS metagenomes, IMG/M metagenomes, and soil clone libraries, respectively (*SI Appendix*, Table S2). Significantly, there is little overlap between the PepM groups found from different environments (e.g., distinct PepM sequences found in aquatic ecosystems, terrestrial ecosystems, or animal-associated microbiota) (*SI Appendix*, Fig. S29). Thus, a realistic estimate of phosphonate diversity in nature would be closer to the sum of the four richness estimates.

Discussion

Biogenic C–P compounds were discovered more than 50 years ago and have been found in a variety of organisms, yet they are often portrayed as rare molecules with highly specialized functions (49). In contrast, our results show that the genetic capacity for phosphonic acid synthesis is common, widespread, and diverse, with more than 5% of sampled microbial cells encoding putative *pepM* genes. These microorganisms are found in every environment and include members of phylogenetically diverse groups. Moreover, these data probably underestimate the prevalence of phosphonate biosynthesis because larger eukaryotes such as mollusks, anemones, and corals, which are known to make phosphonates (7), are not well represented in either genomic or metagenomic datasets. Furthermore, numerous truncated homologs found in the metagenomic data were removed from our count because they lacked the diagnostic PepM catalytic motif. It is very likely that some of these would have contained the PepM motif if the full-length gene sequence had been available. This difference in counting method is also likely to explain the higher PepM abundances (~16%) that we previously reported (8) because PepM homologs were not filtered for the EDKXXXXXNS motif in that study.

Our results also show that phosphonate biosynthesis is highly diverse. Our calculations predict that hundreds of unique phosphonate molecules remain to be discovered in nature. Based on the content of available sequenced gene clusters, it is probable that many of these molecules will contain similar phosphonate moieties (e.g., AEP and HEP) attached to different glycans or lipids. However, many of the gene neighborhoods, including those for all known bioactive phosphonates, lack obvious candidates for gene glycan or lipid synthesis, suggesting that numerous small-molecule phosphonates await discovery. Given the potent bioactivity of this class of molecule, it is likely that pharmaceutically useful compounds will be among this group. In this regard, most of the actinomycete strains examined in this study have either unique *pepM* gene or ones that are similar to known phosphonate antibiotic gene clusters, suggesting that this group represents a rich source to mine for bioactive phosphonate natural products (Fig. 2*A*, red branches).

Materials and Methods

Bacterial Strains, Plasmids, and Culture Conditions. The bacterial strains and plasmids used in this study are listed in *SI Appendix*, Table S3. Growth conditions and molecular biology methods are described in *SI Appendix, Materials and Methods*.

Identification of PEP Mutase Gene Homologs in IMG-Sequenced Genomes, IMG/M Microbiomes, and GOS Metagenomes. PEP mutase gene homologs were identified from all sequenced genomes in the IMG database (39), all microbiomes in the IMG/M database (40), and GOS metagenomes in the CAMERA database (50) as described in *SI Appendix, Materials and Methods*. Genome equivalents and the percentage of genome equivalents with *pepM* in IMG/M microbiomes and GOS metagenomes were determined using the method suggested by Howard et al. (42).

Soil *pepM* Gene Clone Library Construction, Sequencing, and Sequence Analysis. Clone library construction, sequencing, and sequence analyses were performed as described in *SI Appendix, Materials and Methods*.

Strain Isolation, *pepM* Gene Screening, Sequencing, and Sequence Annotations of Phosphonate Biosynthetic Gene Clusters from Actinomycetes. Twenty-five *pepM*-positive actinomycete strains were chosen for cloning and sequencing of their phosphonate gene clusters: 21 isolates were from the genus *Streptomyces* and one each from the genera *Kitasatospora*, *Lechevalieria*, *Micromonospora*, and *Saccharothrix* (*SI Appendix*, Table S3). Strain isolation, *pepM* screening, gene cluster sequencing, and sequence annotations are described in *SI Appendix, Materials and Methods* and Figs. S2–S26.

PEP Mutase Sequence Diversity Analysis. PepM amino acid sequences from IMG-sequenced genomes, IMG/M microbiomes, GOS metagenomes, and soil clone libraries were aligned with ARB software (51), and distance matrices were used as input to Mothur software (52) for clustering sequences into operational taxonomic units (OTUs), generating rarefaction curves, and calculating richness estimates. The PEP mutase maximum-likelihood tree was constructed with FastTree (53) as described in *SI Appendix, Materials and Methods*.

Phosphonate Biosynthetic Gene Cluster Analysis. Phosphonate gene clusters from the National Center for Biotechnology Information (NCBI) sequenced genomes and from 25 actinomycete isolates from this study were compared and analyzed as described in *SI Appendix, Materials and Methods*.

Nucleotide Accession Numbers. Twenty-five actinomycete phosphonate biosynthetic gene cluster sequences have been deposited in GenBank under accession nos. KF386859–KF386883. Partial soil PEP mutase gene sequences have been deposited in GenBank under accession nos. KF386884–KF387474.

ACKNOWLEDGMENTS. We thank Laura Guest and Alvaro Hernandez of the University of Illinois at Urbana–Champaign Biotechnology Center for assistance in sequencing, and Amla Sampat and Joleen Su for technical support. We thank Jisen Zhang for providing scripts for soil data analysis. This work was supported by the National Institutes of Health (GM P01 GM077596).

1. Elser JJ, et al. (2007) Global analysis of nitrogen and phosphorus limitation of primary producers in freshwater, marine and terrestrial ecosystems. *Ecol Lett* 10(12): 1135–1142.
2. Dyhrman ST, Ammerman JW, Van Mooy BAS (2007) Microbes and the marine phosphorus cycle. *Oceanography* 20(2):110–116.
3. Beever RE, Burns DJW (1981) Phosphorus uptake, storage and utilization by fungi. *Advances in Botanical Research*, ed Woolhouse HW (Academic Press, London), Vol 8, pp 127–219.
4. van Veen HW (1997) Phosphate transport in prokaryotes: Molecules, mediators and mechanisms. *Antonie van Leeuwenhoek* 72(4):299–315.
5. Metcalf WW, van der Donk WA (2009) Biosynthesis of phosphonic and phosphinic acid natural products. *Annu Rev Biochem* 78:65–94.
6. Horiguchi M, Kandatsu M (1959) Isolation of 2-aminoethane phosphonic acid from rumen protozoa. *Nature* 184(Suppl 12):901–902.
7. Horiguchi M (1984) Occurrence, identification and properties of phosphonic and phosphinic acids. *Biochemistry of Natural C-P Compounds*, eds Hori T, Horiguchi M, Hayashi A (Japanese Association for Research on the Biochemistry of C-P Compounds, Shiga, Japan), pp 24–52.
8. Metcalf WW, et al. (2012) Synthesis of methylphosphonic acid by marine microbes: A source for methane in the aerobic ocean. *Science* 337(6098):1104–1107.
9. Horiguchi M (1984) Some physiological aspects of phosphonic and phosphinic acids. *Biochemistry of Natural C-P Compounds*, eds Hori T, Horiguchi M, Hayashi A (Japanese Association for Research on the Biochemistry of C-P Compounds, Shiga, Japan), pp 104–115.
10. Seto H, Kuzuyama T (1999) Bioactive natural products with carbon-phosphorus bonds and their biosynthesis. *Nat Prod Rep* 16(5):589–596.
11. McGrath JW, Chin JP, Quinn JP (2013) Organophosphonates revealed: New insights into the microbial metabolism of ancient molecules. *Nat Rev Microbiol* 11(6):412–419.
12. Clark LL, Ingall ED, Benner R (1998) Marine phosphorus is selectively remineralized. *Nature* 393:426.
13. Kolowith LC, Ingall ED, Benner R (2001) Composition and cycling of marine organic phosphorus. *Limnol Oceanogr* 46(2):309–320.
14. Young CL, Ingall ED (2010) Marine dissolved organic phosphorus composition: Insights from samples recovered using combined electrodialysis/reverse osmosis. *Aquat Geochem* 16(4):563–574.
15. Tate KR, Newman RH (1982) Phosphorus fractions of a climosequence of soils in New Zealand tussock grassland. *Soil Biol Biochem* 14(3):191–196.
16. Turner BL, Baxter R, Mahieu N, Sjogersten S, Whitton BA (2004) Phosphorus compounds in subarctic Fennoscandian soils at the mountain birch (*Betula pubescens*)-tundra ecotone. *Soil Biol Biochem* 36(5):815–823.
17. Cade-Menun BJ, Navaratnam JA, Walbridge MR (2006) Characterizing dissolved and particulate phosphorus in water with ^{31}P nuclear magnetic resonance spectroscopy. *Environ Sci Technol* 40(24):7874–7880.
18. Zhang RY, et al. (2009) Phosphorus composition in sediments from seven different trophic lakes, China: A phosphorus-31 NMR study. *J Environ Qual* 38(1):353–359.
19. Martinez A, Tyson GW, Delong EF (2010) Widespread known and novel phosphonate utilization pathways in marine bacteria revealed by functional screening and metagenomic analyses. *Environ Microbiol* 12(1):222–238.
20. Villarreal-Chiu JF, Quinn JP, McGrath JW (2012) The genes and enzymes of phosphonate metabolism by bacteria, and their distribution in the marine environment. *Front Microbiol* 3:19.
21. Dyhrman ST, et al. (2006) Phosphonate utilization by the globally important marine diazotroph *Trichodesmium*. *Nature* 439(7072):68–71.
22. Ilikchyan IN, McKay RML, Zehr JP, Dyhrman ST, Bullerjahn GS (2009) Detection and expression of the phosphonate transporter gene *phnD* in marine and freshwater picocyanobacteria. *Environ Microbiol* 11(5):1314–1324.
23. Ilikchyan IN, McKay RM, Kutovaya OA, Condon R, Bullerjahn GS (2010) Seasonal expression of the picocyanobacterial phosphonate transporter gene *phnD* in the Sargasso Sea. *Front Microbiol* 1:135.
24. Peck SC, Gao J, van der Donk WA (2012) Discovery and biosynthesis of phosphonate and phosphinate natural products. *Methods Enzymol* 516:101–123.
25. Salhany JM, Yamane T, Shulman RG, Ogawa S (1975) High resolution ^{31}P nuclear magnetic resonance studies of intact yeast cells. *Proc Natl Acad Sci USA* 72(12): 4966–4970.
26. Lebedev AV, Rezvukhin AI (1984) Tendencies of ^{31}P chemical shifts changes in NMR spectra of nucleotide derivatives. *Nucleic Acids Res* 12(14):5547–5566.
27. Boyd RK, et al. (1987) Glycerol 1,2-cyclic phosphate in centric diatoms. Observation by ^{31}P NMR *in vivo*, isolation, and structural determination. *J Biol Chem* 262(26): 12406–12408.
28. Lee JH, et al. (2010) Characterization and structure of DhpI, a phosphonate *O*-methyltransferase involved in dehydrophos biosynthesis. *Proc Natl Acad Sci USA* 107(41):17557–17562.
29. Circello BT, Eliot AC, Lee JH, van der Donk WA, Metcalf WW (2010) Molecular cloning and heterologous expression of the dehydrophos biosynthetic gene cluster. *Chem Biol* 17(4):402–411.
30. Borisova SA, Circello BT, Zhang JK, van der Donk WA, Metcalf WW (2010) Biosynthesis of rhizocticins, antifungal phosphonate oligopeptides produced by *Bacillus subtilis* ATCC6633. *Chem Biol* 17(1):28–37.
31. Eliot AC, et al. (2008) Cloning, expression, and biochemical characterization of *Streptomyces rubellomurinus* genes required for biosynthesis of antimalarial compound FR900098. *Chem Biol* 15(8):765–770.
32. Kim SY, et al. (2012) Different biosynthetic pathways to fosfomycin in *Pseudomonas syringae* and *Streptomyces* species. *Antimicrob Agents Chemother* 56(8):4175–4183.
33. Blodgett JAV, Zhang JK, Metcalf WW (2005) Molecular cloning, sequence analysis, and heterologous expression of the phosphinothricin tripeptide biosynthetic gene cluster from *Streptomyces viridochromogenes* DSM 40736. *Antimicrob Agents Chemother* 49(1):230–240.
34. Woodyer RD, et al. (2006) Heterologous production of fosfomycin and identification of the minimal biosynthetic gene cluster. *Chem Biol* 13(11):1171–1182.
35. Venter JC, et al. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science* 304(5667):66–74.
36. Rusch DB, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 5(3):e77.
37. Jia Y, Lu ZB, Huang K, Herzberg O, Dunaway-Mariano D (1999) Insight into the mechanism of phosphoenolpyruvate mutase catalysis derived from site-directed mutagenesis studies of active site residues. *Biochemistry* 38(43):14165–14173.
38. Zhang G, Dai J, Lu Z, Dunaway-Mariano D (2003) The phosphonopyruvate decarboxylase from *Bacteroides fragilis*. *J Biol Chem* 278(42):41302–41308.
39. Markowitz VM, et al. (2010) The integrated microbial genomes system: An expanding comparative analysis resource. *Nucleic Acids Res* 38(Database issue):D382–D390.
40. Markowitz VM, et al. (2012) IMG/M: The integrated metagenome data management and comparative analysis system. *Nucleic Acids Res* 40(Database issue):D123–D129.
41. Chen CCH, et al. (2006) Structure and kinetics of phosphonopyruvate hydrolase from Variovorax sp. Pal2: New insight into the divergence of catalysis within the PEP mutase/isocitrate lyase superfamily. *Biochemistry* 45(38):11491–11504.
42. Howard EC, Sun S, Biers EJ, Moran MA (2008) Abundant and diverse bacteria involved in DMSP degradation in marine surface waters. *Environ Microbiol* 10(9):2397–2410.
43. Brady A, Salzberg SL (2009) Phymm and PhymmBL: Metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* 6(9):673–676.
44. Watanabe Y, et al. (2001) A novel sphingophosphonolipid head group 1-hydroxy-2-aminoethyl phosphonate in *Bdellovibrio stolpii*. *Lipids* 36(5):513–519.
45. Baumann H, Tzianabos AO, Brisson JR, Kasper DL, Jennings HJ (1992) Structural elucidation of two capsular polysaccharides from one strain of *Bacteroides fragilis* using high-resolution NMR spectroscopy. *Biochemistry* 31(16):4081–4089.
46. Shao ZY, et al. (2008) Biosynthesis of 2-hydroxyethylphosphonate, an unexpected intermediate common to multiple phosphonate biosynthetic pathways. *J Biol Chem* 283(34):23161–23168.
47. Sanders HL (1968) Marine benthic diversity: A comparative study. *Am Nat* 102(925): 243–282.
48. Hurlbert SH (1971) The nonconcept of species diversity: A critique and alternative parameters. *Ecology* 52(4):577–586.
49. Hilderbrand RL, Henderson TO (1983) Phosphonic acids in nature. *The Role of Phosphonates in Living Systems*, ed Hilderbrand RL (CRC Press, Boca Raton, FL), pp 5–30.
50. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M (2007) CAMERA: A community resource for metagenomics. *PLoS Biol* 5(3):e75.
51. Ludwig W, et al. (2004) ARB: A software environment for sequence data. *Nucleic Acids Res* 32(4):1363–1371.
52. Schloss PD, et al. (2009) Introducing mothur: Open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Appl Environ Microbiol* 75(23):7537–7541.
53. Price MN, Dehal PS, Arkin AP (2010) FastTree 2: Approximately maximum-likelihood trees for large alignments. *PLoS ONE* 5(3):e9490.