SCIENTIFIC REPORTS

OPEN

SUBJECT AREAS:
ENVIRONMENTAL MICROBIOLOGY
APPLIED MICROBIOLOGY
METAGENOMICS
MICROBIAL ECOLOGY

Received
21 November 2013

Accepted
7 April 2014

Published
15 May 2014

Correspondence and requests for materials should be addressed to X.-L.W. (xiaolei_wu@pku.edu.cn)

Diverse alkane hydroxylase genes in microorganisms and environments

Yong Nie[1,2], Chang-Qiao Chi[1], Hui Fang[1], Jie-Liang Liang[1], She-Lian Lu[1], Guo-Li Lai[1,2], Yue-Qin Tang[1,3] & Xiao-Lei Wu[1]

[1]College of Engineering, Peking University, Beijing 100871, P. R. China, [2]Institute of Engineering (Baotou), College of Engineering, Peking University, Baotou 014030, China, [3]College of Architecture and Environment, Sichuan University, Chengdu, 610065, China.

AlkB and CYP153 are important alkane hydroxylases responsible for aerobic alkane degradation in bioremediation of oil-polluted environments and microbial enhanced oil recovery. Since their distribution in nature is not clear, we made the investigation among thus-far sequenced 3,979 microbial genomes and 137 metagenomes from terrestrial, freshwater, and marine environments. Hundreds of diverse *alkB* and CYP153 genes including many novel ones were found in bacterial genomes, whereas none were found in archaeal genomes. Moreover, these genes were detected with different distributional patterns in the terrestrial, freshwater, and marine metagenomes. Hints for horizontal gene transfer, gene duplication, and gene fusion were found, which together are likely responsible for diversifying the *alkB* and CYP153 genes adapt to the ubiquitous distribution of different alkanes in nature. In addition, different distributions of these genes between bacterial genomes and metagenomes suggested the potentially important roles of unknown or less common alkane degraders in nature.

Bacterial alkane degradation is important for the bioremediation of petroleum-contaminated environments as well as for microbial enhanced oil recovery[1]. Alkane hydroxylases (AHs) are the key enzymes in aerobic degradation of alkanes by bacteria. These enzymes hydroxylate alkanes to alcohols, which are further oxidized to fatty acids and catabolized via the bacterial β-oxidation pathway[2].

The integral-membrane alkane monooxygenase (AlkB)-related AHs are so-far the most commonly found AHs distributed in both Gram-negative and Gram-positive bacteria[3,4]. Rubredoxin and rubredoxin reductase are the essential electron transfer components needed for alkane hydroxylation by AlkB[5]. The substrates of AlkB are generally *n*-alkanes ranging from C10 to C16[6]. However, in some Actinomycetes, AlkB-type AHs could hydroxylate *n*-alkanes with the chain lengths up to C32 when they were fused with rubredoxin protein[7–9]. The cytochrome P450 CYP153 family is another type of AH for degradation of short- and medium-chain-length *n*-alkanes, which are commonly found in alkane-degrading bacteria lacking AlkB[10]. A number of bacteria have multiple AHs which were proven to potentially expand the *n*-alkane range of the host strain[10]. For example, the co-existence of AlkB and CYP153 was found in *Dietzia* sp. DQ12-45-1b[11,12], as well as the co-existence of multiple AHs found in *Amycolicicoccus subflavus* DQS3-9A1^{T} [13]. Some *Rhodococcus* strains were found to contain more than one *alkB* homologous genes, which have different substrate ranges and induction styles[14,15].

It was reported that more than 60 genera of aerobic bacteria and 5 genera of anaerobic bacteria are able to degrade *n*-alkanes[16]. Although AHs have been intensively studied since the first enzyme was identified in 1977[17], only some alkane degrading bacteria have been subjected to AH analysis. Recently, AH genes have been discovered in new bacterial strains[18,19], indicating the presence of unknown *n*-alkane-degrading bacteria as well as unknown AHs in nature that could be important for natural *n*-alkane metabolism as well as for industrial application such as bioremediation and microbial enhanced oil recovery. However, what are these potential *n*-alkane degrading bacteria? How different are their AH systems? How are they distributed in nature? Why do some of them have multiple AHs genes, whereas others do not?

To answer these questions, we searched for the genes coding for AlkB and CYP153 proteins in the microbial genome and metagenome data deposited in GenBank and the Integrated Microbial Genomes (IMG) system. We then evaluated the diversity of AHs among the microorganisms and different environments, as well as the possible origins of multiple AH genes in one strain, such as gene duplication and gene transfer.

Results

Features of *alkB* and CYP153 genes in microbial genomes. *Distribution in genomes.* In the total 2,069 complete and 1,910 draft microbial genomes, 458 genes for AlkB and 130 genes for CYP153 were found in 369 and 87

Table 1 | Distribution of *alkB* and CYP153 genes in microbial genomes

Phylum (Class)	No. of genomes sequenced	No. of genomes containing *alkB*	No. of *alkB* genes found	No. of genomes containing CYP153	No. of CYP153 genes found
Actinobacteria	424	130	162	26	31
Bacteroidetes	220	14	14	ND	ND
Alphaproteobacteria*	348	55	76	38	63
Betaproteobacteria*	230	70	75	3	3
Gammaproteobacteria*	835	93	125	18	32
Deltaproteobacteria*	126	2	2	ND	ND
Spirochaetes	73	4	4	ND	ND
Planctomycetes	11	ND	ND	1	1

ND, not detected.
*: Classes are provided for the phylum Proteobacteria.

genomes respectively. Neither the *alkB* nor the CYP153 gene was found in archaeal genomes. At the phylum level, the *alkB* genes were only found in Proteobacteria, Actinobacteria, Bacteroidetes, and Spirochaetes. Among them, 278 and 162 *alkB* genes were found in 220 Proteobacteria and 130 Actinobacteria genomes (Table 1), which belonged to at least 51 and 23 genera, respectively (Fig. S1). CYP153 genes were found only in Proteobacteria, Actinobacteria and Planctomycetes. About 98 and 31 CYP153 genes were found in 59 Proteobacteria and 26 Actinobacteria genomes (Table 1), which belonged to at least 20 and 8 genera, respectively (Fig. S2). Of the sequenced genomes, relatively more Actinobacteria contained *alkB* or CYP153 genes (30.7% and 6.1%, respectively) than Proteobacteria did (13.3% and 3.6%, respectively).

At the genus level, *alkB* and CYP153 genes were detected in 85 and 30 genera, respectively (Fig. S1 and S2). Among all the 27 genera that had at least four sequenced genomes, *alkB* genes were detected in all or most of the sequenced genomes belonging to *Frankia*, *Gordonia*, *Mycobacterium*, *Rhodococcus*, *Rhodobacter*, *Burkholderia*, *Acinetobacter*, *Legionella* and *Marinobacter*, suggesting that *alkB* genes could be the core genes shared by these genera (Fig. S3). Similarly, CYP153 genes were found in all the sequenced genomes of *Bradyrhizobium*, *Rhodopseudomonas*, *Caulobacter* and *Novosphingobium*, also indicating their potential core roles in these genera (Fig. S4). In contrast, only one, one, and four genomes were found to have *alkB* genes in 10 *Corynebacterium*, 40 *Streptomyces* and 60 *Enterobacter* sequenced genomes, respectively. Moreover, only one genome was found to contain the CYP153 gene in the 34 *Acinetobacter* and 76 *Burkholderia* genomes sequenced, respectively.

Architectures of AlkB and CYP153. Most of the deduced proteins of *alkB* genes detected had only an AH domain belonging to membrane-FADS-like superfamily and required rubredoxin and rubredoxin reductase genes encoded in separate open reading frames in order to function catalytically. However, 23 genes encoding multiple-domain AlkB proteins were found in this work (Fig. 1). Nineteen of them encoded AlkB-rubredoxin fused proteins, having two conserved domains comprising an N-terminal alkane-hydroxylase domain and a C-terminal rubredoxin domain (Table S1). Seventeen of the 19 genes encoding AlkB-rubredoxin fused proteins were found in Actinobacteria like *Streptomyces*, *Aeromicrobium*, *Gordonia*, *Amycolatopsis*, *Janibacter*, *Pseudonocardia, Saccharomonospora* and *Dietzia*. The remaining four genes encoding multiple-domain AlkB proteins were found in *Leptospira*, *Limnobacter* and *Polaromonas*, comprising an N-terminal ferredoxin domain, a ferredoxin reductase domain and a C-terminal AH domain (Table S1), which has never been reported to our knowledge.

Three CYP153 genes encoding similar fusion proteins were also found in *Gordonia araii* NBRC 100433, *G. polyisoprenivorans* NBRC 16320 and *G. polyisoprenivorans* VH2 consisted of an N-terminal cytochrome P450 domain, a ferredoxin reductase domain, and a C-terminal ferredoxin domain.

Presence of multiple alkane hydroxylase genes. Within the 369 *alkB*-containing and 87 CYP153-containing genomes, 73 and 32 genomes were detected to have multiple copies of *alkB* and CYP153 homologous genes (Table S2 and S3). For example, up to six *alkB* homologous genes and five CYP153 homologous genes were found in *Rhodococcus erythropolis* SK121 and *Parvibaculum lavamentivorans* DS-1, respectively. These multiple copies of genes within one genome shared sequence similarities ranging from 27.7% to 99.7% for *alkB* genes and 42.4% to 100% for CYP153 genes.

Furthermore, both the *alkB* and CYP153 genes were simultaneously detected in 38 genomes that belonged to 16 genera, including *Marinobacter*, *Alcanivorax*, *Rhodococcus* and *Mycobacterium* (Table S4). However, the distribution of these two genes was uneven: 10.3% of the total *alkB* containing genomes harbored CYP153 genes, which was in contrast to the 43.7% of the CYP153-containing genomes harboring *alkB* genes (Fig. S5a). These unbalanced distributions were different in different bacterial phyla (Fig. S5b-e). For example, 24 Actinobacteria genomes had both the *alkB* and CYP153 genes, which occupied 18.5% and 88.9% of the total Actinobacteria genomes containing *alkB* and CYP153 genes, respectively.

Phylogenies of *alkB* and CYP153 genes in microbial genomes. The phylogenetic and comparative genomic analyses revealed that the sequences of AlkB AHs could be clustered into eight clusters (I–VIII) (Fig. 2a). Cluster I included all the sequences from Actinobacteria and one sequence from a Gammaproteobacteria *Salinisphaera shabanensis* E1L3A. Cluster II mainly included most of the sequences from Betaproteobacteria and a few sequences from

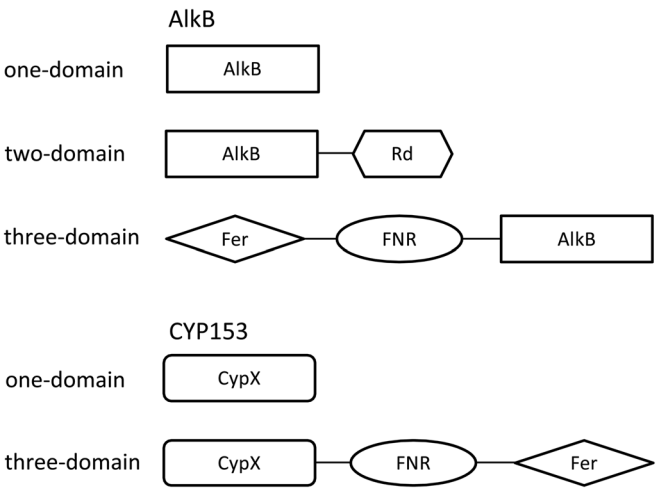

Figure 1 | Conserved domain architecture of AlkB and CYP153. AlkB, AlkB-like alkane hydroxylase; Rd, rubredoxin; Fer, ferredoxin; FNR, ferredoxin reductase; CypX, cytochrome P450.

Gammaproteobacteria like *Marinobacter*, *Pseudomonas*, and *Alcanivorax*. Sequences of Cluster III were all from Gammaproteobacteria, including all the sequences from *Acinetobacter*, and sequences from some *Marinobacter* and *Psychrobacter* species. Sequences in Cluster IV were more diverse, which were from Betaproteobacteria, Gammaproteobacteria, and Alphaproteobacteria. Most of the sequences in Cluster V were from Bacteroidetes, together with a few sequences from Gammaproteobacteria and Spirochaetes. Sequences in Cluster VI were mainly from Gammaproteobacteria, including those from *Pseudomonas aeruginosa*, *Alcanivorax* and *Marinobacter*. Sequences in Cluster VII were all from Alphaproteobacteria and clustered into two groups (Fig. 2a and Fig. S6). Cluster VIII was distinct from the other clusters and included mixed sequences from Alphaproteobacteria, Betaproteobacteria, Gammaproteobacteria, and Spirochaetes. Moreover, sequences in Cluster VIII showed low similarities with sequences in other clusters, For example, two putative *alkB* genes in *Burkholderia* sp. Ch1-1, one of which (ZP_10031959.1) was embedded in Cluster II and another one (ZP_10037453.1) was embedded in Cluster VIII, shared a very low sequence similarity (27.7%) with each other. The functions of the sequences in Cluster VIII are not clear, since none of the gene sequences in this cluster was functionally characterised by experiments such as heterogenous expression or gene knockout analysis. However, their deduced proteins contained all the residues that are conserved in alkane hydroxylases.

The topology of the phylogenetic tree based on the AlkB sequences was largely matched but was different from that based on the 16S rRNA genes. For example, the Proteobacteria AlkB sequences formed a number of distinct clusters, and AlkB sequences from Gammaproteobacteria were distributed among five different clusters (Fig. 2b).

The CYP153 sequences could be clustered into four major clusters (I–IV) (Fig. 3a). Cluster I included two groups with sequences from Actinobacteria and Gammaproteobacteria such as *Alcanivorax*, *Acinetobacter*, and *Marinobacter*, respectively. Cluster II included most of the remaining Gammaproteobacterial sequences and a few sequences from Alphaproteobacteria. Sequences in Clusters III and IV were mainly from Alphaproteobacteria such as *Bradyrhizobium*, *Rhodopseudomonas*, and *Caulobacter*, and mixed with a few sequences from *Mycobacterium* (Actinobacteria), *Polaromonas* (Betaproteobacteria), *Dickeya* (Gammaproteobacteria), *Burkholderia* (Betaproteobacteria), and *Planctomyces* (Planctomycetes). Again, the topology of the CYP153-based phylogenetic tree was different from that of the 16S rRNA gene-based phylogenetic tree. For example, within the Alphaproteobacteria, sequences belonging to the CYP153A group were generally in Cluster III, whereas sequences belonging to the CYP153C and CYP153D groups were in Cluster IV in the CYP153 phylogenetic tree (Fig. 3b).

AH genes could transfer between organisms. For example, the *alkB* operon was found in plasmid OCT in *Pseudomonas putida*, which could easily be subjected to transfer. In *P. mendocina* ymp, a putative *alkB* gene (YP_001185946.1) was located in a predicted genomic island (GI) and had 100% sequence identity with *alkB* genes (CAB54050.1) in the *P. putida* OCT plasmid. In addition, YP_001185946.1 had a much lower G+C content (45.9%) than its host genome (64.7%), also indicating a possible HGT event. Similar HGT could also be found for CYP153 genes. From the clustering in the phylogenetic tree, the potentially same ancient ancestor of CYP153 genes could be found from *Alcanivorax*, *Marinobacter*, *Acinetobacter*, and Actinobacteria (Fig. 3b). Of all the 26 CYP153 gene-containing Actinobacteria genomes (both complete and draft), 18 had over 10% difference in the G+C contents between the CYP153 genes and their host genomes (Table 2). Furthermore, of all the 15 complete Actinobacteria genomes, predicted GIs containing CYP153 genes were found in eight genomes, and plasmids containing CYP153 genes were found in three genomes. All the results indicate the potential common HGT events for CYP153 genes in Actinobacteria.

Distribution of *alkB* and CYP153 genes in metagenomes. Although the sampling sites, metagenomic sequencing and analysis strategies among different samples were obviously different, the average numbers of AH genes in the metagenomes from terrestrial, freshwater and marine samples could more or less reflect their distribution in the three habitats.

From the thus-far available 42 terrestrial, 35 freshwater, and 60 marine metagenomes, the average microbial compositions were calculated (Fig. S7). The results indicated that the terrestrial microbial community was dominated by Proteobacteria (34.9%), Actinobacteria (19.5%), and Cyanobacteria (22%). Proteobacteria (56.9%), Cyanobacteria (12.5%), Bacteroidetes (14.2%) and Actinobacteria (9.6%) were abundant in freshwater environments. In marine metagenomes, bacteria were mainly from Proteobacteria (51.6%), Bacteroidetes (13.4%) and Cyanobacteria (12.1%).

alkB genes in metagenomes. A total of 301, 144, and 524 *alkB* genes were found out of 20,162,506, 4,999,959, and 9,254,226 total proteins predicted in the terrestrial, freshwater and marine metagenomic datasets, respectively. Phylogenetic trees based on these *alkB* gene sequences were constructed with sequences from the genomes and the 28 reference sequences (Table S6). Results indicated that the terrestrial AlkB sequences were mainly clustered with the sequences from Actinobacteria, Gammaproteobacteria and Alphaproteobacteria genomes derived (Fig. S8a), corresponding to Clusters I, VI and VII of the genome AlkB phylogenetic tree, respectively, with some sequences clustered with those from Bacteroidetes in Cluster V (Fig. 2a). The freshwater AlkB sequences were mainly clustered with those from Gammaproteobacteria and Bacteroidetes genomes (Clusters III, IV, V and VI) (Fig. S8b), with some sequences distributed in Clusters VIII. Only two freshwater sequences were related to *alkB* sequences from Actinobacteria in Cluster I. The marine AlkB sequences were mainly distributed in Clusters III, IV, V, and VIII, being closely related to sequences from Gammaproteobacteria and Bacteroidetes. No sequence was found in Cluster I with Actinobacteria, but at least 22 other sequences formed a new cluster distant from any of the seven main clusters (Fig. S8c). In general, the phylogeny of the *alkB* gene sequences revealed that only a few *alkB* sequences retrieved from the three metagenomic databases were closely related (>75% amino acid identities) to the sequences found in microbial genomes or previously identified, suggesting the presence of numerous novel *alkB* genes in the different environments, especially for the marine environment.

Since it is hard to understand the taxonomy of the *alkB* genes from the above phylogenetic trees based on the bacterial genomes and 28 reference sequences, taxonomic analysis was conducted by comparing the deduced proteins of all *alkB* genes against the NR database in GenBank using a BLASTP search and the MEGAN program with the last common ancestor algorithm (Fig. S9). Although it is difficult to assign unknown metagenome-derived AH genes to taxa accurately without examination of flanking gene content because of HGT, here we propose to provide a picture of the potential taxonomic affiliations of metagenome derived AH genes. Our binning approach can be applied in order to facilitate broad comparisons among different samples.

The results showed that only 100 (33.3%), 25 (17.4%) and 52 (9.9%) of the 301, 144, and 524 *alkB* sequences from the terrestrial, freshwater and marine metagenomes, respectively, encoded proteins that were highly similar (>75%) to proteins deposited in the NR database. In contrast, 103 (34.2%), 34 (23.6%), and 290 (55.3%) sequences were < 50% identical to the deposited proteins, respectively (Fig. S10). In addition, only 177 (58.8%), 89 (61.8%), and 363 (69.3%) of the *alkB* sequences in terrestrial, freshwater and marine

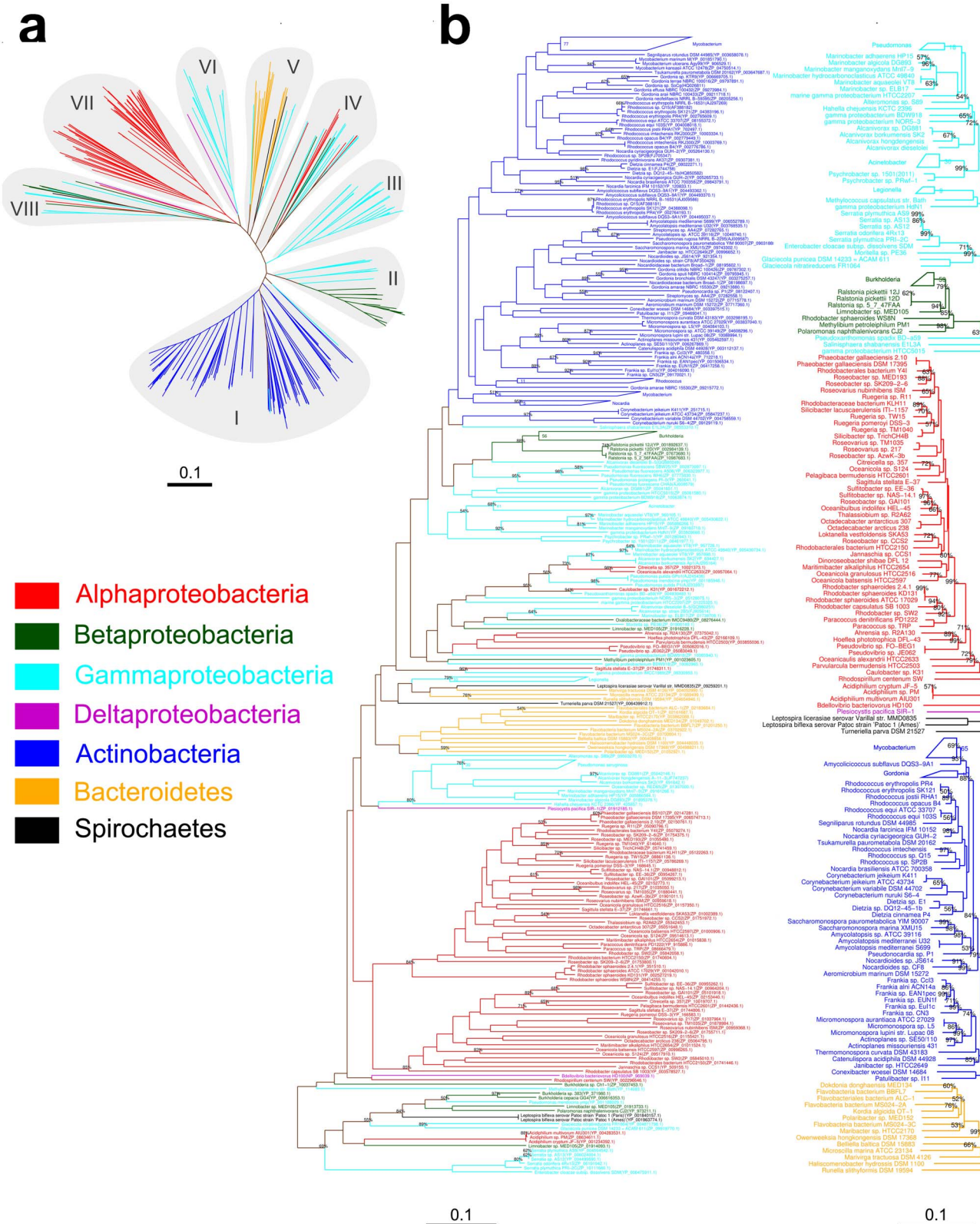

Figure 2 | Phylogenetic distribution of *alkB* genes based on amino acid sequences analysis. A: Major clusters of *alkB* genes. B: Comparison of *alkB* genes and 16S rRNA genes phylogenies. *alkB* (left), 16S rRNA (right). All the sequences were aligned and analyzed by the neighbor-joining method using ARB. The trees were bootstrapped with 1000 replicates. Bootstrap values of >50% are indicated at the respective nodes. The scale bar indicates the percentage of sequence divergence.

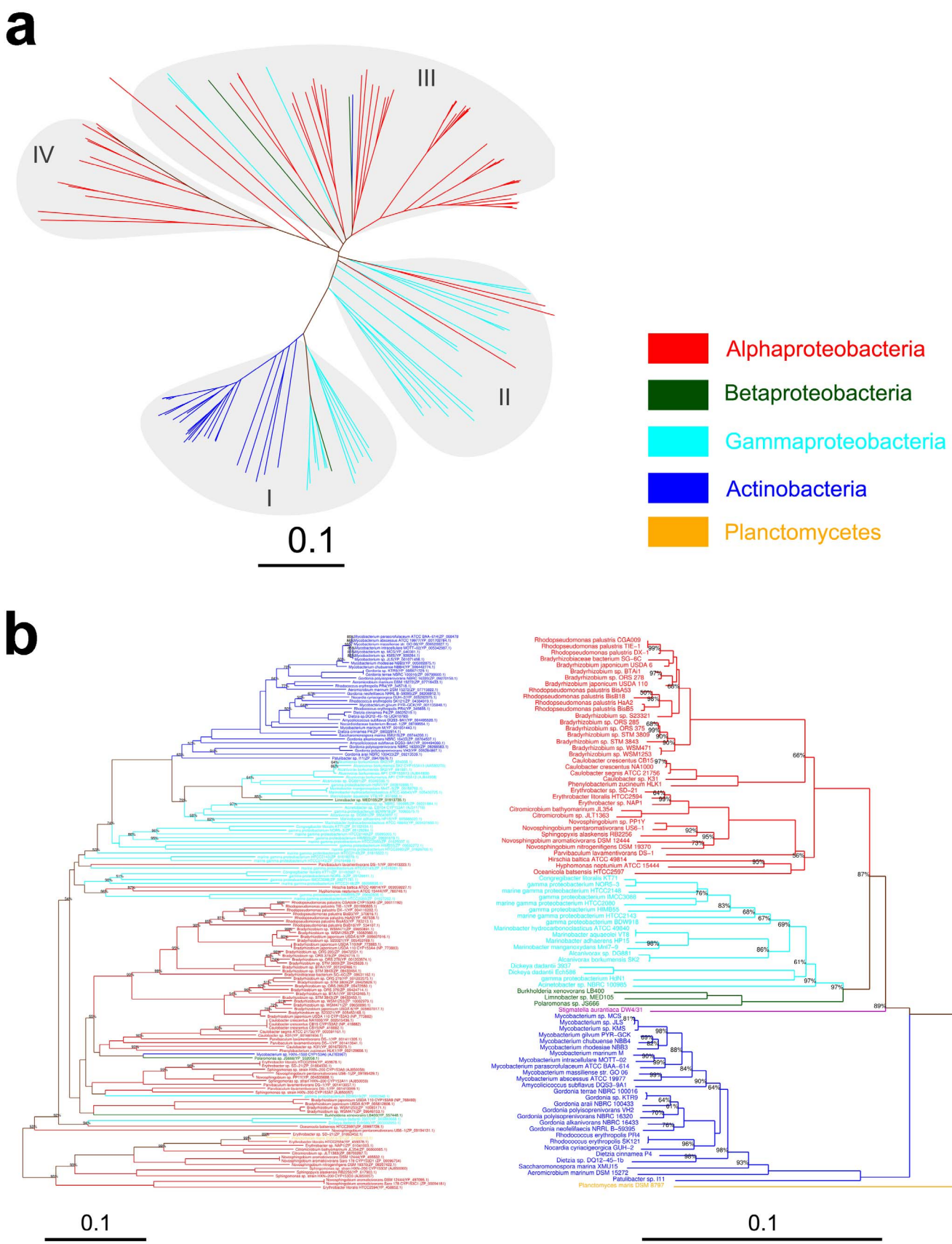

Figure 3 | Phylogenetic distribution of CYP153 genes based on amino acid sequences analysis. A: Major clusters of CYP153 genes. B: Comparison of CYP153 genes (left) and 16S rRNA genes (right) phylogenies. All the sequences were aligned and analyzed by the neighbor-joining method using ARB. The trees were bootstrapped with 1000 replicates. Bootstrap values of >50% are indicated at the respective nodes. The scale bar indicates the percentage of sequence divergence.

Table 2 | Location and G+C contents of CYP153 genes found in Actinobacteria genomes

Accession No.	Source	Location[a]	GIs[b]	CYP153 gene G+C content[c]	G+C content of genome	ratio[c]
YP_004494060.1	*Amycolicicoccus subflavus* DQS3-9A1	C	-	58%	62.2%	1.07
YP_004495520.1	*Amycolicicoccus subflavus* DQS3-9A1	C	+	56.2%	62.2%	1.11
ZP_07715822.1	*Aeromicrobium marinum* DSM 15272 (D)	NA	NA	61.2%	71.1%	1.16
ZP_07716433.1	*Aeromicrobium marinum* DSM 15272 (D)	NA	NA	61.6%	71.1%	1.15
ZP_08022914.1	*Dietzia cinnamea* P4 (D)	NA	NA	61.9%	71%	1.15
ZP_08025219.1	*Dietzia cinnamea* P4 (D)	NA	NA	57.3%	71%	1.24
YP_005284867.1	*Gordonia polyisoprenivorans* VH2	C	-	63.5%	67%	1.06
ZP_09268583.1	*Gordonia polyisoprenivorans* NBRC 16320 (D)	NA	NA	63.9%	66.9%	1.05
ZP_09270150.1	*Gordonia polyisoprenivorans* NBRC 16320 (D)	NA	NA	60.8%	66.9%	1.10
ZP_08206912.1	*Gordonia neofelifaecis* NRRL B-59395 (D)	NA	NA	60.2%	68.2%	1.13
ZP_08764507.1	*Gordonia alkanivorans* NBRC 16433 (D)	NA	NA	61.7%	67.4%	1.09
ZP_09213539.1	*Gordonia araii* NBRC 100433 (D)	NA	NA	64.9%	68%	1.05
ZP_09798600.1	*Gordonia terrae* NBRC 100016 (D)	NA	NA	59%	67.8%	1.15
YP_006671729.1	*Gordonia sp.* KTR9	P	NA	59%	67.5%	1.14
YP_001702784.1	*Mycobacterium abscessus* ATCC 19977	C	-	57.8%	64.1%	1.11
YP_006442774.1	*Mycobacterium chubuense* NBB4	P	NA	58.6%	68.3%	1.17
YP_001135848.1	*Mycobacterium gilvum* PYR-GCK	C	+	57.6%	67.7%	1.18
YP_005342987.1	*Mycobacterium intracellulare* MOTT-02	C	+	57.4%	68.1%	1.19
YP_001071498.1	*Mycobacterium* sp. JLS	C	+	57%	68.4%	1.18
YP_939264.1	*Mycobacterium* sp. KMS	C	+	57.9%	68.2%	1.18
YP_001851443.1	*Mycobacterium marinum* M	C	+	59.7%	65.7%	1.10
YP_006520827.1	*Mycobacterium massiliense str. GO 06*	C	NA	57.8%	64.2%	1.11
YP_640381.1	*Mycobacterium* sp. MCS	C	+	57%	68.4%	1.18
YP_005002875.1	*Mycobacterium rhodesiae* NBB3	C	-	57.6%	65.5%	1.14
ZP_06847816.1	*Mycobacterium parascrofulaceum* ATCC BAA-614 (D)	NA	NA	57.4%	68.5%	1.19
ZP_08199554.1	*Nocardioidaceae bacterium* Broad-1 (D)	NA	NA	58.2%	69.6%	1.20
YP_005262975.1	*Nocardia cyriacigeorgica* GUH-2	C	+	59%	68.4%	1.16
ZP_09470676.1	*Patulibacter* sp. I11 (D)	NA	NA	64.6%	74.1%	1.15
YP_345695.1	*Rhodococcus erythropolis* PR4	P	NA	57%	62.3%	1.09
YP_345718.1	*Rhodococcus erythropolis* PR4	P	NA	60.9%	62.3%	1.02
ZP_04384919.1	*Rhodococcus erythropolis* SK121 (D)	NA	NA	58.9%	62.4%	1.07

[a]: Location of CYP153 genes. C, genes located in chromosome; P, genes located in plasmid; NA, no data available.
[b]: Genes found in predicted genomic islands (GIs). +, genes located in GI; -, genes not located in GI; NA, no data available.
[c]: CYP153 gene versus average genome G+C content'.

metagenomes could be aligned to a phylum. All these results suggested a vast amount of novel *alkB* genes in the different environments.

Among the AlkB sequences assigned to a phylum, 69 (39.0%), 61 (68.5%), and 220 (60.6%) sequences were from Proteobacteria in comparison with Proteobacteria occupying 34.9%, 56.9% and 51.6% of the total metagenomic communities in the terrestrial, freshwater and marine metagenomes, respectively (Fig. 4a and Fig. S7). In addition, 22 (25.8%) and 130 (36.9%) AlkB sequences in the freshwater and marine metagenomes were found to be related to Bacteroidetes, although the relative abundances of Bacteroidetes were only 14.2% and 13.4% in the two environments, respectively (Fig. 4a and Fig. S7). In contrast, 97 (54.8%) and 5 (5.6%) sequences were found related to Actinobacteria, in comparison with Actinobacteria occupying 19.5% and 9.6% of the total metagenomic communities in the terrestrial and freshwater metagenomes, respectively (Fig. 4a and Fig. S7). These results suggest that AlkB sequences from Actinobacteria were enriched in terrestrial metagenomes, whereas those from Bacteroidetes were enriched in aquatic environments. No AlkB sequence from Actinobacteria was found in marine metagenomes.

Furthermore, only about 22.6%, 15.3%, and 30.9% of the total 301, 144, and 524 AlkB sequences in the terrestrial, freshwater and marine metagenomes, respectively, could be assigned to a genus, including *Conexibacter*, *Pseudomonas*, *Mycobacterium*, *Acidiphilium*, and *Polaribacter* (Fig. S9), many of which are not known as alkane degraders. Among them, sequences related to *Polaribacter* were found in all three habitats. Sequences related to *Conexibacter*, *Mycobacterium*, *Hyphomicrobium* and *Rhodococcus* were only found in terrestrial metagenomes, and those related to *Methylophaga*, *Burkholderia*, *Methylomicrobium*, *Leptospira*, *Kordia*, *Haliscomenobacter*, *Roseobacter*, *Ahrensia*, and *Ralstonia* were only found in marine metagenomes, whereas *Marivirga* was unique to the freshwater (Fig. S11).

CYP153 genes in metagenomes. Using similar analyses as applied for the *alkB* genes, 585, 43, and 332 CYP153 homologous genes were found in the terrestrial, freshwater and marine metagenomic datasets, respectively. The relative abundance of candidate CYP153 genes in freshwater were much less than in terrestrial and marine metagenomes (Fig. S12). The phylogenetic tree based on CYP153 sequences showed that the CYP153 sequences were only from Proteobacteria and Actinobacteria. Among them, most terrestrial sequences were in Clusters III and IV, with sequences from Alphaproteobacteria. The remaining ones were distributed in Cluster II, related to Gammaproteobacteria (Fig. S13a). Freshwater CYP153 sequences were mainly distributed in Clusters III and IV, with those related to Alphaproteobacteria (Fig. S13b). Most marine CYP153 sequences were in Cluster II, affiliated with sequences from Gammaproteobacteria. The remaining sequences were mainly in Cluster III, with sequences from Alphaproteobacteria (Fig. S13c). Analysis of deduced protein sequences of CYP153 genes by BLASTP showed that only about 36.9%, 30.2%, and 11.1% of the total 585, 43, and 332 candidate CYP153 genes in the terrestrial, freshwater and marine metagenomic datasets, respectively, were >75% amino acid identical to the deposited genes in the NR database (Fig. S14). About 80.5%, 81.4%, and 75% of the total CYP153 sequences in the respective terrestrial, freshwater and marine meta-

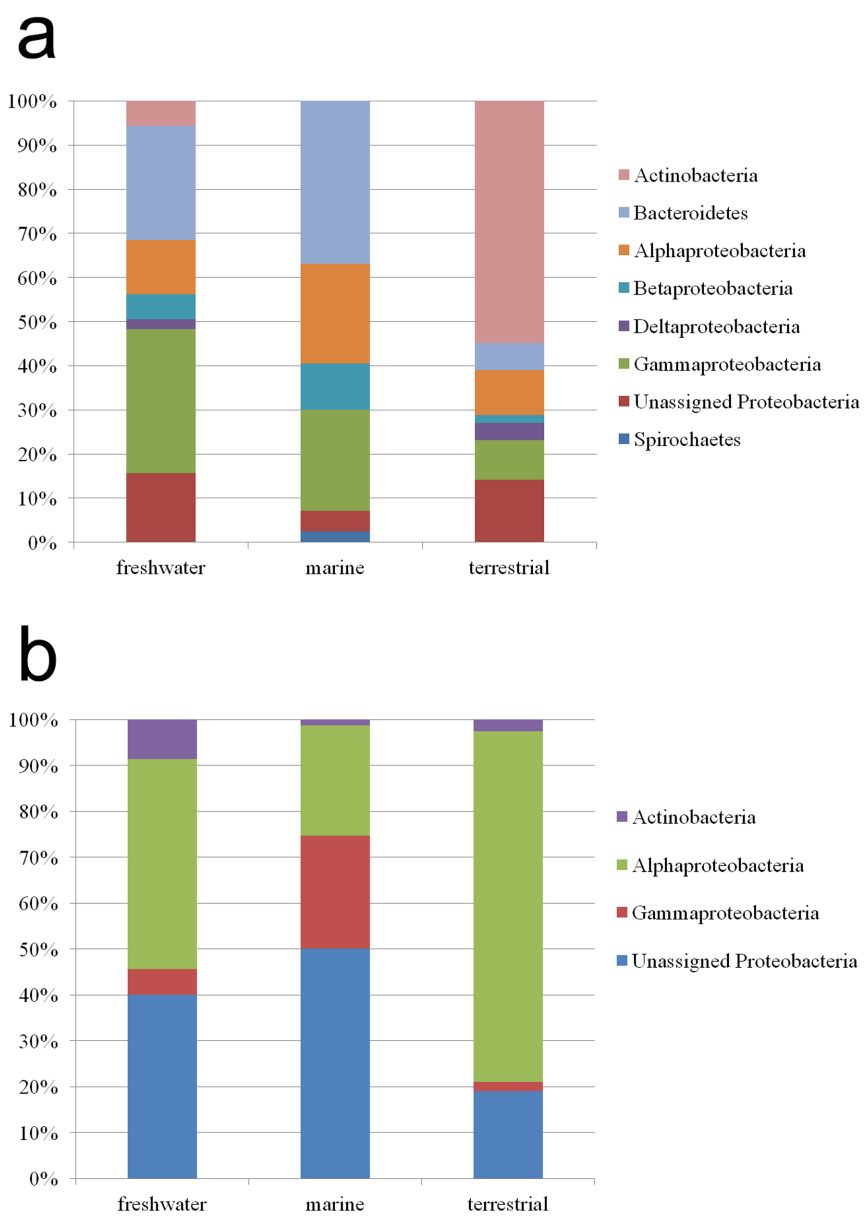

Figure 4 | Taxonomic distribution of alkane hydroxylases in freshwater, marine and terrestrial habitats. A: Taxonomic distribution of *alkB* genes. B: Taxonomic distribution of CYP153 genes.

genomes could be assigned to the phylum level. Among these sequences assigned to a phylum, CYP153 genes related to Alphaproteobacteria were well distributed in all three environments. In contrast, CYP153 sequences related to Gammaproteobacteria were most enriched in marine environments (Fig. 4b). Only about 24.1%, 11.6%, and 10.8% of total 585, 43, and 332 CYP153 sequences could be assigned to the genus level (Fig. S15), including *Bradyrhizobium*, *Caulobacter, Phenylobacterium,* and *Parvibaculum* belonging to Alphaproteobacteria in terrestrial metagenomes, *Sphingopyxis* in freshwater, and *Parvibaculum*, *Hyphomonas*, *Bradyrhizobium*, *Caulobacter*, and *Sphingopyxis* belonging to Alphaproteobacteria in marine metagenomes (Fig. S15).

Discussion

Unexpectedly diverse *alkB* and CYP153 genes were detected in different environments and in so many bacterial genomes, including those that do not have the proven alkane hydroxylation functions like genus such as *Conexibacter*, *Acidiphilium*, *Methylibium*, and *Leptospira* (Table S5). Among these bacteria, some of them including *Acidiphilium*[20], *Parvularcula*[21], *Limnobacter*[22], and *Glaciecola*[23], had been found in oil contaminated environments, whereas many other bacteria, including those with proven alkane degradation abilities, were not originally isolated from petroleum-contaminated environments. Despite that nature is not always petroleum-contaminated, alkanes are in fact observed throughout nature although in low concentrations[24]. They can be produced from fatty acid metabolites of plants, insects, and microorganisms. For example, to protect against water loss and pathogen infection, plants can produce cuticular waxes at the epidermal cells[25], which typically constitute 20–60% of the cuticle mass and are complex mixture of straight-chain C20–C60 aliphatics[26]. Insects produce pheromones with the hydrocarbon backbone[27]. Alkanes have also been reported to be synthesized in diverse microorganisms, including cyanobacteria that were reported to produce high proportions of heptadecane[28–30]. Therefore, alkanes are always observed in natural habitats dominated by cyanobacteria[31].

The ubiquitous presence of alkanes in nature could have played important roles in the evolution of hydroxylase genes in different environments, resulting in the detection of so many *alkB* and

CYP153 genes from both the microbial genomes and the various metagenomes. In addition, the different alkane availabilities, either in amount or in molecular structure, along with different microbial communities in environments could have enriched different *alkB* and CYP153 genes in different habitats. For example, the *alkB* genes were much less diverse in freshwater than in both marine and terrestrial habitats, which might be a consequence of the much less alkane availability in freshwater environments. Similarly, the higher abundance of Actinobacteria in terrestrial than in marine environments[32,33] could be one of the reasons why more *alkB* genes related to Actinobacteria were found in terrestrial metagenomes (Fig. S7). Moreover, the soil microbial communities were shaped by a complex interaction of soil variables, such as salinity, pH, total carbon, and availability of oxygen. Among these variables, the presence of total petroleum hydrocarbons and different bioavailability of alkanes had a significant effect on the structure of alkane-degrading communities[34,35]. The enrichment of Bacteroidetes *alkB* genes in freshwater, as well as the detection of different alkane degraders in the metagenomes is likely attributed to the different availabilities of alkanes and the presence of different bacterial communities. However, further research is needed to investigate the exact reasons. Remarkably, different distributions of *alkB* and CYP153 genes were also found between the microbial genomes and the metagenomes. For example, although a large number of *alkB* sequences were found in *Acinetobacter* genomes, no sequence related to *Acinetobacter* was found in any of the metagenomes. These inconsistencies could result from the difference between the culture-dependent and -independent analytical approaches. A recent research on the effects of different methods used in analysis of soil microbial communities showed unexpected accessibility of the rare biosphere by culturing. Soil bacteria captured by culturing, such as *Pseudomonas*, *Rhodococcus*, *Arthrobacter* and *Flavobacterium* which were abundant among the cultured organisms, were in very low abundance or absent in the culture independent community[36]. In contrast, some bacteria abundant in culture independent community were not cultured. It also suggested that many bacteria harboring *alkB* and CYP153 genes have not been isolated yet (Fig. S9 and S15). However, it can hardly be concluded that these usually isolated alkane-degrading strains that are less dominant in the various environments cannot play important roles in alkane degradation, and vice versa, because the low-abundant bacteria could burst forth under certain environmental stresses, like an oil spill, and be finally important[37].

Although members of *Bacillus* and *Geobacillus* were often detected in oil-related environments, and a number of *Bacillus* and *Geobacillus* strains were reported to be able to utilize long-chain *n*-alkanes as the sole carbon and energy source[18,38,39], neither the *alkB* gene nor the CYP153 gene was found in the 1,077 sequenced Firmicutes genomes, including 139 *Bacillus*, 10 *Geobacillus* and one *Thermobacillus* genomes. Moreover, no *alkB* or CYP153 homologous genes from Firmicutes were found in all the three metagenomic datasets. Although it was reported that *alkB* genes were detected in alkane degraders belonging to *Geobacillus*, the *alkB* gene fragments had remarkable high amino acid sequence similarities with those in *Rhodococcus*, which suggested that *alkB* genes in *Geobacillus* were obtained via HGT from *Rhodococcus* or other Actinobacteria[40]. It therefore seemed that *alkB* and CYP153 were not the key genes for alkane degradation in *Bacillus* and *Geobacillus*. Whether the alkane-degrading Firmicutes have other AHs needed to be further researched. At least one novel soluble long-chain alkane monooxygenase (LadA) was found in *Geobacillus thermodenitrificans* NG80-2, catalyzing the first alkane hydroxylation reaction in the alkane-degrading pathway[18].

The generation of novel functions is critical for microorganisms in order to be able to respond to environmental and evolutional challenges. Gene duplication, HGT, and gene fusion/fission are common evolutionary processes that generate novel genes or functions for rapid adaptation[41–46]. Among the genes, rRNA genes are thought to be the most conserved, and the rRNA-based phylogeny is considered to be robust and consistent with the genome phylogeny[47]. In general, highly similar clustering of Actinobacteria and Betaproteobacteria were found between the 16S rRNA and AlkB trees, as were the orderings of Gammaproteobacteria between the 16S rRNA and CYP153 trees. The results indicated that *alkB*/CYP153 genes might have occurred earlier than the speciation events and were inherited from their ancestors. Major inconsistences between the 16S rRNA and AlkB trees were found for some members from Bacteroidetes, Gammaproteobacteria and Alphaproteobacteria. In contrast, major inconsistencies between the 16S rRNA and CYP153 trees were found for all members from Actinobacteria, which were shown as the outgroup of Proteobacteria in the 16S rRNA tree but clustered with some members from Gammaproteobacteria in the CYP153 tree (Fig. 3). The inconsistencies suggest different origins of these *alkB*/CYP153 genes by different processes.

Paralogs are homologs that arise through gene duplication events. They usually form two groups sharing identical phylogenies in the phylogenetic tree, and are connected by a branch that indicates their last common ancestor[48]. Based on these definitions, both *alkB* and CYP153 paralogs were found in some genomes containing multiple AH genes. For example, *alkB* paralogs were found in *Pseudomonas aeruginosa* (Cluster VI) and Alphaproteobacteria (Cluster VII) (Fig. S6). AlkB sequences from Alphaproteobacteria in Cluster VII were clustered into two groups. Except for the genomes of *Pelagibaca bermudensis* HTCC2601, *Octadecabacter arcticus* 238 and *Rhodobacter capsulatus* SB 1003, genomes in group 2 always had more than two *alkB* genes: one was in group 2 and the others in group 1. Although in two groups, these genes shared similar G+C content with their host genomes, clustered together, and were distant from other clusters of the *alkB* genes from other bacterial classes. Moreover, genomes from *Pseudomonas aeruginosa* strains contained two copies of the *alkB* gene. They clustered into two groups that were distant from other bacterial taxa and shared similar G+C contents with their host genomes, indicating that they might be paralogs separated during gene duplication. Similarly, CYP153 genes in Cluster IV were possibly paralogous genes because of the deep bifurcations and long branch lengths (Fig. 3b), which indicated the more rapid evolutionary rate than orthologous genes.The paralogous genes may have different abilities and functions. For example, *alkW1* and *alkW2* are two paralogous genes encoding AlkB-rubredoxin fusion proteins found in *Dietzia* sp. DQ12-45-1b. Functional research showed that *alkW1* could hydroxylate *n*-alkanes ranging from C14 to C32, whereas *alkW2* was not expressed[8].

HGT is a potent evolutionary force in prokaryotes. Genes acquired by HGT can be predicted by comparing the differences of G+C content, codon usage, phylogenies between the candidate genes and the whole genomes, as well as analyzing the flanking mobile genetic elements[49,50]. It is common for catabolic genes to undergo genetic rearrangements, such as insertions, deletions, duplications and inversions, which is attributable to the presence of elements that possess the ability to mobilise the catabolic genes[51]. One example is *Pseudomonas putida* GPo1 *alkB* gene that located in the OCT plasmid[52,53]. The G+C content of GPo1 *alkB* gene is much lower than that of the host strain and the OCT plasmid, and it is flanked by the insertion sequence ISPpu4, constituting a class 1 transposon, suggesting this gene is part of a mobile element[53] and obtained from some closely related *Alcanivorax* strains[54]. The possible HGT of *alkB* genes may explain why several *alkB* genes from *Pseudomonas* are distributed in several different clusters, which is also proposed before[55]. Interestingly, almost all the Actinobacteria genomes containing CYP153 genes had *alkB* genes, suggesting a potential link between the CYP153 and *alkB* genes in the Actinobacteria. It is reasonable that bacteria need not only AHs, but also enzymes for sensing, taking up, and emulsifying alkanes to degrade them.

Successful HGT of both the *alkB* and CYP153 genes could therefore more easily occur between alkane degraders which originally have the entire alkane degradation-related systems[56].

Gene fusion, in which two previously independent genes fuse to a single contiguous open reading frame, is thought to be an important source of evolutionary novelty. It contributed significantly to the rapid evolution of novel biological functions[57]. Although the substrate length of AlkB homologous AHs were often less than C16, recent research has proven that the AlkB-rubredoxin fusion could enlarge the long-chain *n*-alkane degradation spectrum. The fusion of the two genes could be beneficial either for electron transfer or for substrate-enzyme binding[8]. In general, bacterial P450s usually need ferredoxin and ferredoxin reductase for electron transfer. Self-sufficient P450s with other reductase domains were characterized by their remarkable enzymatic activities[58–60], supporting that gene fusion could be beneficial for the evolution of novel functions. In this study, the fusion of AlkB, ferredoxin and ferredoxin reductase was detected for the first time in *Leptospira*, *Limnobacter* and *Polaromonas*. Although rubredoxin and rubredoxin reductase could be replaced *in vitro* by ferredoxin and ferredoxin reductase, respectively[17,61], further study is needed to address whether ferredoxin and ferredoxin reductase could function as rubredoxin and rubredoxin reductase*in vivo*, or vice versa.

Although rapid evolutionary evidences like gene duplication and HGT could explain the inconsistences of AHs and 16S rRNA phylogenies, some inconsistent branches could not be explained, at least by recent evolutionary events. For example, AlkB sequences related to Bacteroidetes were embedded in the Proteobacteria and clustered with Gammaproteobacteria in the AlkB-based phylogenetic tree (Cluster V). GI prediction and analysis of the G+C content could not reveal an obvious recent HGT event of *alkB* genes both in Bacteroidetes and Gammaproteobacteria. Neither were the paralogs found in these cases. AlkB sequences from *Pseudomonas fluorescens* and *Pseudomonas aeruginosa* separated into two distinct clusters, but no obvious evidence was found to support the recent rapid evolutionary events like gene duplication and HGT in these strains. It suggested that *alkB* genes in these strains might occur after the speciation events. However, the detailed mechanisms need to be further studied.

In a summary, hundreds of putative *alkB* and CYP153 genes, most of which are novel ones with lower identity to the known genes, were retrieved by mining the released microbial genome and metagenome databases. They were found in Proteobacteria, Actinobacteria, Bacteroidetes, Spirochaetes and Planctomycetes, but not in archaea. The rapid evolutionary events like HGT, gene duplication, and gene fusion likely contributed to the diversity of both the *alkB* and CYP153 genes. Moreover, *alkB* and CYP153 genes, with many being unknown and less common, were found distributing differently in the terrestrial, freshwater and marine environments, suggesting their potential contributions to alkane metabolism in nature. Although the functions and evolution of AH genes and the mechanism of how organisms and genes are selected in different habitats need to be further researched, our work provides an overall profile of the *alkB* and CYP153 gene distributions in microbes and environments which can help to understand the gene and microbial functions toward alkane degradation in different environments.

Methods

Generation of protein sequence database. All the available microbial genomic data (up to September of 2012) including their predicted open reading frames were downloaded from the NCBI genome database in September 2012, which consisted of 2,069 complete and 1,910 draft microbial genomes representing 784 different genera. Among them, Archaea had 134 complete and 18 draft genomes, representing 72 genera. A total of 6,534,587 and 8,864,443 protein sequences were obtained from all the complete and draft genomes, respectively. Finally, a microbial proteomic database containing 15,399,030 protein sequences was built.

Search for alkane hydroxylases in the microbial genomes. First, protein sequences of 28 AlkB and 18 CYP153 functionally characterized enzymes were selected as references. They were aligned using the MUSCLE program[62] and the alignment was manually adjusted. The aligned sequences were then used to construct a first profile Hidden Markov Model (pHMM)[63] for AlkB and CYP153, respectively, by using HMMER3 software package[64].

The first AlkB and CYP153 pHMMs were then used to search the AlkB and CYP153 sequences in the microbial proteomic database with 15,399,030 protein sequences by using the HMMER3 software. The resulting positive hits of both AlkB and CYP153 were then aligned with the MUSCLE program and manually filtered with the following criteria: for AlkB proteins, hit sequences without three histidine boxes and HYG motif were excluded[4,65]; for CYP153 proteins, the hit sequences were compared against the bacterial cytochrome P450 database (http://drnelson.uthsc.edu/CytochromeP450.html), and those not only with the best hits but also >40% identity to CYP153 family members in the database were selected[66]. The new CYP153 protein sequences were added to update the bacterial P450 database. Finally, the obtained AlkB and CYP153 protein sequences with their respective reference sequences were aligned and applied to build the second pHMMs, respectively. The second pHMMs were used to again search against the microbial proteomic database for new AlkB and CYP153 proteins as described above, repeating in the same manner until no new sequence was found. The obtained sequences were then used for the following analyses, including phylogenetic and evolutionary analyses. To identify the possible conserved domains, the obtained proteins were analyzed using the Simple Modular Architecture Research Tool (SMART, http://smart.embl-heidelberg.de)[67].

The GenBank accession numbers or Gene ID of all the AH genes identified in this study were shown in Table S7 and Table S8.

Phylogenetic analysis. The 16S rRNA genes were extracted from the genome database, as well as from the Ribosomal Database Project[68] and Silva database[69]. They, along with the AH sequences obtained above, were used to construct the phylogenetic trees using ARB[70] by neighbor-joining algorithm[71]. The stability of tree topology was evaluated by bootstrap resampling with a total of 1,000 replicates in all cases. The 16S rRNA gene from *Methanobacterium curvum*, XylM protein from *Pseudomonas putida* mt-2 and P450cam protein were selected as the outgroups for building the 16S rRNA, AlkB and CYP153 phylogenetic trees, respectively.

Analysis of horizontal gene transfer events. The AlkB and CYP153 phylogenies were compared against the 16S rRNA gene-based trees. The inconsistent species in the two trees were subjected to the analysis of genomic islands (GIs) using IslandViewer (http://www.pathogenomics.sfu.ca/islandviewer)[72]. The G+C contents of these inconsistent sequences were also calculated to compare with their host genomes. Genes located in predicted GIs or with considerable G+C content differences were postulated to have high HGT potentials.

Search for alkane hydroxylases in different metagenomes. The total thus-far available (up to September 2012) 137 assembled metagenome datasets including their open reading frames from the IMG/M database[73] were downloaded, 42, 35 and 60 of which were from terrestrial, freshwater and marine environments, respectively. The phylogenetic information of each sample was also downloaded from the "Radial Tree Distribution" in IMG/M and the average microbial compositions were calculated. For example, for calculating the average microbial composition of terrestrial metagenomes, the number of hits for each different taxon (Phylum level) in each terrestrial metagenome were summarized, and then divided by the number of total hits in all terrestrial metagenomes. In total, 20,162,506, 4,999,959 and 9,254,226 proteins were obtained from these respective selected terrestrial, freshwater and marine metagenomes, respectively. Similarly, the above-generated AlkB and CYP153 pHMMs were used to search against the metagenome protein databases using the HMMER3 software. The positive hits were aligned and manually filtered as described above. The obtained sequences were compared against the NCBI NR database using BLASTP and the BLAST results were imported into MEGAN4[74] for taxonomic analysis using the last common ancestor based algorithm. Phylogenetic trees were built based on the sequences from metagenomic data together with those from microbial genomes using ARB, as described above.

1. Brown, L. R. Microbial enhanced oil recovery (MEOR). *Curr. Opin. Microbiol.* **13**, 316–320 (2010).
2. Rojo, F. Degradation of alkanes by bacteria. *Environ. Microbiol.* **11**, 2477–2490 (2009).
3. Smits, T. H. M., Balada, S. B., Witholt, B. & van Beilen, J. B. Functional analysis of alkane hydroxylases from Gram-negative and Gram-positive bacteria. *J. Bacteriol.* **184**, 1733–1742 (2002).
4. Smits, T. H. M., Rothlisberger, M., Witholt, B. & van Beilen, J. B. Molecular screening for alkane hydroxylase genes in Gram-negative and Gram-positive strains. *Environ. Microbiol.* **1**, 307–317 (1999).
5. van Beilen, J. B. *et al.* Rubredoxins involved in alkane oxidation. *J. Bacteriol.* **184**, 1722–1732 (2002).
6. van Beilen, J. B. & Funhoff, E. G. Alkane hydroxylases involved in microbial alkane degradation. *Appl. Microbiol. Biotechnol.* **74**, 13–21 (2007).
7. Bihari, Z. *et al.* Functional analysis of long-chain n-alkane degradation by *Dietzia* spp. *FEMS Microbiol. Lett.* **316**, 100–107 (2011).

8. Nie, Y., Liang, J., Fang, H., Tang, Y. Q. & Wu, X. L. Two novel alkane hydroxylase-rubredoxin fusion genes isolated from a *Dietzia* bacterium and the functions of fused rubredoxin domains in long-chain *n*-alkane degradation. *Appl. Environ. Microbiol.* **77**, 7279–7288 (2011).
9. Lo Piccolo, L., De Pasquale, C., Fodale, R., Puglia, A. M. & Quatrini, P. Involvement of an alkane hydroxylase system of *Gordonia* sp. strain SoCg in degradation of solid *n*-alkanes. *Appl. Environ. Microbiol.* **77**, 1204–1213 (2011).
10. van Beilen, J. B. *et al*. Cytochrome P450 alkane hydroxylases of the CYP153 family are common in alkane-degrading eubacteria lacking integral membrane alkane hydroxylases. *Appl. Environ. Microbiol.* **72**, 59–65 (2006).
11. Wang, X. B. *et al*. Degradation of petroleum hydrocarbons (C6-C40) and crude oil by a novel *Dietzia* strain. *Bioresour. Technol.* **102**, 7755–7761 (2011).
12. Nie, Y., Liang, J.-L., Fang, H., Tang, Y.-Q. & Wu, X.-L. Characterization of a CYP153 alkane hydroxylase gene in a Gram-positive *Dietzia* sp. DQ12-45-1b and its "team role" with *alkW1* in alkane degradation. *Appl. Microbiol. Biotechnol.* **98**, 163–173 (2014).
13. Nie, Y. *et al*. The genome of the moderate halophile *Amycolicicoccus subflavus* DQS3-9A1^T reveals four alkane hydroxylation systems and provides some clues on the genetic basis for its adaptation to a petroleum environment. *PloS one* **8**, e70986 (2013).
14. Whyte, L. G. *et al*. Gene cloning and characterization of multiple alkane hydroxylase systems in *Rhodococcus* strains Q15 and NRRL B-16531. *Appl. Environ. Microbiol.* **68**, 5933–5942 (2002).
15. Amouric, A. *et al*. Identification of different alkane hydroxylase systems in *Rhodococcus ruberstrain* SP2B, an hexane-degrading actinomycete. *J. Appl. Microbiol.* **108**, 1903–1916 (2009).
16. Prince, R. C. *Petroleum Microbiology* (ASM press, Washington, 2005).
17. Benson, S., Fennewald, M., Shapiro, J. & Huettner, C. Fractionation of inducible alkane hydroxylase activity in *Pseudomonas putida* and characterization of hydroxylase-negative plasmid mutations. *J. Bacteriol.* **132**, 614–621 (1977).
18. Feng, L. *et al*. Genome and proteome of long-chain alkane degrading *Geobacillus thermodenitrificans* NG80-2 isolated from a deep-subsurface oil reservoir. *Proc. Natl Acad. Sci. USA* **104**, 5602–5607 (2007).
19. Li, P., Wang, L. & Feng, L. Characterization of a novel Rieske-type alkane monooxygenase system in *Pusillimonas* sp. T7-7. *J. Bacteriol.* **195**, 1892–1901 (2013).
20. Röling, W. F. M., Ortega-Lucach, S., Larter, S. R. & Head, I. M. Acidophilic microbial communities associated with a natural, biodegraded hydrocarbon seepage. *J. Appl. Microbiol.* **101**, 290–299 (2006).
21. Yergeau, E., Sanschagrin, S., Beaumier, D. & Greer, C. W. Metagenomic analysis of the bioremediation of diesel-contaminated Canadian high arctic soils. *PloS one* **7**, e30058 (2012).
22. Wen, Z., Jizhou, D., Liye, C. & Hongbo, S. Isolation of a nitrate-reducing bacteria strain from oil field brine and the inhibition of sulfate-reducing bacteria. *African Journal of Biotechnology* **10**, 10019–10029 (2011).
23. Brakstad, O. G., Nonstad, I., Faksness, L.-G. & Brandvik, P. J. Responses of microbial communities in Arctic sea ice after contamination by crude petroleum oil. *Microb. Ecol.* **55**, 540–552 (2008).
24. Schirmer, A., Rude, M. A., Li, X., Popova, E. & Del Cardayre, S. B. Microbial biosynthesis of alkanes. *Science* **329**, 559–562 (2010).
25. Samuels, L., Kunst, L. & Jetter, R. Sealing plant surfaces: cuticular wax formation by epidermal cells. *Plant Biol.* **59**, 683 (2008).
26. Heredia, A. Biophysical and biochemical characteristics of cutin, a plant barrier biopolymer. *BBA-Gen. Subjects* **1620**, 1–7 (2003).
27. Tillman, J. A., Seybold, S. J., Jurenka, R. A. & Blomquist, G. J. Insect pheromones—an overview of biosynthesis and endocrine regulation. *Insect Biochem. Mol. Biol.* **29**, 481–514 (1999).
28. Winters, K., Parker, P. & Van Baalen, C. Hydrocarbons of blue-green algae: geochemical signfficance. *Science* **163**, 467–468 (1969).
29. McInnes, A. G., Walter, J. A. & Wright, J. L. Biosynthesis of hydrocarbons by algae: Decarboxylation of stearic acid to N-heptadecane in *Anacystis nidulans* determined by ^{13}C-and ^{2}H-labeling and ^{13}C nuclear magnetic resonance. *Lipids* **15**, 609–615 (1980).
30. Dembitsky, V. & Srebnik, M. Variability of hydrocarbon and fatty acid components in cultures of the filamentous cyanobacterium *Scytonema* sp. isolated from microbial community "black cover" of limestone walls in Jerusalem. *Biochemistry (Moscow)* **67**, 1276–1282 (2002).
31. Shiea, J., Brassell, S. C. & Ward, D. M. Mid-chain branched mono-and dimethyl alkanes in hot spring cyanobacterial mats: A direct biogenic source for branched alkanes in ancient sediments? *Org. Geochem.* **15**, 223–231 (1990).
32. Giovannoni, S. J. & Stingl, U. Molecular diversity and ecology of microbial plankton. *Nature* **437**, 343–348 (2005).
33. Venter, J. C. *et al*. Environmental genome shotgun sequencing of the Sargasso Sea. *science* **304**, 66–74 (2004).
34. Schulz, S. *et al*. Plant litter and soil type drive abundance, activity and community structure of *alkB* harbouring microbes in different soil compartments. *ISME J.* **6**, 1763–1774 (2012).
35. Powell, S., Bowman, J., Ferguson, S. & Snape, I. The importance of soil characteristics to the structure of alkane-degrading bacterial communities on sub-Antarctic Macquarie Island. *Soil Biol. Biochem.* **42**, 2012–2021 (2010).
36. Shade, A. *et al*. Culturing captures members of the soil rare biosphere. *Environ. Microbiol.* **14**, 2247–2252 (2012).
37. Yooseph, S. *et al*. Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* **468**, 60–66 (2010).
38. Kato, T., Haruki, M., Imanaka, T., Morikawa, M. & Kanaya, S. Isolation and characterization of long-chain-alkane degrading *Bacillus thermoleovorans* from deep subterranean petroleum reservoirs. *J. Biosci. Bioeng.* **91**, 64–70 (2001).
39. Wang, L. *et al*. Isolation and characterization of a novel thermophilic *Bacillus* strain degrading long-chain *n*-alkanes. *Extremophiles* **10**, 347–356 (2006).
40. Tourova, T. *et al*. *alkB* homologs in thermophilic bacteria of the genus *Geobacillus*. *Mol. Biol.* **42**, 217–226 (2008).
41. Moore, R. C. & Purugganan, M. D. The early stages of duplicate gene evolution. *Proc. Natl Acad. Sci. USA* **100**, 15682–15687 (2003).
42. Kondrashov, F. A., Rogozin, I. B., Wolf, Y. I. & Koonin, E. V. Selection in the evolution of gene duplications. *Genome Biol.* **3**, 8.1–8.9 (2002).
43. Zhang, J. Evolution by gene duplication: an update. *Trends Ecol. Evol.* **18**, 292–298 (2003).
44. Jones, C. D. & Begun, D. J. Parallel evolution of chimeric fusion genes. *Proc. Natl Acad. Sci. USA* **102**, 11373–11378 (2005).
45. Gogarten, J. P. & Townsend, J. P. Horizontal gene transfer, genome innovation and evolution. *Nature Rev. Microbiol.* **3**, 679–687 (2005).
46. Pasek, S., Risler, J.-L. & Brézellec, P. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics* **22**, 1418–1423 (2006).
47. Daubin, V., Gouy, M. & Perriere, G. A phylogenomic approach to bacterial phylogeny: evidence of a core of genes sharing a common history. *Genome Res.* **12**, 1080–1090 (2002).
48. Gogarten, J. P. & Olendzenski, L. Orthologs, paralogs and genome comparisons. *Curr. Opin. Genet. Dev.* **9**, 630–636 (1999).
49. Syvanen, M. Horizontal gene transfer: evidence and possible consequences. *Annu. Rev. Genet.* **28**, 237–261 (1994).
50. Garcia-Vallvé, S., Romeu, A. & Palau, J. Horizontal gene transfer in bacterial and archaeal complete genomes. *Genome Res.* **10**, 1719–1725 (2000).
51. Tan, H.-M. Bacterial catabolic transposons. *Appl. Microbiol. Biotechnol.* **51**, 1–12 (1999).
52. Kok, M. *et al*. The *Pseudomonas oleovorans* alkane hydroxylase gene. Sequence and expression. *J. Biol. Chem.* **264**, 5435–5441 (1989).
53. van Beilen, J. B. *et al*. Analysis of *Pseudomonas putida* alkane-degradation gene clusters and flanking insertion sequences: evolution and regulation of the *alk* genes. *Microbiology* **147**, 1621–1630 (2001).
54. van Beilen, J. B. *et al*. Characterization of two alkane hydroxylase genes from the marine hydrocarbonoclastic bacterium *Alcanivorax borkumensis*. *Environ. Microbiol.* **6**, 264–273 (2004).
55. Giebler, J., Wick, L. Y., Schloter, M., Harms, H. & Chatzinotas, A. Evaluating the assignment of *alkB* aerminal restriction fragments and sequence types to distinct bacterial taxa. *Appl. Environ. Microbiol.* **79**, 3129–3132 (2013).
56. Cai, M. *et al*. Complete genome sequence of *Amycolicicoccus subflavus* DQS3-9A1^T, an actinomycete isolated from crude oil-polluted soil. *J. Bacteriol.* **193**, 4538–4539 (2011).
57. Patthy, L. *Origin and Evolution of New Gene Functions* (Springer Netherlands, 2003).
58. Nodate, M., Kubota, M. & Misawa, N. Functional expression system for cytochrome P450 genes using the reductase domain of self-sufficient P450RhF from *Rhodococcus* sp. NCIMB 9784. *Appl. Microbiol. Biotechnol.* **71**, 455–462 (2006).
59. De Mot, R. & Parret, A. A novel class of self-sufficient cytochrome P450 monooxygenases in prokaryotes. *Trends Microbiol.* **10**, 502–508 (2002).
60. Munro, A. W., Girvan, H. M. & McLean, K. J. Cytochrome P450-redox partner fusion enzymes. *BBA-Gen. Subjects* **1770**, 345–359 (2007).
61. Peterson, J. A., Basu, D. & Coon, M. J. Enzymatic ω-oxidation. *J. Biol. Chem.* **241**, 5162–5164 (1966).
62. Edgar, R. C. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.* **32**, 1792–1797 (2004).
63. Eddy, S. R. Profile hidden Markov models. *Bioinformatics* **14**, 755–763 (1998).
64. Eddy, S. R. Accelerated profile HMM searches. *PLoS Comput. Biol.* **7**, e1002195 (2011).
65. Shanklin, J., Whittle, E. & Fox, B. G. Eight histidine residues are catalytically essential in a membrane-associated iron enzyme, stearoyl-CoA desaturase, and are conserved in alkane hydroxylase and xylene monooxygenase. *Biochemistry* **33**, 12787–12794 (1994).
66. Nelson, D. R. *et al*. P450 superfamily: update on new sequences, gene mapping, accession numbers and nomenclature. *Pharmacogenetics* **6**, 1–42 (1996).
67. Letunic, I., Doerks, T. & Bork, P. SMART 7: recent updates to the protein domain annotation resource. *Nucleic Acids Res.* **40**, D302–D305 (2012).
68. Cole, J. *et al*. The Ribosomal Database Project: improved alignments and new tools for rRNA analysis. *Nucleic Acids Res.* **37**, D141–D145 (2009).
69. Quast, C. *et al*. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res.* **41**, D590–D596 (2013).
70. Ludwig, W. *et al*. ARB: a software environment for sequence data. *Nucleic Acids Res.* **32**, 1363–1371 (2004).
71. Saitou, N. & Nei, M. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Mol. Biol. Evol.* **4**, 406–425 (1987).

72. Langille, M. G. I. & Brinkman, F. S. L. IslandViewer: an integrated interface for computational identification and visualization of genomic islands. *Bioinformatics* **25**, 664–665 (2009).
73. Markowitz, V. M. *et al*. IMG/M: a data management and analysis system for metagenomes. *Nucleic Acids Res.* **36**, D534–D538 (2008).
74. Huson, D. H., Mitra, S., Ruscheweyh, H. J., Weber, N. & Schuster, S. C. Integrative analysis of environmental sequences using MEGAN4. *Genome Res.* **21**, 1552–1560 (2011).

Acknowledgments

This study was supported by the National Natural Science Foundation of China (31225001, 31300108) and the National High Technology Research and Development Program ("863"Program: 2012AA02A703, 2014AA021505).

Author contributions

Y.N. and X.L.W. designed the work. Y.N., H.F. and C.C.Q. performed the data analysis. Y.N. and X.L.W. wrote the main manuscript. J.L.L., S.L.L., G.L.L. and Y.Q.T. performed the data collection. All authors discussed and reviewed the manuscript.

Additional information

Supplementary information accompanies this paper at http://www.nature.com/scientificreports

Competing financial interests: The authors declare no competing financial interests.

How to cite this article: Nie, Y. *et al*. Diverse alkane hydroxylase genes in microorganisms and environments. *Sci. Rep.* **4**, 4968; DOI:10.1038/srep04968 (2014).