Systems biology approach to elucidation of contaminants biodegradation in complex samples- integration of high-resolution analytical and molecular tools

Caroline Gauchotte-Lindsay[1*], Thomas J. Aspray[2§], Mara Knapp[3] and Umer Z. Ijaz[1],

1.Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow G12 8QQ, UK. 2. School of Energy, Geoscience, Infrastructure and Society, Heriot-Watt University, Edinburgh EH14 4AS, UK. 3. Department of Civil and Environmental Engineering, University of Strathclyde, Glasgow G1 1XQ, UK.

* Corresponding author

§ now at: ERS Ltd, Westerhill Road, Bishopbriggs, Glasgow G64 2QH, UK.

Supplementary Information

1. Detailed Methodology

1.1. *Pressurised Liquid Extraction of Semivolatile organic compounds*

Pressurised liquid extraction (PLE) was performed using an ASE 350 system (Dionex, Camberley, UK) equipped with 10 mL the extraction solvents.

The extraction cells were lined with 2 filter papers (to ensure unwanted particulate matter did not collect in the extract) and packed with 3 g silica gel 60 (10% deactivated w/w using deionised water) to provide simultaneous sample extraction, cleanup and fractionation. The samples were air dried for 5 days and sieved to 2 mm. A portion of each soil sample (approximately 0.25 g) is pre-treated by mixing it with 0. 25g sodium sulphate and added to the extraction cell. The remaining space is filled with sand. 150 µl of the surrogate solution (deuterated analogues: Napthalene-d8, Fluorene-d10, Anthracene-d10, Fluoranthene-d10, and Chrysene-d12) at 2000µg/ml is added to the cell.

Three separate extractions were employed to sequentially extract the same cell using solvents of increasing polarity. To obtain the first fraction, hexane (50 % cell volume, 60s) was used to extract the cell. The oven temperature was maintained at 150 °C with the cells heated for 10 minutes prior to extraction. The second fraction was eluted with hexane:toluene in a 8:12 ratio (50 % cell volume, 60s). The oven temperature was maintained at 150 °C with the cells heated for 10 minutes prior to

extraction. The final fraction was extracted using toluene (70% cell volume, 30s) at 150 °C (with 10 min heating time). The first and the second fractions were collected together and concentrated to 1.15 mL prior to analysis using a Büchi Syncore® Analyst (Oldham, UK). Samples were stored in 2.5ml GC vials at -80C. 150 µl of the internal standard solution (Phenanthrene-d10) at 2000µg/ml were added to the sample vials prior to analysis.

1.2. GC-MS analysis of soil extracts

Fourteen PAHs (Naphthalene, Acenaphthylene, Fluorene, Phenanthrene, Anthracene, Fluoranthene, Pyrene, Benzo(a)anthracene, Benzo(b)fluoranthene, Benzo(k)fluoranthene,Benzo(a)pyrene, Indeno(1,2,3-cd)pyrene, Dibenzo(a,h)anthracene, Benzo(g,h,i)perylene) along with the surrogates were quantified in the samples by gas chromatography coupled with mass spectrometry (GC-MS) using a Thermo Trace GC coupled with a DSQII mass spectrometer. The gas chromatograph was fitted with a 30 m Zebon SemiVolatiles capillary column (0.25 mm ID, 0.25 µm film thickness) supplied by Phenomenex. The helium flow was kept constant at 1.4 mL/min. The initial oven temperature was set to 50 °C and held for 2 minutes before the temperature was ramped at 5 °C/min to 295 °C and then 15 ^{o}C/min to 325 $^{\circ}$C, with a final temperature hold time of 3 minutes. The split/splitless injector was used in splitless mode and its temperature was set at 280 $^{\circ}$C. One microlitre of sample was injected using a Triplus (Thermo Scientific) autosampler. Quantification was carried out in TIC or SIM mode depending on the concentrations.

1.3. GCxGC-TOFMS analysis of soil extracts

All GCxGC TOFMS analyses were performed using a LECO (St. Joseph, Michigan) time of flight mass spectrometer, model Pegasus 4D, connected to an Agilent 7890A gas chromatograph equipped with a LECO thermal modulator. The TOF ion source was fixed at 230 °C with a detector voltage of 1650 V, applied electron ionization voltage of -70 eV and a scan rate of 200 spectra/second between mass 45 and 500u. The solvent delay was 392s.

The column set comprised of a TR50-MS (30 m x 0.25 mm i.d. x 0.25 µm film thickness) supplied by Thermo as the primary column and a Rxi-5Sil (2 m x 0.25 mm i.d. x 0.25 µm film thickness) supplied by Thames Restek (Buckinghamshire, UK) as the secondary column, connected via a Thames Restek Press-tight® connector.

All extracts were analyzed with the primary oven temperature programmed at 10 °C/min from 60 °C (2 min isotherm) to 120 °C, 3 °C/min to 310 °C (10 min

isotherm). The secondary oven and modulator temperatures were maintained at a 20 °C offset relative to the primary oven. The modulation period was 6 s with a 1.3 s hot pulse time. Helium was used as the carrier gas, with a flow rate of 1.0 mL/min. The split/splitless injector was used in splitless mode and its temperature was set at 250°C. An MPS2 twister autosampler (Gerstel, GmbH & Co., Germany) was used to inject 1 μL of sample per run.

Data was processed in the LECO ChromaTOF software (Version 4.71.0.0). Baseline correction was carried out with an offset of 1 and auto smoothing. The peak finding parameters were: peak widths of 4s and 0.2s in the first and second dimension respectively, a minimum signal over noise ratio of 50 and a minimum match of 800 to combine peaks. The integration approach chosen was traditional. A classification method was employed to filter out the peaks from solvent and column bleed. Peak areas were integrated for TIC values and a maximum of 2000 peaks was set. The resulting peak tables including the top library match name, the two retention times, the quantification mass and the peak true spectrum (relative abundance to base ion between 0u and 500u) were exported as csv files for further processing.

To align peaks the R code R2DGC[11] was used. The exported csv files were first pre-processed with the PrecompressFiles function, to merge split peaks. Alignment was carried out twice using the ConsensusAlign function with the following parameters: the seed sample was chosen as the CH samples with the most peaks in its peak table, autoTuneMatchStringency was set to TRUE, similarityCutoff to 80, missingPeakFinderSimilarityLax to 0.70 and missingValueLimit to 0.1 and then 1. Alignment using two seed samples, one from COV and one from CH, was attempted but was not successful. Similarly, R2DGC offers the possibility to use retention index calculated by using compounds present in all samples but when attempted the results were less successful than without references. One output of ConsensusAlign is an alignment table (samples x compounds) that was used further for the statistical analysis.

1.4. Transition metal analysis

Soil samples were air dried in a fume hood for 72 h, after which they were disaggregated with pestle and mortar and returned to dry in the fume hood for an additional 72 h. Once dry, samples were sieved to obtain the fraction below 2 mm. A sub-portion of approximately 5 g of each sample was placed in the muffle furnace at 105 °C for 12 hours followed by another 12 h at 450 °C, and weighed to measure the

loss on ignition (LOI). A sub-portion of ~ 1g was then digested with aqua regia (10 ml of aqua regia 1:1 HCl [37%]: HNO3 [69%]) at 95 °C in a digestion block (DigiPREP Jr. ®, SCP Science) for 4 h, reduced to ~3 ml, filtered (Whatman N°42) and made to 100 ml with Milli-Q water.

Concentrations of Lead, Iron, Cadmium, Chromium , Zinc, Copper, Nickel were determined by inductively coupled plasma optical emission spectrometry (ICP-OES).

COV samples were analysed in an external laboratory (Concept Life Sciences, 16 Langlands Place, Kelvin South Business Park, East Kilbride, Scotland, G75 0YF) while remaining samples were analysed in the SUERC ICP-OES facility with iCAP 7000 ICP-OES (Thermo Fisher Scientific) with use of appropriate standards and a calibration set from 1 to 20 mg/L. Three repetitions were made per sample, and an averaged concentration was obtained. Concentrations were then adjusted using the [450 °C ash: air dried] weight ratio for each sample.

1.5. Quantitative PCR

Genomic DNA was extracted from soil samples (0.25 g fresh weight) using the PowerSoil DNA Isolation Kit (MoBIO Laboratories, Inc., USA) according to the manufacturer's protocol. Eluted DNA was quantified spectrophotometrically using a NanoDrop 2000 instrument (Thermo Scientific).

Quantitative PCR (qPCR) was carried out on an Applied Biosystems Step One instrument using the previously reported primer pairs for alkB[12] and PAH RHD GN and PAH RHD GP. The qPCR reactions were performed in duplicate for each sample with the Applied Biosystems software quality control (QC) check used to accept/reject replicate homogenicity. Reactions were performed in 20 µl volumes containing; 10 µl PerfeCTa SYBR® green ROC (Quantabio), 2 µl each of forward and reverse primer, 0.8 µl of 20 mg/ml bovine serum albumin (Roche), 3.2 µl nuclease-free water (Ambion®) and 2 µl of extracted sample DNA, standard DNA or nuclease-free water in the case of no template controls. Further quality controls included verification of melt curves and gel electrophoresis to confirm single amplification products with all three primer sets and samples.

Standards for qPCR were prepared by extracting DNA from *Pseudomonas putida* strain PG (9816, NCIMB Ltd, Aberdeen UK) and *Rhodococcus* sp. strain MJL100 (12038, NCIMB Ltd, Aberdeen, UK) following growth of monocultures under aseptic conditions using QIAamp DNA Mini Kit (Qiagen, Germany) according to the manufacturer's protocol. Amplification of single size products was verified by

standard endpoint PCR and gel electrophoresis approaches. The products were cleaned using the Wizard SV Gel and PCR Clean-Up System (Promega Corporation, USA) according with manufacturer´s protocol.

1.6. 16S metataxogenomic sequencing library preparation of soil microbiome

Genomic DNA was extracted from soil samples (0.25 g fresh weight) using the MP FastDNA® SPIN kit for Soil (MP Biomedicals, Inc., USA) according to the manufacturer's protocol. Extracted DNA was quantified using the Broad-Range Qubit Assay (LifeTechnologies) and stored at -20C until further use.

16S libraries encompassing the V3 and V4 regions were generated by Glasgow Polyomics. In brief, the V3 and V4 regions of bacterial 16S were amplified using Kapa HiFi Hotstart readymix (2x) (Kapa Biosystems, Wilmington, MA, USA) with the addition of primers specific for the V3 and V4 regions of 16S (based on the standard Illumina 16S primers), which contain an overlap sequence making the primers compatible with the Nextera XT indexing reagents (Illumina, San Diego, CA, USA). Samples were then amplified using a 5 min 95 ˚C hotstart followed by 26 cycles of 95 ˚C for 30 s and 60 ˚C for 1 minute with a final elongation step of 60 ˚C for 5 min.

The resulting amplicons were purified using bead extraction (SPRI select beads, Beckman Coulter, Brea, CA, USA), using 0.9x beads followed by 80% ethanol washes and resuspension in 20µL of 10mM Tris buffer. The amplicons were quantified using the High Sensitivity DNA Qubit system and profiles were obtained from an Agilent 2100 Bioanalyser using High Sensitivity DNA reagents (Agilent, Santa Clara, CA, USA).

Samples were then standardized to 10ng per reaction and amplified in the presence of Nextera XT v2 indexes using Kapa Hifi Hotstart readymix (2x) for 8 cycles. The resulting indexed libraries were then purified as before using SPRI select beads and quantified using the Qubit system. Final library profiles were obtained from the Agilent 2100 Bioanalyser.

The libraries were combined in equimolar ratios and sequenced on a MiSeq (Illumina, San Diego, CA, USA) instrument using a paired end, 2x300bp, sequencing run. Samples were sequenced with an average of 50,000 reads per sample.

Possible contamination of reagents was controlled by running a negative control (purified water instead of a DNA sample) through the whole analysis in conjunction with the samples.

1.7. Bioinformatics

We used VSEARCH v2.3.4 (steps documented in http://github.com/torognes/vsearch/wiki/VSEARCH-pipeline) to generate the abundance table by constructing operational taxonomic units (OTUs), a proxy for species. Prior to using VSEARCH, the paired-end reads were preprocessed. Briefly, the paired-end reads were trimmed and filtered using Sickle v1.200[13] by using a sliding window approach and trimming the reads where the average base quality drops below 20. Only the reads that were above 10 bp length were kept after trimming. Next, BayesHammer[14] was used from the Spades v2.5.0 assembler, which error-corrected the paired-end reads. Following this, pandaseqv(2.4)[15] was used to assemble the forward and reverse reads into a single sequence spanning the entire V4 region with a minimum overlap of 10 bp. The preprocessed reads (overlapped) from each sample were pooled together while barcodes were added to keep track of which sample the read originated from. The reads were then dereplicated, sorted in order of decreasing abundance and singletons were discarded. Next, the reads were clustered based on 97% similarity followed by a removal of clusters which had chimeric models built from more abundant reads (--uchime_denovo option in vsearch). To remove any chimeras that may have been missed, particularly in the case that they had parents that were absent from the reads or were present in very low abundance, a reference-based chimera filtering step (--uchime_ref option in vsearch) using a gold database (https://www.mothur.org/w/images/f/f1/Silva.gold.bacteria.zip) was applied. Finally, the OTU table was generated by matching the original barcoded reads against clean OTUs (a total of 2,234 OTUs for n=26 samples) at 97% similarity (a proxy for species-level separation) with summary read statistics for samples as follows: [1st Quantile: 45,116, Median: 49,398, Mean: 52,557, 3rd Quantile: 65,156, Max: 95,350]. The assign_taxonomy.py script from the Qiime workflow[16] was used to taxonomically classify the representative OTUs against the SILVA SSU Ref NR database release v123 database. After, the OTUs were multisequence aligned using MAFFT v 7.3[17] and were used in FastTree v2.1.7[18] to generate the phylogenetic tree in NEWICK format. The biom file for the OTUs was then generated by combining the abundance table with taxonomy information using make_otu_table.py from the Qiime workflow. All prokaryotic KEGG organisms are available in Tax4Fun[19] for SILVA v123 and KEGG database release 64.0. In Tax4Fun, the ultrafast protein classification (UProC) tool[20] was used to generate metabolic functional profiles after

the data was normalised for 16S rRNA gene copy numbers. Through Tax4Fun, we recovered: 6,599 KEGG orthologs (enzymes) using fctProfiling=TRUE in Tax4Fun() function; and 284 KEGG pathways according to the MoP-Pro approach using fctProfiling=FALSE in Tax4Fun() function. Although the Tax4Fun-based metabolic predictions are limited by the taxa available in the reference database, it provides a statistic called fraction-of-taxonomic-units-unexplained (FTU), which reflects the quantity of sequences that are assigned to a taxonomic unit but are not transferable to KEGG reference organisms.

Table S1- List of compounds that were found in all 68 peak tables. Except for parent PAHs and deuterated surrogates that were matched with standards, identifications were tentative using matches to the NIST library. * indicates match scores lower than 800. RT1= retention time in the first dimension and RT2= retention time in the second dimension.

Compound identification	RT1,RT2 (s,s)	Compound identification	RT1,RT2 (s,s)
C3 Benzene a	410 , 1.855	1-Methyldecylbenzene	1400 , 3.485
C4 Benzene a	422 , 1.890	1-Pentylheptylbenzene	1424 , 3.560
C3 Benzene b	446 , 1.855	1-Butyloctylbenzene	1436 , 3.555
Phenol	452 , 1.665	2,2-Diphenylpropane	1442 , 2.635
Benzaldehyde	458 , 1.670	Benzylxylene*	1472 , 2.670
C3 Benzene (1u)	482 , 1.810	1-Ethyldecylbenzene	1508 , 3.605
C4 Benzene b	494 , 1.905	Fluorene	1532 , 2.630
C3 Benzene (2u)	506 , 1.780	Diethyl Phthalate	1556 , 2.395
C4 Benzene (1u)	518 , 1.915	1-Methylundecylbenzene	1598 , 3.590
C1 Phenol	548 , 1.795	1-Pentyloctylbenzene	1610 , 3.660
Acetophenone	572 , 1.810	1-Butylnonylbenzene	1628 , 3.655
D8-Naphthalene	722 , 2.115	GCxGCMS_569	1958 , 2.615
Naphathlene	728 , 2.115	D10-Phenanthrene	2012 , 2.630
Propanoic acid, 2-methyl-, 2,2-dimethyl-1-(2-hydroxy-1-methylethyl)propyl ester	818 , 2.670	GCxGCMS_608	2048 , 2.565
Propanoic acid, 2-methyl-, 3-hydroxy-2,4,4-trimethylpentyl ester	854 , 2.705	4H-Cyclopenta[def]phenanthrene	2300 , 2.675
2-Methylnaphthalene	884 , 2.370	Fluoranthene	2600 , 2.685
1-Methylnaphthalene	932 , 2.370	Cyclopenta(def)phenanthrenone	2612 , 2.580
1-Butylhexylbenzene	1070 , 3.225	GCxGCMS_788	2690 , 2.675
2,4-Di-tert-butylphenol	1154 , 2.795	Pheleno[1,9-bc]thiophene	2708 , 2.630
Propanoic acid, 2-methyl-, 1-(1,1-dimethylethyl)-2-methyl-1,3-propanediyl ester	1196 , 3.245	Pyrene	2732 , 2.660
1-Pentylhexylbenzene	1238 , 3.430	Methylpyrene a	2846 , 2.725
1-Butylheptylbenzene	1250 , 3.415	Methylpyrene b*	2912 , 2.705
1-Propyloctylbenzene	1268 , 3.445	Methylpyrene c	2966 , 2.715
Biphenylene	1274 , 2.480	Benzo[ghi]fluoranthene	3218 , 2.705
1,1-Diphenylethane	1274 , 2.560	Benz[a]anthracene a	3308 , 2.705
1-Ethylnonylbenzene	1316 , 3.470	D12-Chrysene	3332 , 2.685
Acenaphthene	1322 , 2.545	Benz[a]anthracene b	3344 , 2.680
Benzyltoluene*	1364 , 2.595	Methylchrysene	3470 , 2.775
Dibenzofuran	1376 , 2.605	Perylene	3962 , 2.725

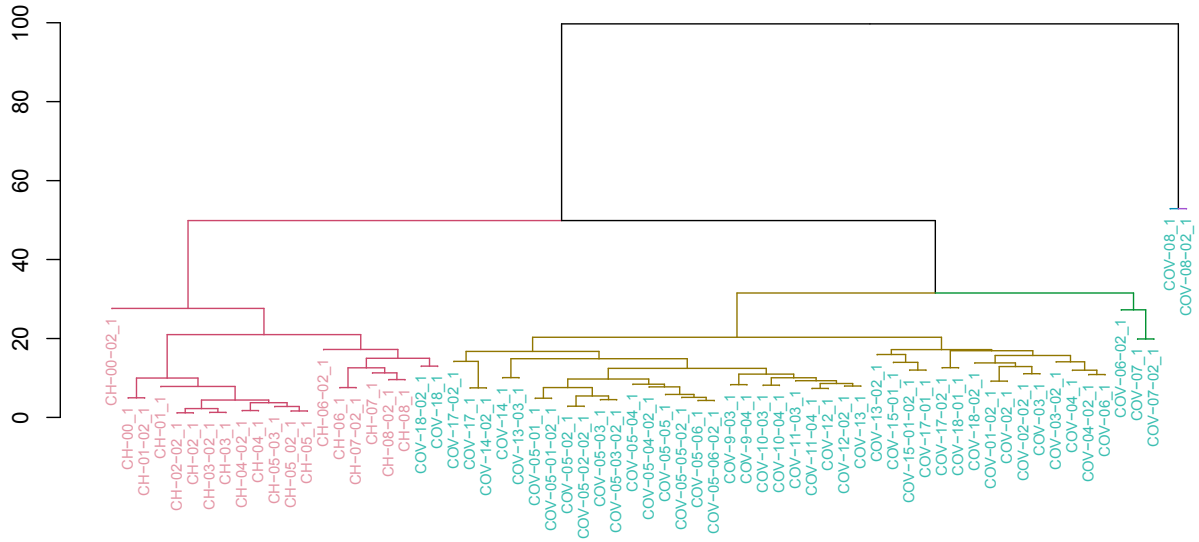

Figure S1- Hierarchical clustering of the GCxGC samples based on the 961 compounds selected during alignment of the data. XX-NN_1 and XX-NN-02_1 signify two instrumental replicates (XX=site, NN=sample number). COV-05 was extracted by PLE six times, COV-05-0N thus describes the Nth extraction of the same soil sample.