**Functional soil metagenomics: elucidation of polycyclic aromatic
hydrocarbon degradation potential following 12 years of *in situ*
bioremediation**

Márcia Duarte¹, Agnes Nielsen¹, Amélia Camarinha-Silva^{1,3}, Ramiro Vilchez-Vargas^{1,4},
Thomas Bruls², Melissa L. Vos-Oxley¹, Ruy Jauregui^{1,5} and Dietmar H. Pieper^{1*}

¹Microbial Interactions and Processes Research Group, HZI – Helmholtz Centre for Infection
Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany.

²Institut de Génomique, Genoscope, UMR8030 (CNRS, CEA, Université d'Evry), Evry,
France.

*For correspondence. E-mail Dietmar.Pieper@helmholtz-hzi.de; Tel. (+49) 531 6181 4200.

Running title: Functional soil metagenomics

³Present affiliation: Institute of Animal Nutrition, University of Hohenheim Emil-Wolf-Str. 10,
70599 Stuttgart, Germany.

⁴Department of Gastroenterology, Hepatology and Infectious Diseases, Otto-von-Guericke-
University, Magdeburg, Germany.

⁵AgResearch Grasslands, Tennent drive, Palmerston North, New Zealand

Originality-Significance Statement

In the current work, we explored by function based screening the biodegradation potential of a soil subjected to extensive bioremediation. We extracted from the metagenome >200 fosmid inserts expressing key catabolic activities and obtained by sequencing not only information of those key aromatic ring cleavage activities, but extensive information on the complete net for aromatic degradation. As current databases are overloaded with mis-annotations, we used a curated database for aromatic degradation reactions for careful annotations, and extended the analysis to further aromatic degradation reactions. The fact that most of the encoded enzymes found in this study were only distantly related to those previously reported suggests that the diversity of the key catabolic genes is much greater than expected and has significantly increased our knowledge of the functional degradative potential of an *in situ* microbial community.

Summary

A culture-independent function-based screening approach was used to assess the microbial aerobic catabolome for polycyclic aromatic hydrocarbons degradation of a soil subjected to 12 years of *in situ* bioremediation. A total of 422,750 fosmid clones were screened for key aromatic ring-cleavage activities using 2,3-dihydroxybiphenyl as substrate. Most of the genes encoding ring-cleavage enzymes on the 768 retrieved positive fosmids could not be identified using primer-based approaches and, thus, 205 fosmid inserts were sequenced. Over two hundred extradiol dioxygenase encoding genes of three different superfamilies could be identified, where approx. 60% of the encoded enzymes shared only <70% of protein sequence identity to those previously described. Additional key genes of aromatic metabolic pathways were identified, including a high abundance of Rieske non-heme iron oxygenases that provided detailed information on enzymes activating aromatic compounds and enzymes involved in activation of the side chain of methylsubstituted aromatics. The gained insights indicated a complex metabolic network acting at the site under study and underlined the great potential of an approach that combines an activity-screening, a pooling strategy for cost-effective high-throughput sequencing of fosmid clones and a phylogenomic-routed and manually curated database to carefully identify key proteins dedicated to aerobic degradation of aromatic compounds.

Introduction

Although several aromatic compound-degrading bacteria have been isolated and studied under laboratory conditions (Peng *et al.*, 2008), the majority of environmental bacteria cannot be readily cultivated using standard techniques and thus, remains unexplored (Eaton, 1994; Amann *et al.*, 1995). This vast and yet unstudied microbial diversity represents a huge natural resource for the isolation of new genes, enzymes and metabolic pathways (Culligan *et al.*,

2014; Liebl *et al.*, 2014) and needs to be understood to assess the aromatic degradation potential in contaminated environments and the metabolic web of biodegradation reactions. Within the past years, culture-independent surveys of catabolic gene diversity in contaminated environments have been reported, involving PCR-based screenings (Witzig *et al.*, 2006; Ding *et al.*, 2010) or DNA hybridization (Vilchez-Vargas *et al.*, 2013). However, these sequence-based strategies rely on conserved nucleotide sequences used for the design of primers or probes and fail to identify genes encoding novel enzymes unrelated or only distantly related to known enzymes (Steele *et al.*, 2009; Liebl *et al.*, 2014).

A metagenomic approach can overcome the aforementioned drawbacks through non-selective cloning of sequences from a given environment (Culligan *et al.*, 2014). Cost-effective and highly scalable functional screenings of metagenomic libraries led to the identification of genes with an economical or biomedical value (Lam *et al.*, 2015). The success of this approach is, however, highly dependent on DNA isolation, cloning and expression procedures (Salonen *et al.*, 2010; Delmont *et al.*, 2011; Liebl *et al.*, 2014).

Despite the advantages, functional metagenomics dedicated to the analysis of aromatic degradation has only been applied in few studies, whereas metagenomic surveys, which can only cover organisms and genes present in high abundance in a given environment, are often attempted (Guazzaroni *et al.*, 2013). The use of functional screening to retrieve aromatic catabolic activities has so far, to our knowledge, been restricted to the study of extradiol dioxygenases with high activity against catechol and thus, to the degradation of simple aromatic compounds such as benzene/toluene (Brennerova *et al.*, 2009) and phenols (Suenaga *et al.*, 2007) or to petroleum contaminated sites not stressed with a high level of polycyclic aromatic hydrocarbon (PAH) contamination (Terron-Gonzalez *et al.*, 2016). These studies indicated screens for extradiol dioxygenases as promising tools to identify the diversity of such enzymes in contaminated environments.

Thus, we reasoned that 2,3-dihydroxybiphenyl would be a useful screening substrate to gain an insight into aromatic ring-cleavage enzymes involved in PAH degradation that are selected in contaminated sites. Cloning into fosmid vectors combined with next generation sequencing should then enable not only identification of extradiol dioxygenase encoding genes but also of genes encoding peripheral enzymes activating aromatics. In this study, we used soil from a formerly PAH-contaminated site that has been subjected to 12 years of successful *in situ* bioremediation. Hence, the use of this soil sample as source of genomic DNA for a functional metagenomic approach represents an extraordinary opportunity to analyze the *in situ* catabolic diversity for aromatic degradation. Using AromaDeg, a web-based resource targeting aerobic degradation of aromatics that comprises updated and manually curated databases grounded in phylogenetic analyses of protein sequences of key catabolic protein families (Duarte *et al.*, 2014), we were capable to obtain a detailed overview of the catabolic potential of the site under study.

Results and discussion

Bacterial community analysis as bias indicator for DNA extraction methodologies

The DNA extraction protocol used to prepare large-insert clone libraries comprised an enzymatic lysis to obtain sufficiently intact metagenomic DNA. The efficiency of lysis and the representativeness of the isolated DNA was evaluated by comparison of amplicon libraries comprising the V5/V6 16S rRNA gene region, prepared using either high-molecular weight metagenomic DNA or genomic DNA extracted using a highly efficient mechanical lysis. Both amplicon libraries contained the same 513 phylotypes dominated by *Proteobacteria* (See Fig. S1 and Dataset S1). As expected, differences were observed in the relative abundance of some bacterial phyla, and specifically actinobacterial sequence reads were of lower abundance in the library prepared by enzymatic lysis (Fig. S1). However, the fact that even reads originating

from *Mycobacteria* were observed in this library (Dataset S1), shows the efficiency of the selected enzymatic lysis and of the high-molecular weight DNA purification.

Screening of metagenomic libraries

The generated metagenomic library containing 422,750 fosmid clones (corresponding to approx. 17 Gb of sequence information or >3700 bacterial genome equivalents when assuming a mean genome size of 4.5 Mb) was subjected to activity-based screening for aromatic ring-cleavage, where 768 clones turned yellow when sprayed with 10 mM 2,3-dihydroxybiphenyl. This number of positive clones corresponds to one active clone per 22Mb of sequence screened. This amount of clones was higher than previously reported in activity screening studies using 2,3-dihydroxybiphenyl as substrate, where one clone per 155 Mb of screened DNA was observed (Terron-Gonzalez *et al.*, 2016) and similar to values observed for catechol 2,3-dioxygenase activities in contaminated environments (varying from one clone per 3.6Mb to 1 clone per 35 Mb of screened DNA (Suenaga *et al.*, 2007; Brennerova *et al.*, 2009)) and reflects not only a high abundance of ring cleavage activities in the analyzed soil but also high soil potential for the degradation of aromatic compounds.

The use of 9 primer sets targeting 5 main groups of genes encoding extradiol dioxygenases (EXDOs), including those reported to exhibit high activity with 2,3-dihydroxybiphenyl (Table S1), yielded specific PCR products only for 24 of the 768 positive clones (Table S2). Twenty-two of these were obtained using EXDO-9 primers targeting proteobacterial catechol 2,3-dioxygenases. Only one specific amplification product each was obtained, using EXDO-3 and EXDO-4 primers targeting dihydroxynaphthalene dioxygenases and related enzymes in *Proteobacteria*. Of the previous reports on the function-based screening of extradiol dioxygenases, only one applied a primer-based approach and could classify 72% of extradiol dioxygenase encoding genes (Brennerova *et al.*, 2009), whereas others (Suenaga *et al.*, 2007;

Terron-Gonzalez *et al.*, 2016) claimed the majority of genes identified to be only distantly related to previously observed ones. Clearly, the diversity of extradiol dioxygenases is too broad to be identified by approaches based on available sequence information (Vilchez-Vargas *et al.*, 2010).

Gauging a pooling strategy for high-throughput sequencing of metagenomic clones

As the majority of EXDOs comprised in the library could not be identified by PCR-screening, we aimed to identify those through high-throughput sequencing. To avoid the preparation of costly libraries per single fosmid, the sequencing of multiple fosmids pooled in single libraries was tested as an affordable sequencing strategy. Twenty fosmid clones (including those that yielded PCR products using EXDO-3, EXDO-4 and EXDO-9 primers and nine other randomly selected active clones) (Table S2) were pooled and sequenced.

Among these 20 sequenced fosmid inserts, six could be fully assembled. With the single exception of the 37,267-bp contig A73-109, which did not encode any aromatic degradation-related genes (in fact, a repeated activity-based screening indicated the activity to cleave 2,3-dihydroxybiphenyl to be lost from the fosmid insert), all the other five fully assembled fosmids encoded proteins related to aromatic degradation. A 45,669-bp-long contig (A179-233, Dataset S2) showed only three base pairs difference to the well-studied naphthalene degradation plasmid pDTG1 of *Pseudomonas putida* strain NCIB 9816-4, including its complete upper pathway (responsible for the conversion of naphthalene to salicylate), bordered by transposases, resolvases, and integrases. A 40,140-bp contig (A374-418, Dataset S2) comprised a ~15kb region containing aromatic degradation genes (A401-418, Dataset S2) with 83% identity to the recently sequenced genome of *Polycyclovorans algicola* strain TG408 (GenBank: NZ_JOMH000000000) (Gutierrez *et al.*, 2013), comprising 5 pairs of genes encoding α -subunits of Rieske non-heme iron oxygenases (RNIOs) and the respective β -

subunits devoid of genes encoding ferredoxins or ferredoxin reductases. A 38,086-bp contig (A34-72, Dataset S2) comprised only two genes probably involved in aromatic degradation (A61-60, Dataset S2), among them an extradiol dioxygenase, which shared 82% identity with 2,3-dihydroxybiphenyl dioxygenase BphC2 (GenBank: AAB58815) from *S. xenophagum* strain BN6 (Heiss *et al.*, 1995). A 38,486-bp contig (A3-33, Dataset S2) comprised four genes encoding enzymes probably involved in carbazole degradation confined in a 3-kb cluster (A22-A25, Dataset S2), with a *carAaCBaBb* gene organization as previously described for the marine carbazole degrading unclassified α -*Proteobacterium* OC9 (Maeda *et al.*, 2010). A 34,828-bp contig (A276-312, Dataset S2) comprised five genes encoding aromatic degradation (A312-308, Dataset S2) among them a LigB-type EXDO distantly related to 2,3-dihydroxyphenylpropionate dioxygenases of the protocatechuate family (Duarte *et al.*, 2014), sharing less than 55% sequence identity with its closest relatives (GenBank: WP_028222141). Additionally to these fully-assembled fosmids, various other contigs showed >99% sequence identity to a ~35 kb region of the genome of the BTEX-degrading bacterium *Pseudoxanthomonas spadix* BD-a59 comprising the gene encoding a putative catechol 2,3-dioxygenase (GenBank: YP_004930496, contigs A320 to A360-361). PCR screening revealed that it was present in 4 of the 20 fosmids (Table S2) and in 17 of the entire collection of 768 clones. In fact, the community structure analysis revealed a 16S rRNA gene sequence identical to that of *P. spadix* BD-a59 (Phylotype 14, Dataset S1) to be abundant (~0.7% of relative abundance) such that it may be assumed that the presence of organisms similar to *P. spadix* BD-a59 and its arsenal of degradation genes might be of high importance for this soil.

Smaller contigs (data not shown), corresponding to the remaining 10 fosmids, encoded twelve enzymes of the vicinal oxygen chelate superfamily of EXDOs, two EXDOs of the LigB superfamily and fifteen α -subunits of RNIOs from 4 different families. Hence, the analysis showed that the employed methodology for pooled sequencing was successful and a total of

200 randomly selected metagenomic clones (including 14 clones that had have yet to be fully assembled) were subjected to sequencing.

Overview of EXDO and RNIO encoding genes identified from 205 fosmids

Sequence analysis of 205 fosmid inserts (including the main dataset of 200 sequenced fosmids plus five fully-assembled fosmids described above) using the AromaDeg database (Duarte *et al.*, 2014) resulted in the identification of 202 ORFs encoding EXDOs. Phylogenomic analysis identified them to be affiliated to three distinct families of the vicinal oxygen chelate superfamily (termed EXDO-VC1, -VC2 and -VC3, see Table 1), to the LigB superfamily, and to the cupin superfamily (Table 1). Only 21 of them showed $\geq 98\%$ protein sequence identity to previously reported enzymes. In contrast, the majority (117 out of 202) had less than 70% protein sequence identity to previously described proteins and often formed novel lineages, suggesting a striking difference between the aromatic-degradation genes identified by functional-based screening and those previously characterized from isolates or identified by screening using primers targeting gene groups known from isolates. Of special importance, forty-one ORFs encoded EXDOs of the LigB superfamily. Such enzymes had not been observed in catechol based screenings which has been attributed to their scarcity in the environment (Suenaga *et al.*, 2007). However, such enzymes have been identified in metagenome based screenings recently (Terron-Gonzalez *et al.*, 2016), and the high number observed in the contaminated environment analyzed here suggest the high importance of these enzymes in biodegradation.

Sequencing of fosmid inserts also enabled the identification of genes encoding enzymes activating aromatic compounds. This allows to deduce putative substrates of the encoded metabolic routes providing information exceeding that obtained by using sequence based approached to detect solely extradiol dioxygenases. The AromaDeg database (Duarte *et al.*,

2014) revealed the presence of 187 ORFs encoding α -subunits of RNIOs that could be affiliated to four distinct families (Table 1) where 58 ORFs encoded α -subunits of the biphenyl family (a family that comprises most of the dioxygenases reported thus far as being involved in the activation of polycyclic aromatics), including 25 oxygenases of cluster XXIV (naphthalene or polycyclic aromatic dioxygenases of *Proteobacteria*) (Table 1, Fig. 1 and Fig. S2).

Nine other RNIO α -subunits belong to the phthalate oxygenase family (Nakatsu *et al.*, 1995), members of which are typically active without requiring β -subunits. This family also comprises carbazole dioxygenases and three genes encoding probable carbazole dioxygenases (Fig. S3, cluster I) were identified, however, all gene products were only distantly related to those previously described (<68% protein sequence identity).

Fifty-one ORFs encoded α -subunits of the salicylate family (Table 1 and Fig. S4). Salicylate hydroxylases transform salicylate into either catechol (salicylate 1-hydroxylases) or gentisate (salicylate 5-hydroxylases). From the identified 51 ORFs, only 2 could be affiliated with salicylate 5-hydroxylases (Fig. S4). The remaining either encoded enzymes clustering with salicylate 1-hydroxylases previously reported in Sphingomonads or *Cycloclasticus* spp. or constitute new branches closely related to probable salicylate 1-hydroxylases or anthranilate dioxygenases (Fig. 2 and Fig. S4). Only two ORFs were identified (C868 and F214) (Fig. 3A) that probably encode flavoprotein salicylate 1-hydroxylases (You *et al.*, 1991; Bosch *et al.*, 1999; Pérez-Pantoja *et al.*, 2010). Rieske non-heme iron salicylate 1-hydroxylases have been poorly studied so far (Pinyakong *et al.*, 2003; Jouanneau *et al.*, 2007) but their high abundance and diversity in the metagenomic library indicates their importance for aromatic degradation in the soil under study.

Sixty-nine ORFs encoding α -subunits of RNIOs were affiliated to the benzoate family. Again, most of the encoded enzymes form deep branching clusters in the phylogeny of this family and several of those clustered together in a new branch distantly related to benzoate

dioxygenases (sharing 42 -52% of protein sequence identity with BAE46757, AB62286 or S23477 benzoate dioxygenases of *Burkholderia cepacia*, *Rhodococcus jostii* or *Acinetobacter calcoaceticus*, respectively) (Fig. S5). If they have a function in transformation of the structural analogous salicylate remains to be elucidated.

Gene clusters encoding toluene monooxygenases

Seven fosmids that showed a PCR amplification product with at least two of the EXDO-12, EXDO-13 and TolM_EXDO primer sets (targeting a toluene monooxygenase and three EXDOs) were sequenced and three of the resulting contigs (B91-123, G206-245 and J133-146) showed 99.9% of DNA sequence identity with the genome of the recently isolated *Rhodocyclaceae* bacterium PG1-Ca6. A further four (C445-480, E138-164, H19-57 and K675-704) showed >82% DNA sequence identity with this genome. All seven fosmid inserts showed a similar gene organization, however, genes encoding a toluene monooxygenase were integrated in a probable polycyclic aromatic catabolic gene cluster of only four of these inserts (B91-123, G206-245, H19-57 and J133-146) (Fig. 3B). They were localized downstream of genes encoding an EXDO with high similarity with DbtC of the dibenzothiophene degrader *Burkholderia* sp. DBT1 (Di Gregorio *et al.*, 2004) (cluster XII of EXDO-VC2, Fig. 4 and Fig. S6). Three additional EXDO encoding genes were typically observed upstream of this gene (clusters XVIII, XVII and XX of EXDO-VC1, Fig. S7). Enzymes related to EXDO-VC1 clusters XVIII and XVII were surprisingly abundant in the fosmid library but these clusters do not comprise enzymes of proven function. All of these fosmids (with the exception of J133-146) comprised a putative xylene monooxygenase encoding gene (Fig. 3B).

In the four fosmid inserts that encoded a toluene monooxygenase, the respective genes are followed by those encoding the α - and β -subunits of a putative biphenyl dioxygenase (cluster XXVI, Fig. S2) and a salicylate 1-hydroxylase (Fig. S4, cluster I and Dataset S2, A374-418). A

conserved structure further downstream indicated the presence of various genes encoding enzymes of aromatic degradation including an EXDO of the protocatechuate family of LigB type enzymes sharing 57% protein sequence identity with PhnC from the phenanthrene-degrading *Burkholderia* sp. strain RP007) (Laurie and Lloyd-Jones, 1999) (Fig. 5 and Fig. S8, cluster XII) and the α - and β -subunits of an additional biphenyl family RNIO related to those involved in the degradation of phenanthrene (62-63% of sequence identity to PhnAc of *Acidovorax* sp. NA3) (Singleton *et al.*, 2009) (Fig. S2, cluster XXIV). Despite the overall relatively conserved gene structure, 4 slightly different variants of the gene cluster were observed.

Catabolic clusters encoding 2,3-dihydroxy-p-cumate dioxygenases

Primers designed based on genes encoding enzymes related to 2,3-dihydroxy-*p*-cumate dioxygenases (Table S1) were able to identify 37 fosmid inserts as carrying such genes (Table S2, Fig. 6). Seven of these were sequenced and encoded seven putative 2,3-dihydroxy-*p*-cumate dioxygenases (B3, C763, E250, F164, G204, H335 and L588) (Fig. S9, cluster I). Two further fosmids encoded putative 2,3-dihydroxy-*p*-cumate dioxygenases (B356 and K318), where the encoding genes exhibited 1-2 mismatches with the EXDO-10 primer.

All nine proteins shared only 52-57% of sequence identity with 2,3-dihydroxy-*p*-cumate dioxygenase from *P. putida* F1 (Eaton, 1996) and were observed in a conserved gene cluster organization (Fig. 3C) where the EXDO (CmtC) encoding genes are preceded by genes encoding the α - and β -subunits of a *p*-cumate 2,3-dioxygenase (Fig. S5, cluster VIII) and followed by a putative *cis*-2,3-dihydroxy-2,3-dihydro-*p*-cumate dehydrogenase (CmtB) encoding gene (Eaton, 1996) (Fig. 6).

A further three fosmids encoded EXDOs distantly related those above (F68, D518 and L545), but only F29-70 comprised a gene cluster encoding all 4 subunits of *p*-cumate 2,3-

dioxygenase supposedly essential for activity (Eaton, 1996). Although the p-cumate degradative pathway has received limited attention in the past, the high number and diversity of genes encoding enzymes of this pathway suggests that it is harbored by several bacteria within the soil under study (an average of one p-cumate pathway per 300 Mb of screened DNA), which indicates its importance in a complex environmental sample under *in situ* conditions. As enzymes of this pathway have been reported to be of a broad substrate specificity (Eaton, 1996), they may also have a function in the metabolism of methylsubstituted or carboxylated PAH intermediates.

Catabolic gene clusters encoding LigB type EXDOs

Another abundant catabolic gene organization observed on five fosmids (Figs 3D1 and 3D2), comprises two genes encoding the β -subunits of LigB type EXDOs (Fig. S8, cluster XII), and an overall of 18 LigB proteins sharing >70% of protein sequence identity were observed on different contigs. All these LigB enzymes shared >53% of sequence identity with PhnC of *Burkholderia sartisoli* RP007 (AAD09870) (Fig. 4 and Fig. S8, cluster XII), which was shown to be involved in naphthalene and phenanthrene degradation (Laurie and Lloyd-Jones, 1999), indicating that LigB encoding genes observed here are also crucial for the degradation of PAHs. The presence of two slightly distinct variants of the same enzyme family may be indicative for evolving gene clusters and evolving substrate specificities of the encoded proteins.

Genes encoding the α -subunits of putative salicylate 1-hydroxylases (Fig. S4, cluster XIV) (exemplified by D302 and F147 – Figs 3D1 and 3D2, respectively) and genes encoding the α - and β -subunits of a PAH dioxygenase (exemplified by D301-300 and F148-149 – Figs 3D1 and 3D2, respectively) where the α - subunits share >60% sequence identity with PhnA1 (BAC81541) (Fig. S2, cluster XXIV) of *Cycloclasticus* sp. A5 (Kasai *et al.*, 2003) are localized

between the LigB encoding genes. Collectively, the gene cluster encodes enzymes crucial for PAH-degradation.

The complex gene clusters also encode proteins involved in the degradation side-chain oxidation of methyl-substituted aromatics, namely the two-component xylene/p-cymene monooxygenases (Fig. 6), which consist of a hydroxylase related to AlkB alkane hydroxylase and a reductase, (Worsey and Williams, 1975; Eaton, 1996). A reductase encoding gene (e.g. F416, 46% of sequence identity with *p*-cymene monooxygenase reductase CymAb of *Pseudomonas chlororaphis* (Dutta and Gunsalus, 1997) is encoded directly downstream of a benzoate family β -subunit encoding gene (e.g. F417) and a total of eighteen CymAb encoding genes could be identified. The respective hydroxylase encoding genes were encoded roughly 1.6 kb downstream in the opposite direction. Phylogenetic analysis clearly placed the encoded enzymes in a cluster of enzymes capable to hydroxylate the methyl-substituent of methyl-substituted aromatics (Fig. S10), which evidently seems to be widespread in the environment analyzed here.

In addition, this gene region contained a pathway capable of hydroxylation of the aromatic ring. Directly upstream of the gene cluster encoding LigB type EXDOs, two subunits of a two-component *p*-cresol methylhydroxylase, an enzyme catalyzing the hydroxylation of *p*-cresol to *p*-hydroxybenzyl alcohol and further dehydrogenation to *p*-hydroxybenzaldehyde (Peters *et al.*, 2007) are encoded (Figs 6 and 3D) and the flavoprotein (e.g. F143) shares significant similarity (44% of identity) with validated *p*-cresol methylhydroxylase such as P09788 of *P. putida* NCIMB 9866 (Cronin *et al.*, 1999) (Fig. S11). Further transformation of *p*-hydroxybenzaldehyde may be catalyzed by the aldehyde dehydrogenases encoded by the subsequent gene, which exhibits 41% sequence identity with vanillin dehydrogenase from *Pseudomonas* sp. HR199 (Overhage *et al.*, 1999). The following gene (e.g. F140, Fig. 3D2) encodes a protein of the 3-octaprenyl-4-hydroxybenzoate carboxy-lyase/UbiD superfamily

sharing 54% of sequence identity with 4-hydroxybenzoate/vanillate decarboxylase of *Streptomyces* sp. D7 (Chow *et al.*, 1999; Lupa *et al.*, 2005). It could be thus speculated that the function of this gene cluster is for formation of phenol from cresol (Fig. 6). Further downstream, two α -subunits of benzoate dioxygenases (e.g. F137 and F136, cluster XI, Fig. S5) are encoded that are followed by genes encoding two subunits of a second p-cresol methylhydroxylase (e.g. F135-134, Fig. 3D2 and Fig. S11). This novel gene organization has never been described and although the associated fosmid inserts share a common gene organization, the genes are not identical and not due to clonal expansion.

A catabolic gene cluster encoding EXDOs and p-cresol methylhydroxylase

The degradation of aromatics via hydroxylation of the methyl substituent seems of particular importance in this soil, as a further set of catabolic gene clusters encoding two subunits of a p-cresol methylhydroxylase was observed in five other fosmid inserts (e.g. B248-280, Fig. 3E) where the encoded subunits (e.g. B261-260) shared significant similarity with p-cresol methylhydroxylase of *P. putida* (e.g. 39-41% identity with flavoprotein AAA80318 and 30-33% of identity with cytochrome AAA80319). As described above, these genes are followed by those encoding an aldehyde dehydrogenase and a 3-octaprenyl-4-hydroxybenzoate carboxylase/UbiD superfamily protein. Such a gene organization seems thus to be widespread in the soil under study, however, its actual function in aromatic degradation remains to be elucidated.

Approximately 11kb upstream, two probable catechol 2,3-dioxygenases affiliated to clusters XVII (e.g. B276) and XVIII (e.g. B275) of the EXDO-VC1 family (Fig. S7) were encoded in all five fosmids (e.g. B276-275, Fig. 3E). As these are the only EXDOs encoded on these probably completely assembled fosmid inserts, it can be assumed that the p-cresol methylhydroxylase gene cluster was just fortuitously isolated. The presence of adjacent genes encoding two catechol 2,3-dioxygenases affiliated to clusters XVII/XVIII (Fig. S7) seems to be

of high importance as it was not exclusive to these five fosmids and 16 other fosmid inserts comprised two subsequent genes encoding such catechol 2,3-dioxygenases.

Catabolic gene cluster encoding members of the gentisate dioxygenase family

Only two fosmids encoded proteins of the gentisate dioxygenase family (Fig. S12), which however were not targeted by the screening approach. F460-496 encodes a quite diverse arsenal of aromatic degradation enzymes, including EXDOs of all three superfamilies (Fig. 3F) with a unique gene organization. Approximately half of this 36-kb fosmid insert (F460-475) shared high DNA sequence identity with the previously reported phenanthrene catabolic gene regions of *Delftia* sp. Cs1-4 (CP002735, DelCs14_1744-1765), and *Alcaligenes faecalis* AFK2 (*phnAd* to *phnF* genes, AB024945).

Gene organizations comprising family EXDO-VC enzymes

Among the EXDOs, EXDO-VC1 enzymes of the subfamilies previously termed I.2.A-C (Eltis and Bolin, 1996) (cluster XX and XXII, Fig. S7) have received special attention, and specifically enzymes of subfamily I.2.C (cluster XX) were of high abundance (22 of 768 fosmid clones). Fifteen genes encoding cluster XXII enzymes (I.2 A and B) (Eltis and Bolin, 1996) were identified among the 205 sequenced fosmid clones, with three of them (G506, B368 and L614) closely related to previously characterized enzymes from Sphingomonads (e.g. 90-94% protein sequence identity with BAB07893) of *S. paucimobilis* - (Lu *et al.*, 2000) and two others (C871 and F212) closely related to catechol 2,3-dioxygenases of *Pseudomonas* spp. (e.g. 99% protein sequence identity with the catechol 2,3-dioxygenase encoded on the plasmid pDTG1 from *P. putida* strain NCIB 9816-4 - NP_863103) (Dennis and Zylstra, 2004), where both inserts encode a complete salicylate degradation pathway comprising a flavoprotein salicylate hydroxylase. However, a total of eight fosmid inserts contained genes only distantly

related to those identified from cultured organisms (B508, D571, F120, F453, G488, G521, H293 and L578) but closely related (92-94% of protein sequence identity) to enzymes encoded by a set of fosmid inserts recently identified by function-based screening using catechol as substrate (e.g. AB266122) (Suenaga *et al.*, 2007), indicating these genes to be of importance in different contaminated sites.

Among 20 EXDOs affiliated to cluster XII of family EXDO-VC2 (involved in naphthalene degradation and thus expected to be enriched in communities capable to degrade PAHs) (Fig. 4, Fig. S6), only seven were localized in a complex genetic environment as described above. Two cluster-XII EXDOs (L289 and K88) were encoded in a threonine synthesis gene cluster preceded by an aspartate aminotransferase and a homoserine dehydrogenase and followed by a threonine synthase encoding gene. Such an organization of a cluster-XII EXDO has also recently been reported from studies of a fosmid library constructed from a jet fuel contaminated site (Brennerova *et al.*, 2009). This indicates a respective gene organization to be of advantage for organisms in a contaminated site, even though a link between amino acid synthesis and aromatic degradation is not obvious. Five further cluster-XII EXDOs (J384, L793, F333, D207 and H103) were encoded in fosmid inserts where a gene encoding a putative diacylglycerol kinase followed the EXDO encoding gene and the abundance of this gene cluster may indicate a possible function in the environment under study.

Nine fasmids encoded a 2,3-dihydroxybiphenyl 1,2-dioxygenase affiliated to cluster XXII (see Fig. S3) where four (H463, H226, G447, D510) are localized in fosmid inserts that comprise a cluster (see Fig. 3G) similar to the archetype *bph* biphenyl catabolic gene cluster of *Burkholderia xenovorans* LB400 (Hofer *et al.*, 1993) (Fig. 7). Three further 2,3-dihydroxybiphenyl-1,2-dioxygenases (D66, F236 and J251, Fig. S6) share highest similarity (86-87% protein sequence identity) with PG1C_07515 of *Rhodocyclaceae bacterium* PG1-Ca6. Interestingly, various EXDO-VC2 form novel phylogenetic clusters (see clusters highlighted in

red, Fig. S6). For instance H189, L89, G308, J293 and K114 share only 43-51% protein sequence identity with putative biphenyl 2,3-dioxygenases encoded in the genome of *Bradyrhizobiaceae* (e.g. WP_018320654 of *Bradyrhizobium* sp. WSM2793, see cluster XI, Fig. S6). In these five fosmid inserts, the EXDO encoding genes are preceded by genes encoding enzymes exhibiting approximately 30% of sequence identity with 2-hydroxymuconic semialdehyde hydrolases such as DmpD of *Pseudomonas* sp. CF600 (CAA36993). Hence, it seems likely that the substrates for ring cleavage are *meta*-substituted catechols, which necessitate hydrolysis of the ring-cleavage product (Harayama *et al.*, 1987).

Twenty-four enzymes of family EXDO-VC3 could be identified, comprising twelve 2,3-dihydroxy-*p*-cumate 3,4-dioxygenases described above. Six fosmid inserts comprised genes encoding enzymes that are affiliated to cluster XIV, which includes 2,3-dihydroxybiphenyl dioxygenase BphC2 of *S. xenophagum* BN6 (Fig. S9) (Heiss *et al.*, 1997) and thus are assumed to function also as 2,3-dihydroxybiphenyl dioxygenases. Among these, A60 and G28 encode EXDOs of 81-84% protein sequence identity, with BphC2 of *S. xenophagum* BN6 where besides this enzyme, only a fumarylacetoacetate hydrolase family protein could be of evident function for aromatic degradation in the fosmid inserts.

One-domain EXDOs

EXDOs of the vicinal chelate superfamily are typically composed of homologous N- and C-terminal domains (Eltis and Bolin, 1996). Only a few so-called one-domain EXDOs have been described thus far, comprising BphC2 and BphC3 2,3-dihydroxybiphenyl dioxygenases from *Rhodococcus globerulus* P6 and BphC from *S. xenophagum* BN6 (Asturias and Timmis, 1993; Heiss *et al.*, 1995). A total of ten gene products showed similarity to the one-domain EXDOs, where E435 exhibited 32% of sequence identity with BphC from *S. xenophagum* BN6 and all other nine gene products (encoded by B435, D51, D434, D617, E88, J348, K282, L305 and

L689) exhibited 37-48% of sequence identity to BphC2 and BphC3 from *R. globerulus* P6. The natural function of these one-domain enzymes is still unknown (McKay *et al.*, 2003).

Only in case of L689, the genetic context (Fig. 3H) allows to deduce a possible function as the EXDO is framed by genes probably involved in the formation of ring-cleavage substrates. These are genes encoding the β - and α -subunits of a RNIO, where the α -subunit is closely related to biphenyl dioxygenases of clusters XXXI and XXXII (Fig. S2) and a gene encoding a protein of the 4-hydroxyphenylacetate 3-hydroxylase family (Pérez-Pantoja *et al.*, 2010), which shares 33% sequence identity with phenol hydroxylase from *Geobacillus thermoglucosidasius* A7 (Duffner *et al.*, 2000).

Conclusions

The activity-based screening employed here allowed the detection of a surprisingly high number of DNA fragments encoding aromatic ring-cleavage activities corresponding to 1 clone per 22 Mb screened DNA. The sequence analysis of 205 fosmid inserts gave novel insights into the distribution and organization of aromatics degradation genes across the microbial community of an *in situ* remediated soil. Strikingly, gene organizations observed in well-studied bacterial isolates were extremely rare. Besides 202 genes that encoded enzymes capable of 2,3-dihydroxybiphenyl cleavage (affiliated to EXDO-VC and LigB superfamilies), this study also identified 189 genes encoding RNIOs that could be affiliated to four families clearly indicating the wide range of potential substrates that can be targeted by the microbial community of the environment of interest here. The huge set of RNIOs related to naphthalene dioxygenases of *Proteobacteria* and of EXDOs related to dihydroxynaphthalene dioxygenases suggests unidentified *Proteobacteria* to be predominantly involved in PAH-degradation and the high abundance of Rieske non-heme iron salicylate 1-hydroxylases indicate those as previously unexplored key enzymes of PAH metabolism. Furthermore, various genes encoding enzymes

putatively involved in p-cumate metabolism and those encoding cresol methylhydroxylases or xylene/cymene monooxygenases indicate side-chain processing reactions to be of high importance in the soil under study.

The fact that most of the encoded enzymes found in this study were only distantly related to those previously reported suggests that the diversity of the key catabolic genes is much greater than expected and has significantly increased our knowledge of the functional degradative potential of an *in situ* microbial community.

Experimental procedures

Soil sample

The soil sample originates from a wood impregnation facility area - JDZ Soběslav, Czech Republic where 360,000 tons of soil highly contaminated with PAHs (with an average PAH concentration of 11,000 mg.kg⁻¹), phenol, heavy metals and oil hydrocarbons were remediated *in situ* over a period of 12 years via soil excavation (to promote oxygenation) and off-site biodegradation. Sampling was performed by borrowed pit at a depth of 1.8 m, where pollution by total PAHs remained at 157 mg kg⁻¹ of dry soil (as detailed in Table S3).

DNA extraction and library preparations

For bacterial community analysis DNA was extracted from quadruplicates of 0.5 g of homogenized and sieved soil using the FastDNA Spin Kit for Soil (MP Biomedicals, Solon, OH, USA) following the manufacturer's instructions.

High-molecular-weight soil DNA was recovered from four independent extractions of homogenized and sieved soil. For each extraction, 5 g of soil were suspended in 5 ml of suspension buffer (Vilchez *et al.*, 2007) and incubated at 37°C for 10 min. After incubation, 7.5 mL lysis buffer (Vilchez *et al.*, 2007) was added and samples were incubated at 37°C for 1h.

487 Further DNA extractions were performed as previously described (Vilchez *et al.*, 2007). Pooled
488 high-molecular-weight DNA was separated on a 1.0% Certified Low-Melt agarose gel (Bio-
489 Rad Laboratories, Hercules, CA, USA) at 40V, overnight. The gel area containing DNA
490 fragments of approximately 40 kb was excised, extracted by GELase treatment (Epicentre®,
491 Madison, WI, USA) and concentrated using a Centricon YM-100 (Millipore, Bedford, MA,
492 USA). DNA fragments of approximately 40 kb were end-repaired and cloned into the fosmid
493 copy-control pCC1FOS™ vector (Epicentre®, Madison, WI, USA) and ligation, packaging
494 into phage particles and transfection in *E. coli* EPI300-T1^R were performed according to the
495 recommended protocol. Primary libraries were stored transfected into *E. coli* EPI300-T1^R.
496 Amplicon libraries targeting the hypervariable V5/V6 region of the 16S rRNA gene, were
497 prepared, sequenced as previously described (Burbach *et al.*, 2016). The data-set quality
498 filtering was performed as previously described ((Burbach *et al.*, 2016). Reads were then
499 trimmed conservatively to 140 nt and the paired ends were subsequently matched to give 280
500 nt. The Mothur program unique.seqs was used to collapse the reads. The paired-end reads were
501 merged to fully cover the V5–V6 region and no mismatch was allowed in the overlapping of
502 the forward and reverse reads. All reads were clustered allowing 2 mismatches (>99% sequence
503 identity). Phylotypes were assigned to a taxonomic affiliation based on the naïve Bayesian
504 classification with a threshold of 80%. Phylotypes were then manually analysed against the
505 RDP database using the Seqmatch function to define the discriminatory power of each
506 sequence read. A species name was assigned to a phylotype when only 16S rRNA gene
507 fragments of previously described isolates of that species showed ≤ 2 mismatches with the
508 respective representative sequence read. Similarly, a genus name was assigned to a phylotype
509 when only 16S rRNA gene fragments of previously described isolates belonging to that genus
510 and of 16S rRNA gene fragments originating from uncultured representatives of that genus
511 showed ≤ 2 mismatches.

512

513 *Activity-based screening*

514 *E. coli* EPI300-T1^R harboring fosmid clones were plated at a density of approximately 7
515 colonies per cm² on LB agar plates, supplemented with 12.5 µg ml⁻¹ chloramphenicol and
516 CopyControl™ Induction Solution (Epicentre®, Madison, WI, USA). Plates were incubated at
517 37°C for 36 h and then sprayed with filter-sterilized 2,3-dihydroxybiphenyl (10 mM). Colonies
518 that turned yellow due to extradiol cleavage were purified and stored as glycerol cultures at
519 -80°C.

520

521 *PCR screening of activity-selected clones*

522 Selected phylogenomic clusters of extradiol dioxygenases of the vicinal oxygen chelate
523 superfamily were used as targets for PCR amplification by nine degenerate primer sets (Table
524 S2) as detailed in the Supplementary Experimental procedures. To evaluate the uniqueness of
525 clones exhibiting 2,3-dihydroxybiphenyl 1,2-dioxygenase activity, primers were designed to
526 specifically target the genes identified after sequencing 20 selected clones (Table S2).

527

528 *Development of PCR primers and screening of activity-selected clones*

529 The CDS of proteins of clusters II, XII, XVI and XXII (Duarte *et al.*, 2014) of extradiol
530 dioxygenases of the vicinal oxygen chelate superfamily (EXDO-VC), which typically exhibit
531 2,3-dihydroxybiphenyl 1,2-dioxygenase activity, were collected and used for primer design
532 resulting in the eight degenerate primer sets shown in Table S1, including 3 primer sets
533 previously described (Brennerova *et al.*, 2009). A ninth degenerate primer set (EXDO-9,
534 previously reported as EXDO-D by (Brennerova *et al.*, 2009) targeting enzymes with a
535 preference for monocyclic aromatics, specifically those affiliated to the phylogenomic cluster
536 XX of family EXDO-VC1 (Duarte *et al.*, 2014), was also used in this study (Table S1).

Additionally, eleven primer sets were designed targeting genes encoding key enzymes of aromatic degradation identified in the first 20 sequenced clones (Table S2) and used to evaluate the uniqueness of clones exhibiting 2,3-dihydroxybiphenyl 1,2-dioxygenase activity and address a careful pooling of the 200 clones for sequencing (avoiding similar fosmid inserts in the same pooled set as this would otherwise result in poor assembly).

PCR-based screening was performed on 96 DNA pools, each containing the DNA of 8 selected active fosmid clones as template followed by screening of individual clones. All amplifications were performed as detailed in Table S1, at an annealing temperature of 55°C, except for the primer set EXDO-2 (targeting *dbtC* genes encoding EXDOs involved in dibenzothiophen degradation), which was performed at an annealing temperature of 57°C. DNA of *Escherichia coli* EPI300-T1^R devoid of fosmid copy-control pCC1FOSTM was used as template to visualize possible unspecific amplifications with host genomic DNA. PCR products were separated on 1.5% agarose gels and visualized by ethidium bromide staining.

Fosmid pooling, DNA extraction, purification and sequencing

Selected clones were grown overnight at 30°C in LB medium supplemented with 12.5 µg ml⁻¹ chloramphenicol and CopyControlTM Induction Solution (Epicentre®, Madison, WI, USA). Fosmid DNA was extracted in decuplicates (10x 4 mL) from equal amounts of cells representing 20 different clones using the Qiaprep® Spin Miniprep kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The DNA from the decuplicate extractions was pooled, embedded into 1.0% Certified Low-Melt agarose (Bio-Rad Laboratories, Hercules, CA, USA) and subjected to electrophoresis overnight at 40V. The gel area containing DNA fragments of approximately 40 kb was excised, extracted by GELase treatment (Epicentre®, Madison, WI, USA) and concentrated using a Centricon YM-100 (Millipore, Bedford, MA, USA). Vestigial genomic DNA was removed using the Plasmid-safe

ATP-dependent DNase (Epicentre®, Madison, WI, USA) according to the manufacturer's recommendations for midi-preparation, followed by overnight ethanol precipitation at -20°C. The pellet obtained after centrifugation was dried, resuspended in 160 µL of nuclease-free water and DNA quantified using Quant-iT PicoGreen dsDNA reagent (Invitrogen, Darmstadt, Germany).

In a pilot experiment, fosmid DNA from 20 active clones was fragmented with NEBNext dsDNA Fragmentase (New England BioLabs, Ipswich, MA, USA) to yield 100-800 bp DNA fragments, purified by overnight ethanol precipitation at -20°C, and used for the preparation of a single library using the TrueSeq DNA Sample Prep Kit v2 (Illumina, California, USA). Fragments of 250-650bp in size were excised after agarose gel electrophoresis and purified using the QIAquick Gel extraction Kit (Qiagen, Hilden, Germany). The library was subjected to 148 nt paired-end sequencing on a GAIIx Genome Analyzer (Illumina, California, USA) using the Paired End Cluster Generation Kit v4 and TruSeq SBS Kit v5 (Illumina, California, USA) which generated 3.16 million sequence reads. Image analysis and base calling were accomplished using the Illumina CASAVA Pipeline (version 1.8.1).

Fosmid DNA from the 10 sets of 20 fosmids each was (including the 14 fosmids that were not fully assembled in the pilot experiment comprising the sequencing of 20 clones and 186 selected positive clones, including representative amounts of those that yielded a positive amplification with primer sets EXDO-TolM, EXDO-10, EXDO-16 and LigB-2, see Table S2), sonicated into short fragments (~400 bp) by Covaris DNA shearing with microTUBEs according to the manufacturer's instructions (Covaris, Woburn, MA, USA), purified with the NucleoSpin® Gel and PCR Clean Up kit (Macherey Nagel, Düren, Germany), end-repaired using the NEBNext® End Repair Module (New England Biolabs, Ipswich, MA, USA) and used for the preparation of 10 libraries using the NEBNext® DNA Library Preparation kit (New England Biolabs, Ipswich, MA, USA) and purified with the Ampure® XP beads (Beckman

Coulter, Inc., Indianapolis, IN, USA). Libraries were subjected to 100 nt paired-end sequencing on a HiSeq™ 2000 (Illumina, California, USA) using the TruSeq SBS kit v3 HS (200 cycles), and the TruSeq PE cluster kit v3 cBot HS (Illumina, California, USA) which generated an average of 21 million sequence reads per Index. Image analysis and base calling were accomplished using the Illumina Pipeline (RTA version 1.13.48).

Bioinformatic analysis and annotation of the sequenced metagenomic DNA

The obtained forward reads were processed using a quality filter program that runs a sliding window of 10% of the read length over each read and calculates the local average score based on the Illumina quality line of the fasta file and trims the 3'-ends of the reads that fall below a Phred quality score of 10 (<http://bioinformatics.ucdavis.edu/index.php/Trim.pl>). Only reads with a minimum length of 50 nucleotides were further analysed. All reads that had an N character in their sequence or more than 8 homopolymer stretches were discarded using mothur's "trim.seqs" command (Schloss *et al.*, 2009).

The high quality forward reads were depleted of sequences of the pCC1Fos vector and the genome of *E. coli* K12 DH10B (GenBank: CP000948) and assembled using Velvet (version 1.2.07) (Zerbino and Birney, 2008). Contiguous stretches with less than 1000 nucleotides were discarded. Remaining contigs were annotated using the RAST server (Aziz *et al.*, 2008; Overbeek *et al.*, 2014). As the first set of 20 fosmids contained multiple fosmid inserts exhibiting high similarity to the genome of *Pseudoxanthomonas spadix* BD-a59, reads that showed >99% sequence identity to the *P. spadix* BD-a59 genome were removed from that dataset and analysed separately. Also, reads that exhibited high similarity to the plasmid pDTG1 of *Pseudomonas putida* were removed from the dataset and analysed separately. The remaining sequences were re-assembled using Velvet (Zerbino and Birney, 2008) as described above and annotated using the RAST server. Genes encoding enzymes of superfamilies

possibly involved in the biodegradation of aromatics were further manually curated using AromaDeg (Duarte *et al.*, 2014).

Nucleotide sequence accession numbers

The nucleotide sequences reported in this study were deposited in the Genbank database under the accession numbers KX575869 - KX576454.

Conflict of Interest

The authors declare no conflict of interest.

Acknowledgements

This work was financially supported by the European Community Project MAGICPAH (FP7-KBBE-2009-245226). We gratefully acknowledge Silke Kahl for technical assistance, Diego Chaves Moreno and Walter-Israel Barrantes-Bustinza for bioinformatic support.

References

- Amann, R.I., Ludwig, W., and Schleifer, K.H. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143-169.
- Asturias, J.A., and Timmis, K.N. (1993) Three different 2,3-dihydroxybiphenyl-1,2-dioxygenase genes in the gram-positive polychlorobiphenyl-degrading bacterium *Rhodococcus globerulus* P6. *J Bacteriol* **175**: 4631-4640.
- Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A. *et al.* (2008) The RAST Server: rapid annotations using subsystems technology. *BMC Genomics* **9**: 75.

636 Bosch, R., Moore, E.R.B., Garcia-Valdes, E., and Pieper, D.H. (1999) NahW, a novel,
 637 inducible salicylate hydroxylase involved in mineralization of naphthalene by *Pseudomonas*
 638 *stutzeri* AN10. *J Bacteriol* **181**: 2315-2322.

639 Brennerova, M.V., Josefiova, J., Brenner, V., Pieper, D.H., and Junca, H. (2009) Metagenomics
 640 reveals diversity and abundance of *meta*-cleavage pathways in microbial communities from soil
 641 highly contaminated with jet fuel under air-sparging bioremediation. *Environ Microbiol* **11**:
 642 2216-2227.

643 Burbach, K., Seifert, J., Pieper, D.H., and Camarinha-Silva, A. (2016) Evaluation of DNA
 644 extraction kits and phylogenetic diversity of the porcine gastrointestinal tract based on Illumina
 645 sequencing of two hypervariable regions. *MicrobiologyOpen* **5**: 70-82.

646 Chow, K.T., Pope, M.K., and Davies, J. (1999) Characterization of a vanillic acid non-
 647 oxidative decarboxylation gene cluster from *Streptomyces* sp. D7. *Microbiology* **145**: 2393-
 648 2403.

649 Cronin, C.N., Kim, J., Fuller, J.H., Zhang, X., and McIntire, W.S. (1999) Organization and
 650 sequences of p-hydroxybenzaldehyde dehydrogenase and other plasmid-encoded genes for
 651 early enzymes of the p-cresol degradative pathway in *Pseudomonas putida* NCIMB 9866 and
 652 9869. *DNA Seq* **10**: 7-17.

653 Culligan, E.P., Sleator, R.D., Marchesi, J.R., and Hill, C. (2014) Metagenomics and novel gene
 654 discovery. *Virulence* **5**: 399-412.

655 Delmont, T.O., Robe, P., Clark, I., Simonet, P., and Vogel, T.M. (2011) Metagenomic
 656 comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods* **86**:
 657 397-400.

658 Dennis, J.J., and Zylstra, G.J. (2004) Complete sequence and genetic organization of pDTG1,
 659 the 83 kilobase naphthalene degradation plasmid from *Pseudomonas putida* strain NCIB 9816-
 660 4. *J Mol Biol* **341**: 753-768.

661 Di Gregorio, S., Zocca, C., Sidler, S., Toffanin, A., Lizzari, D., and Vallini, G. (2004)
 662 Identification of two new sets of genes for dibenzothiophene transformation in *Burkholderia*
 663 sp. DBT1. *Biodegradation* **15**: 111-123.

664 Ding, G.C., Heuer, H., Zuhlke, S., Spiteller, M., Pronk, G.J., Heister, K. *et al.* (2010) Soil type-
 665 dependent responses to phenanthrene as revealed by determining the diversity and abundance
 666 of polycyclic aromatic hydrocarbon ring-hydroxylating dioxygenase genes by using a novel
 667 PCR detection system. *Appl Environ Microbiol* **76**: 4765-4771.

668 Duarte, M., Jauregui, R., Vilchez-Vargas, R., Junca, H., and Pieper, D.H. (2014) AromaDeg, a
 669 novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database*
 670 (Oxford) **2014**: :bau118.

671 Duffner, F.M., Kirchner, U., Bauer, M.P., and Muller, R. (2000) Phenol/cresol degradation by
 672 the thermophilic *Bacillus thermoglucosidasius* A7: cloning and sequence analysis of five genes
 673 involved in the pathway. *Gene* **256**: 215-221.

674 Dutta, T.K., and Gunsalus, I.C. (1997) Reductase gene sequences and protein structures: p-
 675 cymene methyl hydroxylase. *Biochem Biophys Res Commun* **233**: 502-506.

676 Eaton, R.W. (1994) Organization and evolution of naphthalene catabolic pathways: sequence of
 677 the DNA encoding 2-hydroxychromene-2-carboxylate isomerase and trans-o-
 678 hydroxybenzylidenepyruvate hydratase-aldolase from the NAH7 plasmid. *J Bacteriol* **176**:
 679 7757-7762.

680 Eaton, R.W. (1996) p-Cumate catabolic pathway in *Pseudomonas putida* Fl: cloning and
 681 characterization of DNA carrying the cmt operon. *J Bacteriol* **178**: 1351-1362.

682 Eltis, L.D., and Bolin, J.T. (1996) Evolutionary relationships among extradiol dioxygenases. *J*
 683 *Bacteriol* **178**: 5930-5937.

684 Guazzaroni, M.E., Herbst, F.A., Lores, I., Tamames, J., Pelaez, A.I., Lopez-Cortes, N. *et al.*
 685 (2013) Metaproteogenomic insights beyond bacterial response to naphthalene exposure and
 686 bio-stimulation. *Isme J* **7**: 122-136.

687 Gutierrez, T., Green, D.H., Nichols, P.D., Whitman, W.B., Semple, K.T., and Aitken, M.D.
 688 (2013) *Polycyclovorans algicola* gen. nov., sp. nov., an aromatic-hydrocarbon-degrading
 689 marine bacterium found associated with laboratory cultures of marine phytoplankton. *Appl*
 690 *Environ Microbiol* **79**: 205-214.

691 Harayama, S., Mermod, N., Rekik, M., Lehrbach, P.R., and Timmis, K.N. (1987) Roles of the
 692 divergent branches of the meta-cleavage pathway in the degradation of benzoate and
 693 substituted benzoates. *J Bacteriol* **169**: 558-564.

694 Heiss, G., Muller, C., Altenbuchner, J., and Stolz, A. (1997) Analysis of a new dimeric
 695 extradiol dioxygenase from a naphthalenesulfonate-degrading sphingomonad. *Microbiology*
 696 **143**: 1691-1699.

697 Heiss, G., Stolz, A., Kuhm, A.E., Muller, C., Klein, J., Altenbuchner, J., and Knackmuss, H.J.
 698 (1995) Characterization of a 2,3-dihydroxybiphenyl dioxygenase from the
 699 naphthalenesulfonate-degrading bacterium strain BN6. *J Bacteriol* **177**: 5865-5871.

700 Hofer, B., Eltis, L.D., Dowling, D.N., and Timmis, K.N. (1993) Genetic analysis of a
 701 *Pseudomonas* locus encoding a pathway for biphenyl/polychlorinated biphenyl degradation.
 702 *Gene* **130**: 47-55.

703 Jouanneau, Y., Micoud, J., and Meyer, C. (2007) Purification and characterization of a three-
 704 component salicylate 1-hydroxylase from *Sphingomonas* sp. strain CHY-1. *Appl Environ*
 705 *Microbiol* **73**: 7515-7521.

706 Kasai, Y., Shindo, K., Harayama, S., and Misawa, N. (2003) Molecular characterization and
 707 substrate preference of a polycyclic aromatic hydrocarbon dioxygenase from *Cycloclasticus* sp.
 708 strain A5. *Appl Environ Microbiol* **69**: 6688-6697.

709 Lam, K.N., Cheng, J., Engel, K., Neufeld, J.D., and Charles, T.C. (2015) Current and future
710 resources for functional metagenomics. *Front Microbiol* **6**: 1196.

711 Laurie, A.D., and Lloyd-Jones, G. (1999) The *phn* genes of *Burkholderia* sp. strain RP007
712 constitute a divergent gene cluster for polycyclic aromatic hydrocarbon catabolism. *J Bacteriol*
713 **181**: 531-540.

714 Liebl, W., Angelov, A., Juergensen, J., Chow, J., Loeschcke, A., Drepper, T. *et al.* (2014)
715 Alternative hosts for functional (meta)genome analysis. *Appl Microbiol Biotechnol* **98**: 8099-
716 8109.

717 Lu, J., Nomura, N., Nakajima-Kambe, T., and Nakahara, T. (2000) Cloning and expression of
718 genes encoding meta-cleavage enzymes from 4,6-dimethyldibenzothiophene-degrading
719 *Sphingomonas* strain TZS-7. *Biochim Biophys Acta* **1492**: 493-498.

720 Lupa, B., Lyon, D., Gibbs, M.D., Reeves, R.A., and Wiegel, J. (2005) Distribution of genes
721 encoding the microbial non-oxidative reversible hydroxyarylic acid decarboxylases/phenol
722 carboxylases. *Genomics* **86**: 342-351.

723 Maeda, R., Ishii, T., Ito, Y., Zulkharnain, A.B., Iwata, K., and Omori, T. (2010) Isolation and
724 characterization of the gene encoding the chloroplast-type ferredoxin component of carbazole
725 1,9a-dioxygenase from a putative *Kordiimonas* sp. *Biotechnol Lett* **32**: 1725-1731.

726 McKay, D.B., Prucha, M., Reineke, W., Timmis, K.N., and Pieper, D.H. (2003) Substrate
727 specificity and expression of three 2,3-dihydroxybiphenyl 1,2-dioxygenases from *Rhodococcus*
728 *globerulus* strain P6. *J Bacteriol* **185**: 2944-2951.

729 Nakatsu, C.H., Straus, N.A., and Wyndham, R.C. (1995) The nucleotide sequence of the
730 Tn5271 3-chlorobenzoate 3,4-dioxygenase genes (*cbaAB*) unites the class IA oxygenases in a
731 single lineage. *Microbiology* **141**: 485-495.

732 Overbeek, R., Olson, R., Pusch, G.D., Olsen, G.J., Davis, J.J., Disz, T. *et al.* (2014) The SEED
 733 and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST).
 734 *Nucleic Acids Res* **42**: D206-214.

735 Overhage, J., Priefert, H., Rabenhorst, J., and Steinbuchel, A. (1999) Biotransformation of
 736 eugenol to vanillin by a mutant of *Pseudomonas* sp. strain HR199 constructed by disruption of
 737 the vanillin dehydrogenase (*vdh*) gene. *Appl Microbiol Biotechnol* **52**: 820-828.

738 Peng, R.H., Xiong, A.S., Xue, Y., Fu, X.Y., Gao, F., Zhao, W. *et al.* (2008) Microbial
 739 biodegradation of polyaromatic hydrocarbons. *FEMS Microbiol Rev* **32**: 927-955.

740 Pérez-Pantoja, D., Donoso, R., Junca, H., González, B., and Pieper, D.H. (2010)
 741 Phylogenomics of aerobic bacterial degradation of aromatics. In *Handbook of Hydrocarbon
 742 and Lipid Microbiology*. Timmis, K. (ed): Springer Berlin Heidelberg, pp. 1355-1397.

743 Peters, F., Heintz, D., Johannes, J., van Dorsselaer, A., and Boll, M. (2007) Genes, enzymes,
 744 and regulation of para-cresol metabolism in *Geobacter metallireducens*. *J Bacteriol* **189**: 4729-
 745 4738.

746 Pinyakong, O., Habe, H., and Omori, T. (2003) The unique aromatic catabolic genes in
 747 sphingomonads degrading polycyclic aromatic hydrocarbons (PAHs). *J Gen Appl Microbiol*
 748 **49**: 1-19.

749 Salonen, A., Nikkila, J., Jalanka-Tuovinen, J., Immonen, O., Rajilic-Stojanovic, M., Kekkonen,
 750 R.A. *et al.* (2010) Comparative analysis of fecal DNA extraction methods with phylogenetic
 751 microarray: effective recovery of bacterial and archaeal DNA using mechanical cell lysis. *J*
 752 *Microbiol Methods* **81**: 127-134.

753 Schloss, P.D., Westcott, S.L., Ryabin, T., Hall, J.R., Hartmann, M., Hollister, E.B. *et al.* (2009)
 754 Introducing mothur: open-source, platform-independent, community-supported software for
 755 describing and comparing microbial communities. *Appl Environ Microbiol* **75**: 7537-7341.

756 Singleton, D.R., Ramirez, L.G., and Aitken, M.D. (2009) Characterization of a polycyclic
 757 aromatic hydrocarbon degradation gene cluster in a phenanthrene-degrading *Acidovorax* strain.
 758 *Appl Environ Microbiol* **75**: 2613-2620.

759 Steele, H.L., Jaeger, K.E., Daniel, R., and Streit, W.R. (2009) Advances in recovery of novel
 760 biocatalysts from metagenomes. *J Mol Microbiol Biotechnol* **16**: 25-37.

761 Suenaga, H., Ohnuki, T., and Miyazaki, K. (2007) Functional screening of a metagenomic
 762 library for genes involved in microbial degradation of aromatic compounds. *Environ Microbiol*
 763 **9**: 2289-2297.

764 Terron-Gonzalez, L., Martin-Cabello, G., Ferrer, M., and Santero, E. (2016) Functional
 765 metagenomics of a biostimulated petroleum-contaminated soil reveals an extraordinary
 766 diversity of extradiol dioxygenases. *Appl Environ Microbiol* **82**: 2467-2478.

767 Vilchez, R., Pozo, C., Gómez, M.A., Rodelas, B., and González-López, J. (2007) Dominance of
 768 sphingomonads in a copper-exposed biofilm community for groundwater treatment.
 769 *Microbiology* **153**: 325-337.

770 Vilchez-Vargas, R., Junca, H., and Pieper, D.H. (2010) Metabolic networks, microbial ecology
 771 and 'omics' technologies: towards understanding in situ biodegradation processes. *Environ*
 772 *Microbiol* **12**: 3089-3104.

773 Vilchez-Vargas, R., Geffers, R., Suarez-Diez, M., Conte, I., Waliczek, A., Kaser, V.S. *et al.*
 774 (2013) Analysis of the microbial gene landscape and transcriptome for aromatic pollutants and
 775 alkane degradation using a novel internally calibrated microarray system. *Environ Microbiol*
 776 **15**: 1016-1039.

777 Witzig, R., Junca, H., Hecht, H.J., and Pieper, D.H. (2006) Assessment of toluene/biphenyl
 778 dioxygenase gene diversity in benzene-polluted soils: links between benzene biodegradation
 779 and genes similar to those encoding isopropylbenzene dioxygenases. *Appl Environ Microbiol*
 780 **72**: 3504-3514.

781 Worsey, M.J., and Williams, P.A. (1975) Metabolism of toluene and xylenes by *Pseudomonas*
782 *putida* (arvilla) mt-2: evidence for a new function of the TOL plasmid. *Journal of Bacteriology*
783 **124**: 7-13.

784 You, I.S., Ghosal, D., and Gunsalus, I.C. (1991) Nucleotide sequence analysis of the
785 *Pseudomonas putida* PpG7 salicylate hydroxylase gene (nahG) and its 3'-flanking region.
786 *Biochemistry* **30**: 1635-1641.

787 Zerbino, D.R., and Birney, E. (2008) Velvet: algorithms for de novo short read assembly using
788 de Bruijn graphs. *Genome Res* **18**: 821-829.

789 .

790 **Table 1.** Total numbers of extradiol dioxygenases and Rieske non-heme iron oxygenases encoded in 205 sequenced fosmid inserts according to AromaDeg (Duarte *et al.*,
791 2014).

Enzyme family	cluster	function	Substrate	Representative sequence	number	ORF name (previously described gene products exhibiting >98% of sequence identity are given in parentheses)	
Rieske non-heme iron oxygenases (RNIOs)	Salicylate	I	Salicylate 1-hydroxylases	Salicylate	BAC65453 <i>Sphingomonas</i> sp. P2	19	B93=G225=J135 (AJP49471 of PG1-Ca6*); C2=E445; E15=K695; H177=K9, A408, B177, C459, F127, H39, H165, H408, K160, L351, L706
		III	RNIOs related to salicylate 5-hydroxylases	Unknown	ZP_06687231 <i>A. piechaudii</i>	1	G350
		New	RNIOs of unknown function (probably salicylate 5-hydroxylases)	Unknown	-	1	K164
		XIV	Salicylate 1-hydroxylases	Salicylate	BAC65433 <i>Sphingomonas</i> sp. P2	17	B81=E461=L220; A405, C845, D302, D581, E260, E270, F4, F147, G475, G478, G510, J7, J382, L723
		XV	Salicylate 1-hydroxylases	Salicylate	BAC65426 <i>Sphingomonas</i> sp. P2	10	G228 (AJP48192 of PG1-Ca6*); E158=K692; A416, B367, C456, D283, G517, H36, J154
		New	RNIOs of unknown function (probably salicylate 1-hydroxylases)	Unknown	-	1	G501
		XVI	Probable salicylate 1-hydroxylases	Probably salicylate	YP_008374324 <i>C. zancles</i> 7 ME	1	E2
		XVII	Probable salicylate 1-hydroxylases	Probably salicylate	YP_002361789 <i>M. silvestris</i> BL2	1	B173
		Total				51	
	Benzoate	II	RNIOs of unknown function	Unknown	WP_008734266 <i>A. pacificus</i>	5	C616=D470; F488, G281, J259
		III	RNIOs of unknown function	Unknown	YP_006837858 <i>Cycloclasticus</i> sp. P1	2	D634, J564
		IV	Ibuprofen-CoA dioxygenases and related enzymes	Ibuprofen-CoA	ABK92277 <i>Sphingomonas</i> sp. Ibu-2	2	D349, K359
		VIII	p-Cumate dioxygenases	p-Cumate	YP_001268212 <i>P. putida</i> F1	11	B1=F162=H337=L590; B358, B361, D515, D516, F66, K316, L457
		IX	2,4-D and 2,4,5-T oxygenases	2,4-Dichlorophenoxyacetate, 2,4,5-trichlorophenoxyacetate	AAB39767 <i>B. cepacia</i>	8	B165=D540=F429=H263=K763=L378; E452, G266

Rieske non-heme iron oxygenases (RNIOs)	Benzoate (cont.)	XI	Benzoate and 2-Chloro-benzoate dioxygenases	Benzoate, 2-Chlorobenzoate	YP_709342 <i>P. putida</i> KT2440	38	B373=J77; B505=J574; C455=H33; C783=E133=F339=G157=H315=L473; D244=E47; D312=H527; D313=H528; G229=J155; J192=L231; J193=L230; B208, B418, E159, E432, F129, F136, F137, G60, G61, G427, G513, J562, K21, K563, K691, L64
		New	RNIOs of unknown function	Unknown	-	3	D244, E47, J569
		Total				69	
	Phthalate	I	Carbazol and diphenylamin dioxygenases	Carbazol and Diphenylamine	BAA31266 <i>P. stutzeri</i>	3	A25, E3, E53
		II	1-Oxo-2,3-dihydroquinoline monooxygenases and related enzymes	1-Oxo-2,3-Dihydroquinoline	CAA73203 <i>P. putida</i>	2	B131, L747
		XI	3-Chlorobenzoate dioxy-genases and related enzymes	3-Chlorobenzoate	Q44256 <i>C. testosteroni</i>	1	K163
		New	RNIOs of unknown function	Unknown		3	J253 (AJP49503 PG1 Ca6*), J209, K58
		Total				9	
	Biphenyl	II	Naphthalene dioxygenases of <i>Actinobacteria</i>	Naphthalene	ABH01029 <i>R. opacus</i>	1	J502
		XIII	RNIOs of <i>Actinobacteria</i>	Unknown	WP_007015351 <i>N. pentaromativorans</i>	1	K169
		XVII	Diterpenoid dioxygenases and related enzymes	Abietane diterpenoids	AAD21063 <i>P. abietaniphilia</i>	1	G176
		XVIII	Angular dioxygenases of dibenzofuran metabolism and related enzymes	Dibenzofuran, fluorene	BAE530401 <i>Paenibacillus</i> sp. YK5	2	H558, K721
		XXIV	Naphthalene dioxygenases of <i>Proteobacteria</i> and related enzymes	Naphthalene, phenanthrene, biphenyl, among others	O07824 <i>P. fluorescens</i>	25	A214=C911 (NP863072 of <i>P. putida</i> NCIB 9816-4); F224 (BAD344471 of <i>Sphingomonas</i> sp. A4); L468 (CAT034469 of <i>S. paucimobilis</i> EPA505); G233=J159 (AJP48195 of PG1-Ca6*); B80=C621=E460=H137=J8=L219; C451=H29; D301=F148 =H77; D380=F511; A412, E163, F424, G47, G188, K687
		XXVI	Biphenyl dioxygenases of <i>Proteobacteria</i>	Biphenyl	ABM79809 <i>S. yanoikuyae</i>	8	B95=G223=J137 (AJP48190 of PG1-Ca6*); E153=K697; A402, F379, H41
		XXVII	Tetralin, biphenyl, ethylbenzene dioxygenases and related enzymes	Tetralin, biphenyl, ethylbenzene	ACL31215 <i>Rhodococcus</i> sp. TFB	2	B36, F377
		XXX	RNIOs of unknown function	Unknown	WP_020122046 <i>S. canus</i>	7	D190, J522, J523, K227, K228, K229, K408
		XXXI	RNIOs of unknown function	Unknown	YP_001261395 <i>S. wittichii</i> RW1	1	L687

Extradiol dioxygenases	Biphenyl (cont.)	XXXIX	Phenylpropionate dioxy-genases of <i>Proteobacteria</i>	Phenylpropionate	CAA86018 <i>E. coli</i>	1	C556
		XL	Biphenyl/benzene dioxygenases	Biphenyl, benzene and related substrates	YP_556409 <i>B. xenovorans</i> LB400	4	H233 (YP_973181 of <i>P. naphthalenivorans</i> CJ2); D514, G453, H233, H469
		New	RNIOs of unknown function	Unknown	-	5	A310=E297; D242, D482, E358
		Total					58
	EXDO-VC1	XIV	Enzymes related to catechol 2,3-dioxygenase of <i>Bacillus</i> sp. JF8	Catechol	BAD08308 <i>Bacillus</i> sp. JF8	1	J506
		XVI	Probable catechol 2,3-dioxygenases of α - <i>Proteobacteria</i>	Probably catechol	ZP_03499467 <i>R. etli</i> Kim 5	1	C576
		XVII	Probable catechol 2,3-dioxygenases	Probably catechol	YP_003590156 <i>B. tusciae</i> DSM 2912	23	B109=G209=J414 (AJP48179 of PG1-Ca6*); D101 (AJP48861 of PG1-Ca6*); B275=C408=E185=H110=J227=K268=L659; B557=E402=J375=L557; C828=F353; C466, D8, D101, G336, H9, H55, H424
		XVIII	Probable catechol 2,3-dioxygenases	Probably catechol	YP_003590157 <i>B. tusciae</i> DSM 2912	23	B110=J413; B276=C407=E186=J226=K269=L660; B556=E401=J374=L556; C504=F352; C467, D7, D102, E145, G208, H8, H57, H111, H423
		XX	Proteobacterial catechol 2,3-dioxygenases	Catechol	AAD05250 <i>P. putida</i> GJ31	12	B107=G211=J416 (AJP48180 of PG1-Ca6*); B185=E340=K540 (YP_004930496 of <i>P. spadix</i> BD-a59); J72 (YP_004930521 of <i>P. spadix</i> BD-a59); C464, E147, H53, K703, L70
		XXI	Actinobacterial / <i>Firmicutes</i> catechol 2,3-dioxygenases	Catechol	AAC38323 <i>G. thermoleovorans</i>	1	F291
		XXII	Proteobacterial catechol 2,3-dioxygenases	Catechol	CAA54847 <i>P. putida</i>	15	C871 (NP_863103 of <i>P. putida</i> NCIB 9816-4); E212 (WP_010335345 of <i>S. yanoikuyae</i> B1); B508=F453=L578; D571=G521=H293; B171, B368, C871, F120, F212, G488, G506, L614
		New	EXDO-VC1 of unknown function	Unknown	-	3	H56, J24, L812
		New	EXDO-VC1 of unknown function	Unknown	-	1	C671
		Total					80
	EXDO-VC2	XII	Dihydroxynaphthalene dioxygenases and related enzymes in <i>Proteobacteria</i>	Probably dihydroxynaphthalene, dihydroxyphenanthrene, among others	2EI2 <i>Pseudomonas</i> sp. C18	20	A218=C907 (2EI2 of <i>P. putida</i> C18); B102=G215=J145 (AJP48184 of PG1-Ca6*); F469 (AEF88782 of <i>Delftia</i> sp. Cs1-4); F2 (BAE93945 of <i>Shingomonas</i> sp. A4); G514 (AAB06725 of <i>S. xenophagum</i> BN6); C461=H49; J384=L793; A414, D207, E150, F333, H103, K88, K700, L289
		XIII	Catechol 2,3-dioxygenases from <i>Actinobacteria</i>	Catechol, maybe others	Q53034 <i>R. rhodochrous</i>	2	J527, K31

EXDO-VC2 (cont.)	XVII	Protobacterial 2,3-dihydroxybiphenyl dioxygenases	2,3-Dihydroxybiphenyl	AAB57639 <i>P. nitroreducens</i> Hbp1	1	G491
	XIX	EXDO-VC2 of unknown function	Unknown	WP_007898103 <i>Pseudomonas</i> sp. GM102	2	B129, L744
	XX	EXDO-VC2 of unknown function	Unknown	WP_006577158 <i>Cupriavidus</i> sp. HPC(L)	3	D277=K600; B354
	XXII	2,3-Dihydroxybiphenyl dioxygenases of Proteo-bacteria and Actinobacteria	2,3-Dihydroxybiphenyl, catechol among others	YP_556403 <i>B. xenovorans</i> LB400	7	G447 (CAA62978 of <i>P. putida</i> OU83); H226 (YP_973175 of <i>P. naphthalenivorans</i> CJ2); D66, D202, F236, H463, J251
	New	EXDO-VC2 of unknown function	Unknown	-	10	D510, F328, G308, H189, H428, H502, J293, K114, K852, L89
	Total					45
EXDO-VC3	I	2,3-Dihydroxy-p-cumate dioxy-genases and 2,3-dihydroxy-benzoate dioxygenases	2,3-Dihydroxy-p-cumate and 2,3-dihydroxybenzoate	YP_001268210 <i>P. putida</i> F1	12	B3=C763=E250; F164=G204=H335=L588; B356, D518, F68, K318, L545
	V	Putative 2,3-dihydroxy-biphenyl dioxygenases	Unknown	WP_007015343 <i>N. pentaromativorans</i> US6-1	1	K172
	XII	EXDO-VC3 of unknown function	Unknown	YP_003060960 <i>H. baltica</i> ATCC 49814	1	D143
	XIV	2,3-Dihydroxybiphenyl dioxygenases and enzymes of methylhydroquinone metabolism	2,3-Dihydroxybiphenyl, methylhydroquinone	BAE46530 <i>Burkholderia</i> sp. NF100	6	A60, B226, G28, G178, G337, K708
	New	EXDO-VC3 of unknown function	Unknown	-	4	C664=D58; D170, K423
	Total					24
LigB Protocatechuate	XII	Extradiol dioxygenases related to PhnC probably transforming polycyclic arene diols	Probably dihydroxyphenanthrene	AAD09870 <i>Burkholderia</i> sp. RP007	24	G232=J158 (AJP48194 of PG1 Ca6*); B77=C583=E457=F151=G50=H7=J11=L216; B82=L221; C463=F146; D298=H14; C452, D303, E162, E462, F421, H30, H152, K688
	XVII	2,2-Dihydroxy-p-cumate dioxygenase of Actinobacteria and related enzymes	2,3-Dihydroxy-p-cumate	ABC33900 <i>Rhodococcus</i> sp. T104	1	F486
	XX	Extradiol dioxygenases possibly involved in fluorene degradation	Unknown	AB270530 <i>Sphingomonas</i> sp. KA1	1	D196
	XXII	Aminobiphenyl-2,3-diol 1,2-dioxygenases	2'-Aminobiphenyl-2,3-diol	ADC31796 <i>Sphingomonas</i> sp. XLDN2-5	4	A22, E51, D247, B363
	XXXII	Putative 2,3-dihydroxy-phenylpropionate dioxygenases	Unknown	ZP_05362403 <i>A. radioresistens</i> SK82	5	F262=L390; J469, C551, H217
	New	Extradiol dioxygenases of unknown function	Unknown	-	6	G415=E119; A311, E296, K179, K420
Total					41	

Extradiol dioxygenases	Gentisate	XII	Gentisate dioxygenases of <i>Actinobacteria</i> and <i>Proteobacteria</i>	Gentisate, Salicylate	ABH01038 <i>R. opacus</i>	1	J491
		XIII	1-Hydroxy-2-naphthoate dioxygenases of <i>Proteobacteria</i>	1-Hydroxy-2-naphthoate	BAA76328 <i>A. faecalis</i>	1	F466
		Total				2	

* PG1 Ca6, *Rhodocyclaceae* bacterium PG1 Ca6

Fig. 4. Evolutionary relationships of dihydroxynaphthalene dioxygenases and related enzymes in Proteobacteria (cluster XII members of family 2 of extradiol dioxygenases of the vicinal oxygen chelate superfamily (EXDO-VC2), typically characterized by a preference for bicyclic substrates, for an overview of evolutionary relationships among members of the complete family, see Supporting Information Fig. S6) where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots. Proteins from databases where a function and/or substrate has been described are indicated by a three- or four-letter code: Dhn, Dihydroxynaphthalene; DbtC, (probably) dihydroxydibenzothiophene; Dhp, (probably) dihydroxyphenanthrene; Dhe, 2,3-dihydroxy-1-ethylbenzene; Dhb, 2,3-Dihydroxybiphenyl; Thn, 1,2-dihydroxy-5,6,7,8-tetrahydronaphthalene; Proteins without proven function are identified by the three-letter code – Non.

Fig. 5. Evolutionary relationships of extradiol dioxygenases related to PhnC probably transforming polycyclic arene diols (cluster XII members of the protocatechuate family of the LigB superfamily extradiol dioxygenases, for an overview of evolutionary relationships among members of the complete family, see Supporting Information Fig. S8) where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots. Proteins from databases where a function and/or substrate has been described are indicated by a three- or four-letter code: PhnC, (probably) dihydroxyphenanthrene. Proteins without proven function are identified by the three-letter code – Non.

Fig. 6. Overview of the degradation potential for monocyclic aromatic compounds present in the soil under study as exemplified for the degradation of toluene, salicylate, 3-hydroxyphenylpyruvate, 2,3-dihydroxybenzoate and p-cymene.

Fig. 7. Overview of the degradation potential for polycyclic aromatic compounds present in the soil under study as exemplified for phenanthrene, naphthalene, biphenyl and carbazole.

Supporting information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Fig. S1. Relative abundances of the dominant bacterial phyla or classes present within amplicon libraries targeting the hypervariable V5/V6 regions of the 16S rRNA gene from metagenomic DNA extracted from the soil sample under study with a mechanical lysis (A) or with an enzymatic lysis (B).

Fig. S2. Evolutionary relationships of α -subunits of members of the biphenyl family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster II - naphthalene dioxygenases of Actinobacteria; cluster XIII - Rieske oxygenases of Actinobacteria; cluster XVII - diterpenoid dioxygenases and related enzymes; cluster XVIII - angular dioxygenases of dibenzofuran or fluorene metabolism and related enzymes; cluster XXIV - naphthalene dioxygenases of Proteobacteria and related

enzymes; cluster XXVI - biphenyl dioxygenases of Proteobacteria; cluster XXVII - tetralin, biphenyl, ethylbenzene, xylene dioxygenases and related enzymes; clusters XXX and XXXI - Rieske oxygenases of unknown function; cluster XXXIX - phenylpropionate dioxygenases of Proteobacteria; cluster XL - biphenyl/benzene dioxygenases. All members that could not be affiliated to any of the clusters previously defined by the Aromadeg database are shown in red.

Fig. S3. Evolutionary relationships of α -subunits of members of the phthalate family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster I - carbazol and diphenylamin dioxygenases; cluster II - 1-oxo-2,3-dihydroquinoline monooxygenases and related enzymes; cluster XI - 3-chlorobenzoate dioxygenases and related enzymes, probable phthalate dioxygenase activity. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S4. Evolutionary relationships of α -subunits of members of the salicylate family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Clusters I, XIV and XV - salicylate 1-hydroxylases; cluster III - Rieske oxygenases related to salicylate 5- hydroxylases; cluster XVI - probable salicylate 1-hydroxylases; cluster XVII - anthranilate dioxygenases of *Burkholderia* and some other

organisms. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S5. Evolutionary relationships of α -subunits members of the benzoate family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Clusters II and III – Rieske oxygenases of unknown function; cluster IV - ibuprofen-CoA dioxygenases and related enzymes; cluster VIII - p-cumate dioxygenases; cluster IX - 2,4-D and 2,4,5-T oxygenases; cluster XI - benzoate and 2-chlorobenzoate dioxygenases. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S6. Evolutionary of members of family 2 of extradiol dioxygenases of the vicinal oxygen chelate superfamily (EXDO-VC2), typically characterized by a preference for bicyclic substrates where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster XII - dihydroxynaphthalene dioxygenases and related enzymes in Proteobacteria; cluster XIII - catechol 2,3-dioxygenases from *Rhodococcus rhodochrous*; cluster XVII - protobacterial 2,3-dihydroxybiphenyl dioxygenases; clusters XIX and XX - extradiol dioxygenases of unknown function; cluster XXII - 2,3-dihydroxybiphenyl dioxygenases of Proteobacteria and Actinobacteria. All

members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S7. Evolutionary relationships of members of family 1 of extradiol dioxygenases of the vicinal oxygen chelate superfamily (EXDO-VC1), typically characterized by a preference for monocyclic substrates where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster XIV - enzymes related to catechol 2,3-dioxygenase of *Bacillus* sp. JF8; cluster XVI - probable catechol 2,3-dioxygenases of α -Proteobacteria; clusters XVII and XVIII - probable catechol 2,3-dioxygenases; cluster XX - proteobacterial catechol 2,3-dioxygenases; cluster XXI - actinobacterial and Firmicutes catechol 2,3-dioxygenases; cluster XXII - proteobacterial catechol 2,3-dioxygenases. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S8. Evolutionary relationships of members of the protocatechuate family of the LigB superfamily extradiol dioxygenases where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte *et al.*, 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster XII - extradiol dioxygenases related to PhnC probably transforming polycyclic arene diols; cluster XVII - 2,2-dihydroxy-*p*-cumate dioxygenase of Actinobacteria and related enzymes; cluster XXII -

Supplementary information

Supporting figures (Fig. S1 – S11) and tables (Table S1 – S3)

Functional soil metagenomics: elucidation of polycyclic aromatic hydrocarbon degradation potential following 12 years of *in situ* bioremediation

Márcia Duarte¹, Agnes Nielsen¹, Amélia Camarinha-Silva^{1,3}, Ramiro Vilchez-Vargas^{1,4}, Thomas Bruls², Melissa L. Wos-Oxley¹, Ruy Jauregui^{1,5} and Dietmar H. Pieper¹

¹Microbial Interactions and Processes Research Group, HZI – Helmholtz Centre for Infection Research, Inhoffenstr. 7, D-38124 Braunschweig, Germany.

²Institut de Génomique, Genoscope, UMR8030 (CNRS, CEA, Université d'Evry), Evry, France.

References

Brennerova, M.V., Josefiova, J., Brenner, V., Pieper, D.H., and Junca, H. (2009) Metagenomics reveals diversity and abundance of *meta*-cleavage pathways in microbial communities from soil highly contaminated with jet fuel under air-sparging bioremediation. *Environ Microbiol* **11**: 2216-2227.

Duarte, M., Jauregui, R., Vilchez-Vargas, R., Junca, H., and Pieper, D.H. (2014) AromaDeg, a novel database for phylogenomics of aerobic bacterial degradation of aromatics. *Database* **2014**: bau118.

Edgar RC. (2004) MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res* **32**: 1792-1797.

Tamura K, Peterson D, Peterson N, Stecher G, Nei M and Kumar S. (2011). MEGA5: Molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Mol Biol Evol* **28**: 2731-2739.

Fig. S1. Evolutionary relationships of α -subunits of members of the biphenyl family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster II - naphthalene dioxygenases of Actinobacteria; cluster XIII - Rieske oxygenases of Actinobacteria; cluster XVII - diterpenoid dioxygenases and related enzymes; cluster XVIII - angular dioxygenases of dibenzofuran or fluorene metabolism and related enzymes; cluster XXIV - naphthalene dioxygenases of Proteobacteria and related enzymes; cluster XXVI - biphenyl dioxygenases of Proteobacteria; cluster XXVII - tetralin, biphenyl, ethylbenzene, xylene dioxygenases and related enzymes; clusters XXX and XXXI - Rieske oxygenases of unknown function; cluster XXXIX - phenylpropionate dioxygenases of Proteobacteria; cluster XL - biphenyl/benzene dioxygenases. All members that could not be affiliated to any of the clusters previously defined by the Aromadeg database are shown in red.

Fig. S2. Evolutionary relationships of α -subunits of members of the phthalate family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster I - carbazol and diphenylamin dioxygenases; cluster II - 1-oxo-2,3-dihydroquinoline monooxygenases and related enzymes; cluster XI - 3-chlorobenzoate dioxygenases and related enzymes, probable phthalate dioxygenase activity. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S3. Evolutionary relationships of α -subunits of members of the salicylate family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Clusters I, XIV and XV - salicylate 1-hydroxylases; cluster III - Rieske oxygenases related to salicylate 5- hydroxylases; cluster XVI - probable salicylate 1-hydroxylases; cluster XVII - anthranilate dioxygenases of *Burkholderia* and some other organisms. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S4. Evolutionary relationships of α -subunits members of the benzoate family of Rieske non-heme iron oxygenases where sequences are available from databases and those identified in the metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Clusters II and III – Rieske oxygenases of unknown function; cluster IV - ibuprofen-CoA dioxygenases and related enzymes; cluster VIII - p-cumate dioxygenases; cluster IX - 2,4-D and 2,4,5-T oxygenases; cluster XI - benzoate and 2-chlorobenzoate dioxygenases. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S5. Evolutionary of members of family 2 of extradiol dioxygenases of the vicinal oxygen chelate superfamily (EXDO-VC2), typically characterized by a preference for bicyclic substrates where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster XII - dihydroxynaphthalene dioxygenases and related enzymes in Proteobacteria; cluster XIII - catechol 2,3-dioxygenases from *Rhodococcus rhodochrous*; cluster XVII - protobacterial 2,3-dihydroxybiphenyl dioxygenases; clusters XIX and XX - extradiol dioxygenases of unknown function; cluster XXII - 2,3-dihydroxybiphenyl dioxygenases of Proteobacteria and Actinobacteria. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S7. Evolutionary relationships of members of the protococatechuate family of the LigB superfamily extradiol dioxygenases where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster XII - extradiol dioxygenases related to PhnC probably transforming polycyclic arene diols; cluster XVII - 2,2-dihydroxy-*p*-cumate dioxygenase of Actinobacteria and related enzymes; cluster XXII - aminobiphenyl-2,3-diol 1,2-dioxygenases; cluster XXIV - actinobacterial extradiol dioxygenases of unknown function; cluster XX - extradiol dioxygenases possibly involved in fluorene degradation; cluster XXXII - putative 2,3-dihydroxyphenylpropionate dioxygenases. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S8. Evolutionary relationships of members of family 3 of extradiol dioxygenases of the vicinal oxygen chelate superfamily (EXDO-VC3) (comprising, among others, enzymes using miscellaneous substrates) where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster I - 2,3-dihydroxy-*p*-cumate dioxygenases and 2,3-dihydroxybenzoate dioxygenases; cluster V - putative 2,3-dihydroxybiphenyl dioxygenases; cluster XII - extradiol dioxygenases of unknown function; cluster XIV - 2,3-dihydroxybiphenyl dioxygenases and enzymes of methylhydroquinone metabolism. All members that could not be affiliated to any of the clusters previously defined by the AromaDeg database are shown in red.

Fig. S9. Evolutionary relationships of integral-membrane non-heme diiron monooxygenases where sequences are available from databases. Enzymes of validated function as alkane monooxygenases (AlkB) or those involved in the oxidation of the methylsubstituent of aromatics (xylene monooxygenases - XylM or cymene monooxygenases - CymAa) are marked by black dots. Proteins identified in the metagenomic library related to alkane monooxygenases are marked with blue dots, whereas those related to xylene and cymene monooxygenases are marked with magenta dots. Sequences were aligned using MUSCLE (Edgar, 2004) and phylogenetic analyses conducted in MEGA5 (Tamura et al., 2011). The evolutionary histories of proteins were inferred using the neighbor-joining method and the p-distance model. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons.

Fig. S10. Evolutionary relationships of the flavoprotein α -subunits of *p*-cresol methylhydroxylases (p-cre), eugenol hydroxylases (eug) and related enzymes where sequences are available from databases. Enzymes of validated function are marked by black dots and proteins identified in the metagenomic library are marked by magenta dots. Sequences were aligned using MUSCLE (Edgar, 2004) and phylogenetic analyses conducted in MEGA5 (Tamura et al., 2011). The evolutionary histories of proteins were inferred using the neighbourjoining method and the p-distance model. All positions containing alignment gaps and missing data were eliminated only in pairwise sequence comparisons.

Fig. S11. Evolutionary relationships of members of the gentisate family of the cupin superfamily extradiol dioxygenases where sequences are available from databases and those identified by functional screening of a metagenomic library. Alignments and annotations were performed by the AromaDeg database (Duarte et al., 2014). Proteins identified to be encoded by the metagenomic library are indicated by dots, where the color of the dot indicates its inclusion in the cluster of sequences shown in the same color. Cluster XII - gentisate dioxygenases of Actinobacteria and Proteobacteria; cluster XIII - 1-hydroxy-2-naphthoate dioxygenases of Proteobacteria.

Table S1. Oligonucleotide primers used in this study.

	Primers	Target	Sequence (5'-3')	Amplicon size (bp)	PCR cycles
Degenerate primer sets	EXDO-1_F	Extradiol dioxygenases involved in cleavage of dihydroxylated polycyclic aromatic hydrocarbons in Actinobacteria (cluster II, EXDO-VC2)	GAAGAGCGAYGAGATYTMCAC	329	30
	EXDO-1_R		AYACTSGRYCGGTCTACTG		
	EXDO-2_F ^a	Dihydroxynaphthalene dioxygenases and related enzymes in Proteobacteria (cluster XII, EXDO-VC2)	TCCGCATGGATTACAACC	423	35
	EXDO-2_R ^a		GATCTGTGGAACGGGCAA		
	EXDO-3_F		GCGMATY MAYCAYYT GATG	244	35
	EXDO-3_R		CCTCRWTVYC RTGRCCGAA		
	EXDO-4_F		GYCTGATGAAGACRGWAGATC	532	30
	EXDO-4_R		GTCKCCRACGTARTAYTCCG		
	EXDO-5_F	Enzymes typically involved in 2,3,2'-trihydroxybiphenyl degradation (e.g. formed from dibenzofuran) (cluster XVI, EXDO-VC2)	GYTCAAGVTSAGCGAYATC	264	30
	EXDO-5_R		GTNMRBACGTAGAARCTGAAC		
	EXDO-6_F	2,3-Dihydroxybiphenyl dioxygenases of Proteobacteria and Actinobacteria (cluster XXII, EXDO-VC2)	GAYRTCATYSACATSMARATG	233	35
	EXDO-6_R		SGGAVAYC RTVTGGTCRTTG		
	EXDO-7_F ^a		GAAAAAGTGGGTTTGATGGAGG	810	30
	EXDO-7_R ^a		CGCTTATGCCKCGTCATCACCC		
	EXDO-8_F ^a		GACCAGGGSWTVGGYCACTA	430	30
	EXDO-8_R ^a		TTRTGNCCCCAGATGCTGAT		
	EXDO-9_F ^{a,b}	Proteobacterial catechol 2,3-dioxygenases (cluster XX, EXDO-VC1)	AAYCCBGABCCNTGGCCNGA	380	35
	EXDO-9_R ^{a,b}		GTYT SVCCNCBGTVADVCCRTGRCG		
Non-degenerate primer sets	EXDO-10_F	EXDOs A512, A556 and A559	CTTGACCGCCTGTTTCGAG	413	32
	EXDO-10_R		GTGACCATGAAGTTGACGTG		
	EXDO-11_F	EXDO A482	GCTGCGTTTCTACACTGACG	310	32
	EXDO-11_R		CTCCCACTCGAAACCTGATG		
	EXDO-12_F	EXDOs A420 and A547	CTCTTTGCCACGATGGATCT	342	32
	EXDO-12_R		CCATAACGTCCGAAGCATCT		
	EXDO-13_F	EXDOs A553 and A592	CATTACACCGATGTGGTTGG	396	32
	EXDO-13_R		CAGAAGGGCATGGTCAAGAT		
	EXDO-14_F	EXDO A60	GCTGGAGCTTTACACCAA	273	32
	EXDO-14_R		GCGAAGGTGAACCTTGCTCTT		
	EXDO-15_F	EXDO A414	GGAAGAGATGGCACAGAAGC	552	32
	EXDO-15_R		CCTGAGGGATTGGAGTGGTA		
	EXDO-16_F	EXDO of <i>Pseudoxanthomonas spadix</i>	AGACAGCGACCTTGACGACT	249	32
	EXDO-16_R		CACATCAACAGGCAATGGTC		
	EXDO-TolM_F	EXDO/TolM A461/A463 and A439/A441	CGTTTTGTGACGGGAGATG	2767	32
	EXDO-TolM_R		CCATCCCGGGTATTTCTCTT		
	LigB-1_F	LigB A22	AGTTGGTTTCGAGCCATTTG	603	32
	LigB-1_R		GTACCCGCTAGTCGCAACAC		
	LigB-2_F	LigB A543	TTATATCAATGCGGCTGTCTG	325	32
	LigB-2_R		TTTCAGTTTGCCCATGACAC		
	LigB-3_F	LigB A311	GACTGGGATCTGGCCTTTCT	187	32
	LigB-3_R		GATCGGCCGGTAATAACGA		

^a Primers have been described (Brennerova et al., 2009).^b Phylogenomic cluster of EXDO-VC1, where characterized enzymes typically have a preference for monocyclic substrates (Duarte et al., 2014).

Table S2. PCR amplifications using as template the fosmid DNA of 768 fosmid clones (including the first 20 sequenced clones)

clone	>30 kb assembly	Degenerate primers			Non-degenerate primers										
		EXDO-3	EXDO-4	EXDO-9	EXDO-10	EXDO-11	EXDO-12	EXDO-13	EXDO-14	EXDO-15	EXDO-16	EXDO-ToIM	LigB-1	LigB-2	LigB-3
1	no*			+							+				
2	no*			+							+				
3	no*			+							+				
4	no*			+							+				
5	yes		+												
6	yes	+								+					
7	yes								+						
8	yes			+									+		
9	yes														+
10	yes														
11	no					+									
12	no			+			+	+				+		+	
13	no			+			+	+				+			
14	no			+			+	+				+			
15	no				+										
16	no				+										
17	no				+										
18	no			+											
19	no														
20	no														
Positive amplification (out of 20)		1	1	9	3	1	3	3	1	1	4	3	1	1	1
Positive amplification (out of 768)		1	1	22	37	6	7	5	1	2	17	5	1	1	1

* 35-kb region with >99% sequence identity to the genome of *Pseudoxanthomonas spadix* BD-a59 (GenBank accession number: CP003093).

Table S3. Polycyclic aromatic hydrocarbon concentrations (mg/kg) in the soil sample used in this study.

PAH	mg/kg
naphthalene	0.15
acenaphthene	1.66
fluorene	1.32
anthracene	2.61
phenanthrene	1.36
fluoranthene	56.40
pyrene	49.90
benzo(a)anthracene	16.40
crysene	21.70
benzo(b)fluoranthene	9.16
benzo(k)fluoranthene	4.51
benzo(a)pyrene	3.73
indene(1,2,3-cd)pyrene	1.66
benzo(g,h,i)perylene	1.12