SUPPORTING INFORMATION

SUPPLEMENTARY METHODS

Microtitre Plate Assay for GFP Expression

To measure GFP fluorescence in microtitre plates, a FLUOStar Optima (BMG Labtech) fluorescent plate reader was fitted with filters for excitation at 485 nm and emission 520 nm. The gain was set at 750. Clear microtitre plates with 96 (Corning) or 384 (Fisher) wells were filled with appropriate media (200 µL or 100 µL, respectively) supplemented with the appropriate antibiotic or inducing compound where required. 96 well plates were inoculated by 1:100 dilution from a starter culture, and 384 well plates were inoculated using a pin replicator (V&P Scientific) from a starter culture. The starter culture was comprised of log-phase liquid culture cells grown in LB for 18 hours at 37 ºC in either a 96-well or 384-well plate.

Chromosomal Knockout of the *spo0A* Gene in Bacillus 6A5 to Make an Asporulant Strain

PCR primers containing restriction sites for BamHI and HindIII were designed to target the middle 374 bp of the 795 bp *spo0A* gene (GI:30022254) from *Bacillus* 6A5 (a.k.a. ATCC 14579). Phusion DNA polymerase (New England Biolabs) was used to amplify the gene according to manufacturer instructions. The PCR fragment was cloned into the corresponding sites of the vector pMUTIN4 and transformed into E. coli DH10b; pMUTIN4/*spo0A* was isolated from E. coli and transformed into 6A5. Note that since pMUTIN4 has an Em^R gene, but does not contain an origin of replication for *Bacillus*, the homologous *spo0A* gene allowed integration of the plasmid – thereby interrupting the targeted gene – and confers resistance to Em. Spore stains were viewed using a Zeiss Axio Imager M1 and photos were captured using the AxioCam

MRm and its corresponding image capture software. Transformation of *Bacillus* 6A5 with pMUTIN4 vector carrying the *spo0A* gene gave rise to colonies resistant to Em.

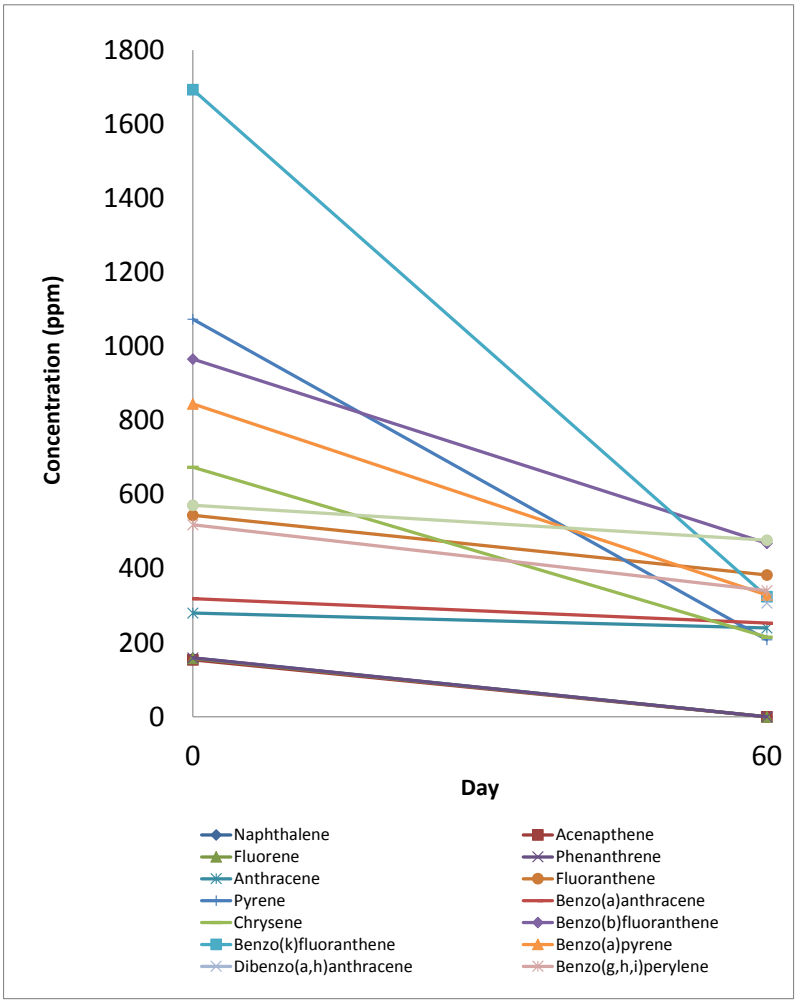
Several attempts were made to recover inducible genes from the *Bacillus* library; however, after flow cytometric sorting, little to no enrichment for the selected populations was seen (Supplementary Figure S6).

Identification of Bacterial Populations and Sorting Based on GFP Fluorescence

The FACSAria was calibrated for quality control on each use by running cytometer setup and tracking (CST) software using CST beads (BD Biosciences; lot # 69922). Setup for sorting was performed using BD Accudrop beads (BD Biosciences; lot # 73991) and the drop delay was determined automatically using the Accudrop experiment template. Data was acquired and analyzed using FACSDiva software (BD Biosciences).

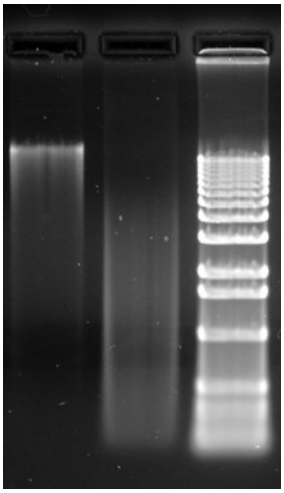
The FSC-A, FSC-W, FSC-H, SSC-A, SSC-W, and SSC-H parameters were used to identify those events that contained a single bacterial cell. The voltages on the FSC photodiode and the SSC photomultiplier were adjusted until a dense cluster of events appeared. The FITC detector was a photomultiplier tube adjusted to a voltage of 600 such that the promoterless GFP containing plasmid (pMMeb) showed basal levels of fluorescence while a constitutively expressed GFP (pMMpos1) was at the upper range of detection. The FSC/SSC width and height were used for doublet gating. Sorts were performed using single-cell mode, keeping the number of events per second ~3000 by adjusting cell density rather than flow rate; flow rate was kept at 1.0 to ensure optimal resolution of populations.

Degradation of PAHs in Soil Bioslurry

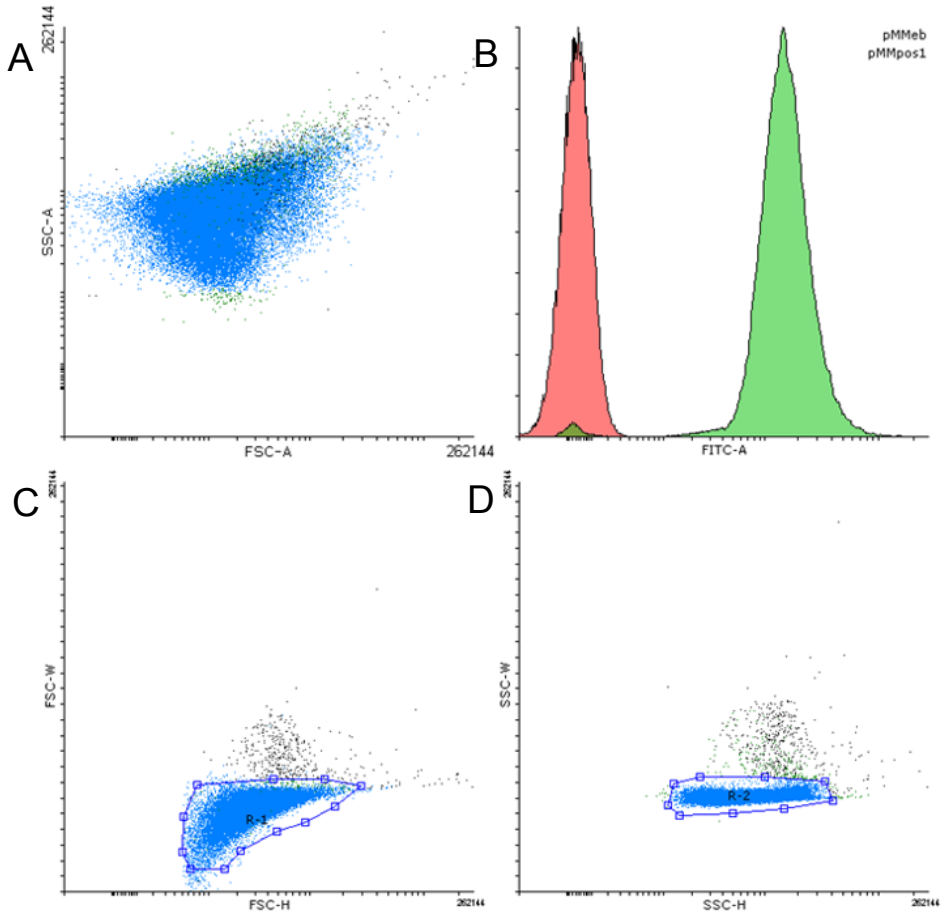


Supplementary Figure S1. Concentrations for Naphthalene, Acenapthene, Fluorene, Phenanthrene, Anthracene, Fluoranthene, Pyrene, Benzo(a)anthracene, Chrysene, Benzo(b)fluoranthene, Benzo(k)fluoranthene, Benzo(a)pyrene, Dibenzo(a,h)anthracene, Benzo(g,h,i)perylene, and Indeno(1,2,3-cd)pyrene measured by HPLC in duplicate bioreactors (average of two values are shown). These concentrations show a significant decrease after 60 days.

Restriction Digestion of Metagenomic DNA

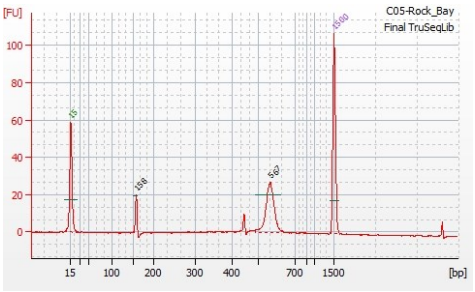


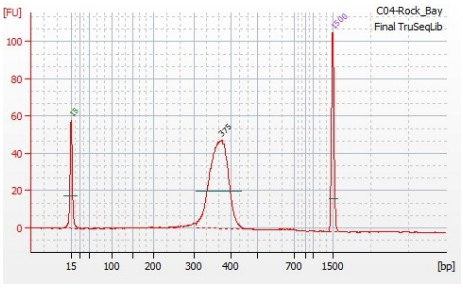
Supplementary Figure S2. Restriction digestion of metagenomic DNA for library preparation. Lane 1, undigested soil DNA from Rock Bay contaminated soil. Lane 2, Sau3A1 digest of metagenomic DNA. Lane 3, 1 kb ladder.



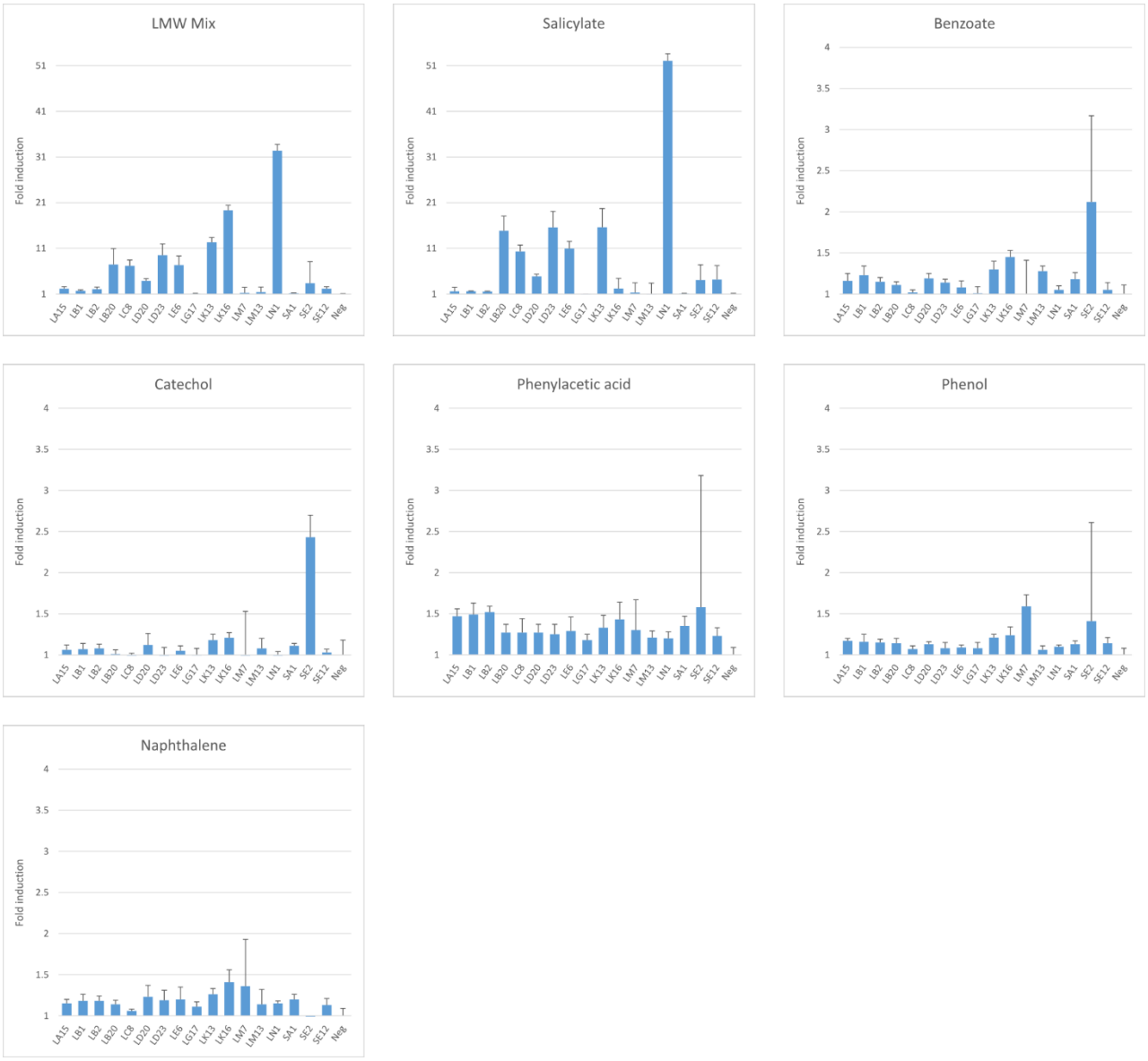
Supplementary Figure S3. Flow cytometric calibration for single-cell sorting of bacteria. Blue dots show identified bacterial populations. Cells outside of this region were excluded from analysis via gating (shown as lines surrounding the populations). A) Dot plot showing FSC and SSC clustering of a population of *E. coli*. B) Histogram plots of GFP expression for pMMeb (promoterless GFP) and pMMpos1 (constitutive GFP) control plasmids. C) Doublet gating on FSC channel. D) Doublet gating on SSC channel.

Supplementary Figure S4. DNA fragmentation data from the production of the A) 447 bp and B) 255 bp (right) TruSeq gDNA libraries from Génome Canada. Insert sizes were calculated by removing the length of the adapter sequences.



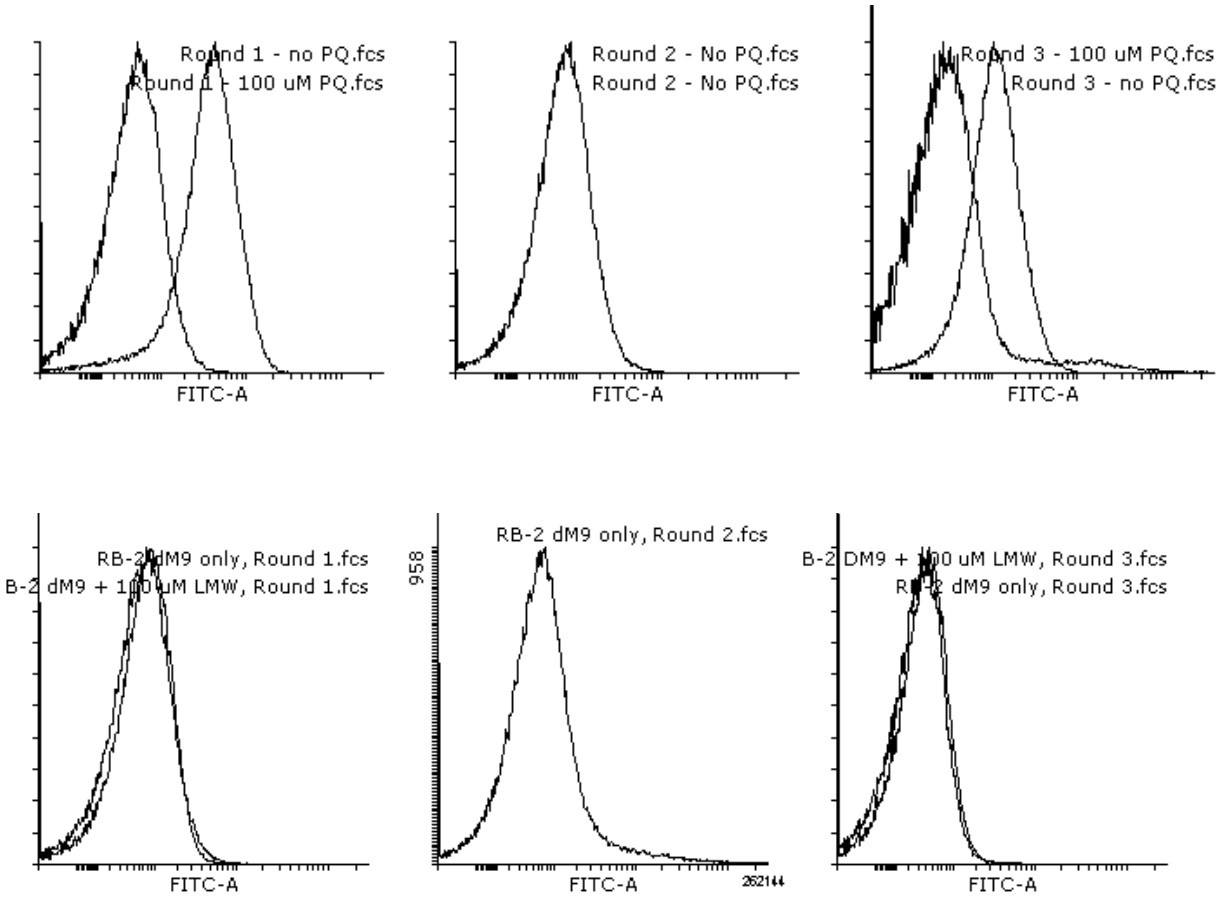


Induction of Aromatic-Inducible Clones by Individual Compounds in Microtitre Plates

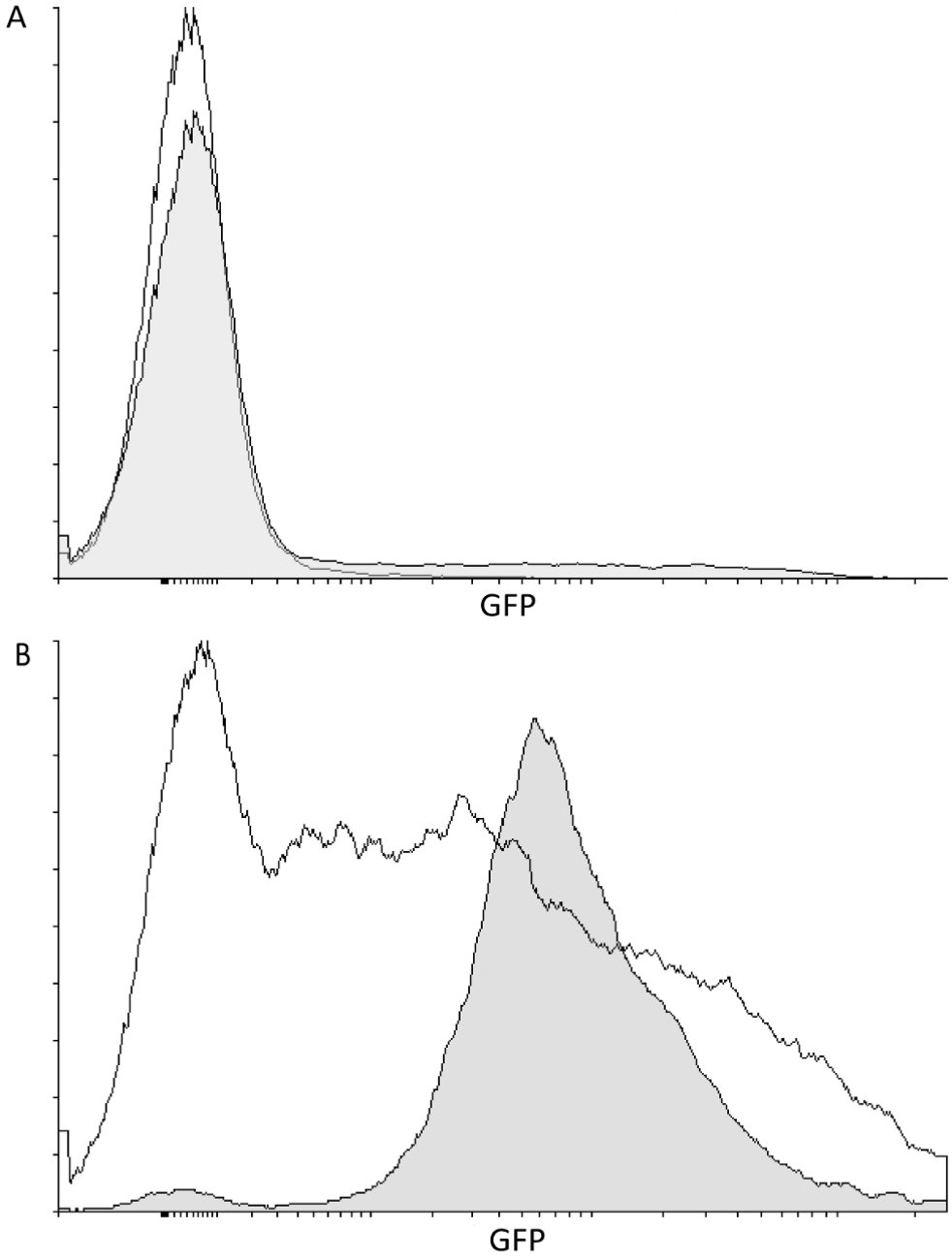


Supplementary Figure S5. Induction of aromatic-inducible clones in microtitre plates using a variety of LMW aromatic compounds. Clones are inducible to different extents relative to one another for each chemical tested here; salicylate was by far the most potent inducer. The y-axes

are scaled to a height dependant on the level of induction, and normalized relative to the negative control (empty vector). Error bars indicate standard deviation of four biological replicates.



Supplementary Figure S6. Sequential rounds of FCM sorting (columns 1-3) on the 6A5 library produce no PQ (top) or LMW-aromatic (bottom) inducible clones. Contrast this with the LMW-inducible clones shown in Fig. 3 for *E. coli* sorting. Middle panels (column 2) should be highly fluorescent but do not show this characteristic. As shown in columns 1 and 3 (A), paraquat drastically increases the auto-fluorescence of *Bacillus*.



Supplementary Figure S7. Patterns of expression in four different clonal populations of GFP expressing cells with similar mean fluorescence values. Single-cell measurements of reporter genes may be advantageous in determining changes in expression, since typical population mean-based measurements (e.g., qPCR) would not be sensitive to sub-population differences

such as those demonstrated here. A) a clone with a typical bell-shaped expression pattern (white, mean=80, median=39) overlaid with a clone expressing GFP with a tail-end, skewed fluorescence (light grey, mean=2,158, median=59) B) a clone with a highly fluorescent but relatively compact expression pattern (light grey, mean=16,114, median=7,309) overlaid with a clone with a highly variable expression pattern that spans the entire range of measureable values (white, mean=16,204, median=784).

Supplementary Table S1. Sequence analysis of inducible genes recovered using the SIGEX protocol from the EXP-1 metagenomic control library and genomic *E. coli* EC-C600 control library. Induction was measured by flow cytometry unless otherwise indicated.

Inducer	Library	Clone	Fold Induction	Homologous Genes/Proteins (Accession)	% A.A. Identity	Homologous Organism	Taxonomy (Phylum: Order)	Comments
IPTG	EC-C600	IPTGH4	5.42*	ND	ND	ND	ND	Cleaves Xgal only in presence of IPTG
		IPTGH8	6.69*	beta-galactosidase (CP002890) transcriptionally fused to GFP	97	*E. coli* UMNF18	Proteobacteria: Gammaproteobacteria	Cleaves Xgal only in presence of IPTG
		IPTGH10	10.19*	beta galactosidase small chain lacZ (CP002729) transcriptionally fused to GFP	100	*E. coli* UMNK88	Proteobacteria: Gammaproteobacteria	Cleaves Xgal only in presence of IPTG
PQ	EC-C600	PQD5	5.88	ND	ND	ND	ND	ND
		PQD9	4.01	pqiB (YP_003228134.1), transcriptionally fused to GFP	98	*E. coli* O26:H11 str. 11368	Proteobacteria: Gammaproteobacteria	Paraquat-inducible protein B; known to be upregulated by superoxide stress (ref)
		PQC1	3.40	ND	ND	ND	ND	ND
		PQE1	3.29	ND	ND	ND	ND	ND
		PQE2	2.79	ND	ND	ND	ND	ND
	EXP-1 Meta-genome	PQG11	70.24	peptidase M16 domain-containing protein (ACU58203.1), transcriptionally fused to GFP, pqqL domain	49	*Chitinophaga pinensis* DSM 2588	Proteobacteria: Deltaproteobacteria	Inducibility abolished in SoxRS-*E. coli*; pqqL known to be upregulated during *E. coli* superoxide stress (ref)
		PQA20	92.15	hypothetical protein MXAN_1454 (ABF93145.1), transcriptionally fused to GFP	38	*Myxococcus xanthus* DK 1622	Proteobacteria: Deltaproteobacteria	Inducibility abolished in SoxRS-*E. coli*; BLAST results indicate that MXAN_1454 shares weak homology with an ABC-type transport system involved in Fe-S cluster assembly (CBK97053.1 from *Eubacterium siraeum* DSM 15702), which indicates a role in oxidative stress response
		PQM17	29.35	hypothetical protein Npun_F1891 (ACC80546.1), transcriptionally fused to GFP	29	*Nostoc punctiforme* PCC 73102	Cyanobacteria: Nostocales	Inducibility abolished in SoxRS-*E. coli*; Unable to determine a function for homologous genes
		PQA4	30.40	Fjo21 (AEI68364.1) upstream of and toward GFP	52	*Myxococcus fulvus* HW-1	Proteobacteria: Deltaproteobacteria	Inducibility abolished in SoxRS-*E. coli*; intergenic region of 333 bp follows Fjo21 before GFP starts. Fjo21 has no known

						function but is part of the COG4270 superfamily (predicted membrane protein). Other close homologs to the A4 clone include DoxX family proteins, an obscure putative oxidoreductase (ref)
	PQG8	33.66	oxidoreductase molybdopterin binding protein (GI:269785296)	33	*Sphaerobacter thermophilus* DSM 20745	Chloroflexi: Sphaerobacteridae	Inducibility abolished in SoxRS-*E. coli*; In addition to the oxidoreductase, this clone shares weak homology with a putative glutathione S-transferase from *Acaryochloris marina* MBIC11017 (plasmid pREB1).
	PQG13	55.11	2-deoxy-D-gluconate 3-dehydrogenase (EEQ93510.1), transcribed divergently from GFP	58	*Ochrobactrum intermedium* LMG 3301	Proteobacteria: Alphaproteobacteria	Two-fold inducible in SoxRS-*E. coli*
	PQK19	42.97	Npun_F1891 (ACC80546.1), transcriptionally fused to GFP	30	*Nostoc punctiforme* PCC 73102	Cyanobacteria: Nostocales	Nearly entire ORF for Npun_F1891 is present; this gene is homologous mainly to proteins of unknown function, and has no conserved domains. Five rounds of PSI BLAST indicate that its closest relatives are mechanosensitive ion channel proteins.
	PQM5	116.72	40 bp intergenic region before GFP, followed by AraC family transcriptional regulator (AEE53667.1), divergently transcribed from GFP; upstream of this is a phosphodiesterase	36	*Haliscomenobacter hydrossis* DSM 1100	Bacteroidetes: Sphingobacteria	Divergently transcribed ORF contains a PRK10572 conserved domain (DNA-binding transcriptional regulator AraC); in *Haliscomenobacter hydrossis*, the gene expected to be transcribed where GFP is fused is a hypothetical protein (YP-004450541.1), and the biological function could not be easily inferred.
	PQL10	8.99	signal peptide protein thioredoxin (YP_411745), ORF divergent from GFP; TonB-like protein upstream from thioredoxin (ABB74352.1)	66	*Nitrosospira multiformis* ATCC 25196	Proteobacteria: Betaproteobacteria	First ORF contains a TlpA-like family thioredoxin domain (part of the thioredoxin superfamily); second ORF contains a Gram-negative bacterial tonB protein domain.
	PQI8	10.09	PAS/PAC sensor-containing diguanylate cyclase /phosphodiesterase (GI: 253997601), transcriptionally fused to GFP	51	*Methylotenera mobilis* JLW8	Proteobacteria: Betaproteobacteria	PAS/PAC sensor is a membrane bound protein that likely senses oxygen-related stress (ref)
	PQD7	18.20	putative ATP-dependent RNA helicase (YP-001658093.1),	29	*Microcystis aeruginosa* NIES-843	Cyanobacteria: Chroococcales	ATP-dependent RNA helicase is similar to CRISPR-associated Cas3 helicase; the upstream hypothetical protein, 4 rounds of

				transcriptionally fused to GFP; upstream, transcribed in the same direction, is hypothetical protein MAE_30780 (YP-001658092.1)				PSI BLAST, appears to be related to putative tranposases.
MERC	EXP-1	PQK9	313.77	merR (divergent) and merT (transcriptionally fused to GFP)	99	*Nitrosomonas europaea*	Proteobacteria: Betaproteobacteria	Inducible by mercury, but discovered by paraquat induction where paraquat contained small amounts of Hg^{2+}. Potential source for novel bioreporter.

* measured by FLUOStar Optima

Supplementary Table S2. Statistical characteristics of various *de novo* assemblies of Illumina-sequenced metagenomic DNA from the Rock Bay PAH-contaminated site. Approximately 125 billion bases of DNA sequence was used to generate these assemblies.

Data Source	**Assembler and Parameters**	**Starting Bases (Gbp)**	**n50 (bp)**	**Longest Contig (kb)**	**Total Assembled Sequence in bp (# contigs)**
All NGS Reads	CLC Assembly Cell, kmer 29	125	3 374	378	703 857 562 (367 294)
	CLC Assembly Cell, kmer 31	125	3 339	398	695 343 104 (364 224)
	IDBA-UD, kmer 20 to100, step size 10, %ID 95	125	5 578 (8 704 scaffolded)	608 (same scaffolded)	406 973 927 (117 116)
	IDBA-UD, kmer 20 to 100, step size 1, %ID 99	125	6 638 (9 121 scaffolded)	608 (same scaffolded)	407 472 907 (114 887)
Subsystem "Metabolism of Aromatic Compounds" annotated by MG-RAST	IDBA-UD, kmer 20-100, step size 1, %ID 95	1.5	1 288 (1 578 scaffolded)	6.1 (9.4 scaffolded)	2 093 243 (1 705)

Supplementary Table S3. Statistical measures of contigs obtained from PRICE directed assemblies. PRICE is used to direct the *de novo* assembly of a metagenome to the regions surrounding a desired subset of initial seed contigs.

Data Source for Seed Contigs	**# Initial Sequences (total bases, n50)**	**Cycles in PRICE**	**# Final Contigs (n50)**	**Longest Contig (kb)**	**Total Bases (Mbp)**
IDBA-UD Assembled "Metabolism of Aromatic Compounds" Reads from MG-RAST (Table 1)	1 835 (2 128 001, 1 269)	20	1 649 (9 587)	63	13.7
SIGEX-recovered Clones Inducible by Aromatic Compounds (Sanger-sequenced)	40 (44 365, 1 135)	20	3 773 (4 662)	33	15.2

Supplementary Table S4. Similarity-based matches between SIGEX-recovered clone sequences and contigs assembled *de novo* using various approaches from NGS data. ND indicates that no match was found.

Clone	IDBA-UD Assembled Contigs		PRICE Assembly of SIGEX Clones		PRICE Assembly of MG-RAST Annotated Aromatic-Metabolic Genes	
	Best Match[1] to contig (length, kb)	Reference-Mapped Pairwise %ID	Best Match[1] to contig (length, kb)	Reference-Mapped Pairwise %ID	Best Match[1] to contig (length, kb)	Reference-Mapped Pairwise %ID
LA15	contig_243 (61.1)	99.0	contig_21 (15.8)	98.5	contig_13 (27.1)	98.5
LB1	contig_243 (61.1)	98.7	contig_137 (9.9)	99.9	contig_13 (27.1)	98.7
LB18 (GfpSeq)	contig_243 (61.1)	100.0	contig_137 (9.9)	100.0	contig_13 (27.1)	100.0
LB18 (pADLeft)	contig_243 (61.1)	91.2	contig_21 (15.8)	90.5	contig_13 (27.1)	91.3
LB2	contig_243 (61.1)	97.9	contig_137 (9.9)	97.3	contig_13 (27.1)	97.9
LB20 (GfpSeq)	contig_14785 (5.5)	97.9	ND	ND	ND	ND
LB20 (pADLeft)	contig_243 (61.1)	94.6	contig_21 (15.8)	94.9	contig_13 (27.1)	94.4
LC17 (GfpSeq)	contig_23284 (6.6)	87.2	**contig_626 (5.6)**	**94.6**	ND	ND
LC17 (pADLeft)	contig_23284 (6.6)	88.1	contig_626 (5.6)	90.9	ND	ND
LD20	contig_23284 (6.6)	86.8	contig_626 (5.6)	87.3	contig_52 (18.4)	75.1

LD23	contig_6160 (18.8)	99.0	**contig_1 (33.2)**	**98.1**	contig_52 (18.4)	98.4
LE6	contig_23284 (6.6)	93.4	contig_626 (5.6)	97.3	contig_52 (18.4)	69.7
LG17	contig_3148 (52.1)	97.0	contig_275 (8.5)	89.1		
LK13	contig_23284 (6.6)	97.2	contig_626 (5.6)	94.2	contig_52 (18.4)	69.7
LK16	contig_23284 (6.6)	99.0	ND	ND	ND	ND
LK9 (GfpSeq)	contig_58390 (1.3)	100.0	**contig_231 (8.8)**	**99.3**	contig_819 (8.4)	93.0
LK9 (pADLeft)	contig_66283 (1.1)	99.6	contig_2326 (2.9)	89.9	contig_819 (8.4)	54.9
LM13	contig_23284 (6.6)	96.0	contig_626 (5.6)	92.8	contig_52 (18.4)	68.1
LM7 (GfpSeq)	contig_18132 (9.9)	96.7	contig_73 (12.3)	93.7	ND	ND
LM7 (pADLeft)	contig_18132 (9.9)	85.7	contig_73 (12.3)	88.2	ND	ND
LN1 (GfpSeq)	contig_23284 (6.6)	94.9	contig_626 (5.6)	96.6	contig_52 (18.4)	72.2
LN1 (pALeft)	contig_5976 (10.6)	86.4	contig_626 (5.6)	82.8	ND	ND
NA1 (GfpSeq)	contig_33223 (2.6)	95.1	**contig_68 (12.6)**	**94.3**	**contig_973 (5.9)**	**95.6**
SA1 (GfpSeq)	contig_6160 (18.8)	95.0	contig_11 (16.8)	96.3	contig_52 (18.4)	95.0
SE2	contig_3075 (17.4)	97.9	contig_6 (19.7)	97.9	ND	ND

(GfpSeq)						
SE12 (GfpSeq)	contig_9794 (17.6)	99.3	contig_46 (13.6)	92.8	ND	ND
Average	18.4	95.0	13.3	94.08	14.5	85.8

[1] SIGEX clones were reference mapped to contigs using Geneious v. 6.1