

Metagenomic Analysis of Contaminated Soils

By

Matthew J. Meier, B.Sc.

A thesis submitted to the Faculty of Graduate and Postdoctoral Affairs in partial fulfillment of the requirements for the degree of

Doctor of Philosophy

in

Biology

Ottawa-Carleton Institute of Biology
Carleton University
Ottawa, Ontario

© 2014, Matthew J. Meier

Abstract

Several metagenomic techniques were applied to analyze the microbial communities found in contaminated sites. A high-throughput metagenomic method called substrate-induced gene expression (SIGEX) was used to screen for genes that were upregulated by xenobiotic pollutants within a metagenomic library. SIGEX uses a promoterless green fluorescent protein (GFP) as a reporter for the transcription of metagenomic DNA that has been cloned in a plasmid library. Through propagation of metagenomic libraries in liquid culture, we used flow cytometry and fluorescence-activated cell-sorting to sort and analyze rare clones with desired expression characteristics. Using microbial DNA isolated from a polycyclic aromatic-contaminated site, clones that were inducible by a variety of aromatic hydrocarbons were recovered using SIGEX. These inducible elements were examined for sequence similarity to known genes, and were found to contain different types of aromatic-metabolizing oxygenases and efflux pumps, along with their respective regulatory genes. Most often, the sequences were those of partial operons containing various *nahG* (salicylate oxygenase) family genes, along with their respective upstream *nahR* regulators. Next-generation sequencing was used to map these small plasmid-based clones (representing relatively small fragments of the metagenome, on the order of 1 to 5 kilobases) to larger contigs (up to 61 kilobases) derived from the *de novo* assembly of 125 gigabases of shotgun-sequenced metagenomic DNA. These contigs were annotated using a variety of gene and protein prediction tools and were found to contain entire operons related

to aromatic hydrocarbon metabolism; this enabled a more complete functional and taxonomic assignment of the sequences recovered through SIGEX analysis. To assess the presence of relevant gene classes that were not recovered with SIGEX, genes with the capacity for xenobiotic metabolism were screened *in silico* within the assembled metagenomic reads using the biodegradation gene database and MG-RAST annotation. The taxonomic relationships between these functional genes were evaluated. As a direct application of this work, we show that SIGEX can be used to aid in the discovery and design of novel whole-cell bioreporters for the detection of xenobiotics.

Acknowledgements

The research presented in this thesis was possible only because of the incredible help I have received throughout the course of my studies.

First and foremost, I would like to thank my supervisor - Dr. Iain Lambert - who has taught me more about doing science the correct way than any textbook or journal article ever could. His enthusiasm and curiosity are a huge asset to all the students he mentors, and his support over the years (in many capacities) has been tremendous.

Within the Carleton and Ottawa Biology faculty there are many individuals I would like to thank who have provided assistance (in the form of teaching, advice, access to equipment/chemicals, and occasionally beer), including Drs. Myron Smith, Alexandre Poulain, Paul White, Ashkan Golshani, Susan Aitken, Bill Willmore, John Vierula, James Cheetham, Alex Wong, Mike Wade, and Patrice Smith. A special thanks to Dr. Suzanne Paterson for her never-ending help in the lab, and the invaluable hands-on experience she has given. I would also like to thank Francina Jackson, Caroline Rose, David Coffey, and Kevan Benn for their company and assistance in the lab (especially Francina for many useful discussions). Furthermore, I thank Biology staff members Tanya Rudd (without whom I would not have had any research materials), Ed Bruggink (for ensuring that the building was standing every day), Joan Mallett, and Glen Kit (with both of whom it has been fun teaching the young ones). I also express my gratitude to

fellow grad students (Caitlin Ritz, Christine Lemieux, Melanie Charlebois, Jen Skanes, Ally Jaworski, Dominique Morneau, Mohsen Hooshyar, Andrew Robinette, Ian Pulsifer, Matt Jessulat, Bahram Samanfar, Edgar Abouassaf, Katayoun Omid, Pratik Lodha, and others), with whom I have shared many memorable experiences. I am grateful to Carleton University and the Natural Sciences and Engineering Research Council of Canada for the funding received toward this project.

I have the fortune of having too many friends to name here; they have helped me remain a (somewhat) socialized person through the course of my PhD; however, those who have contributed significantly to my philosophy on science include: Dr. John Howat, Stephen Poirier, John Stewart, Nicholas Fraser, Elise Vist, Bianca Howat, Tina Delorme, Peter Rehbein, Scott Morrison, George Vladislavljevic, and Dr. Richard Webster.

I am grateful to my parents, Linda and Rolf Meier, for their enduring support and encouragement – and my father in particular, for teaching me about science and nature, for as far back as I can recall, and up to the present day. They have provided me with both the opportunity and the tools to make science a part of my daily life.

Finally, I thank Melissa Dennis – my beautiful lifelong companion – who has given me inspiration, love, and shown limitless patience throughout all of my graduate studies.

Table of Contents

Abstract.....	ii
Acknowledgements.....	iv
Table of Contents.....	vi
List of Tables.....	xii
List of Figures	xiii
List of Appendices.....	xviii
List of Abbreviations.....	xix
Statement of Contribution.....	xxi

Chapter 1. Bacterial Responses to Pollutants in the Context of Metagenomics

1.1 Metagenomics: Access to Uncultured Microbes in the Environment	1
1.2 Bacteria and the Fate of Xenobiotics	3
1.3 Bacterial Gene Expression in the Environment.....	4
1.4 Molecular Methods in Metagenomics.....	5
1.4.1 Functional/Phenotypic Screens.....	7
1.4.2 SIP	8
1.4.3 RFLP, DGGE, and TGGE.....	9
1.4.4 Probe-based Technologies.....	10
1.4.5 Gene and Protein Expression Analysis	10
1.4.6 SIGEX	11
1.4.7 NGS	15
1.4.8 Single-Cell Analysis	16

1.5 Hypotheses and Objectives of this Thesis	17
--	----

Chapter 2. Genetic Regulation of Aromatic Metabolism

2.1 Introduction.....	24
2.1.1 AHs and Biodegradation	24
2.1.2 Mechanisms of Aerobic Biodegradation of Aromatic Compounds.....	28
2.2 Bioavailability and uptake of AHs	33
2.3 Regulation of Genes Involved in LMW AH Catabolism.....	34
2.4 Regulation of Genes Involved in HMW PAH Catabolism.....	35
2.4.1 Regulation in <i>Mycobacterium vanbaalenii</i> PYR-1	36
2.4.2 Regulation in <i>Mycobacterium</i> spp.....	49
2.4.3 Regulation in Sphingomonads	58
2.4.4 Regulation in Pseudomonads.....	67
2.5 Community-based Approaches to Characterize PAH-Degrading Genes	73
2.6 Other Genetic Factors Influencing PAH Degradation	74
2.7 Conclusions.....	76

Chapter 3. Exploration of an Aromatic Hydrocarbon Contaminated Soil Metagenome using Substrate-Induced Gene Expression

3.1 Introduction.....	80
3.2 Methods.....	83
3.2.1 Plasmids, Strains and Growth of Bacteria.....	83
3.2.2 DNA Manipulations and Molecular Methods.....	85
3.2.3 Soil Samples and Treatments.....	85
3.2.4 Metagenomic DNA Isolation	86
3.2.5 Preparation of Vector DNA for Library Creation.....	86
3.2.6 Preparation of Metagenomic Insert DNA	87
3.2.7 DNA Purification by Gene Cleaning	90
3.2.8 Phenol/Chloroform Extraction of DNA.....	90
3.2.9 EtOH Precipitation	91
3.2.10 Plasmid Library Construction.....	91

3.2.11 Transformation of Plasmid Libraries into Hosts.....	91
3.2.12 Chromosomal Knockout of the <i>spo0A</i> Gene in <i>Bacillus</i> 6A5 to Make an Asporulant Strain	92
3.2.13 Libraries Constructed for this Thesis	93
3.2.14 Flow Cytometry (FCM) Analysis	95
3.2.15 Identification of Bacterial Populations and Sorting Based on GFP Fluorescence	95
3.2.16 SIGEX Induction Protocol.....	99
3.2.17 DNA Sequence Analysis	99
3.3 Results	101
3.3.1 Library Construction	101
3.3.2 Proof of Principle: Induction of EXP-1 Metagenomic Library with Paraquat and <i>E. coli</i> Genomic Library with Paraquat and IPTG	101
3.3.3 Induction of PAH-E and PAH-B Libraries with Aromatic Compounds.....	105
3.3.4 Analysis of Aromatic-Inducible Genes Recovered from the PAH-E Library	113
3.3.5 Creation of a Non-Sporulating <i>spo0A</i> Mutant in <i>Bacillus</i> 6A5.....	118
3.3.6 Inductions of <i>Bacillus</i> 6A5 <i>spo0A</i> Rock Bay Metagenomic Library.....	121
3.3.7 Trends in GFP Expression	124
3.3.8 HMW Aromatic Inducers.....	127
3.3.9 Mapping SIGEX Clones to NGS-Derived Scaffolds.....	127
3.4 Discussion	132

Chapter 4. Characterization of Aromatic-Inducible Operons in a Contaminated Soil Metagenome using Next-Generation Sequencing

4.1 Introduction.....	138
4.2 Methods	141
4.2.1 Contaminated Soil Treatment.....	141
4.2.2 Illumina Sequencing	141
4.2.3 SIGEX Experiments	144
4.2.4 Assembly and Analysis of Illumina Sequence Data.....	145
4.2.5 Mapping SIGEX Clones to Metagenomic Contigs	147

4.3 Results	150
4.3.1 MG-RAST Analysis	150
4.3.2 <i>De Novo</i> Assembly of Metagenomic Sequences	153
4.3.3 Directed Assembly of Aromatic-Degrading Features using PRICE	155
4.3.4 Mapping Aromatic-Inducible Clones to Assembled Metagenomic Contigs	157
4.3.5 Determining Biological Roles for Sequences Surrounding SIGEX Clones	162
4.3.6 Variation between Clones Mapped to Identical Metagenomic Contigs	171
4.4 Discussion	176

Chapter 5. Characterization of Microbial Communities in a Contaminated Site using Next-Generation Sequencing

5.1 Introduction	181
5.2 Methods	183
5.2.1 Metagenomic DNA Isolation and Sequencing	183
5.2.2 Assembly of Metagenomic Sequences and MG-RAST Analysis	185
5.2.3 Identification of Biodegradation Genes	185
5.3 Results	187
5.3.1 Overview of Annotations	187
5.3.2 Taxonomic Analysis	190
5.3.3 MG-RAST Analysis of Aromatic Metabolism	201
5.3.4 Biodegradation Gene Identification	214
5.4 Discussion	221

Chapter 6. Applications of SIGEX for the Design of Whole-Cell Bioreporters

6.1 Introduction	227
6.2 Methods	231
6.2.1 Bioreporter Recovery from Metagenomic Samples using SIGEX	231
6.2.2 Bioreporter Constructs	234
6.2.3 Mercury Bioreporter Assay	236
6.2.4 Measurement of Bioreporter Gene Expression	236

6.3 Results	237
6.3.1 Characterization of a Novel Bioreporter Recovered using SIGEX.....	237
6.3.2 Bioreporter Assays in LPS Mutants using Flow Cytometry Analysis	242
6.4 Discussion	247
6.4.1 Using Metagenomic Clones as Novel Bioreporters.....	247
6.4.2 Mutations in the LPS Core Result in Increased Hg Uptake in Bioreporters	248
6.4.3 Magnesium Hampers Hg Uptake in Bioreporters Independent of LPS Truncations	249
6.4.4 Conclusions	250

Chapter 7. Summary and Conclusions

7.1 Summary of Findings.....	252
7.2 Contributions to Scientific Knowledge	257
7.3 Future Directions	260
7.4 Concluding Remarks	261

Appendices..... 262

Appendix A. SIGEX-derived Clones Mapped to NGS Contigs	262
A.1. Contig 243.	263
A.2. Contig 3075.	267
A.3. Contig 3148.	268
A.4. Contig 3721.	271
A.5. Contig 6160	274
A.6. Contig 5976.	275
A.7. Contig 9794.	276
A.8. Contig 14785.	277
A.9. Contig 18132.	278
A.10. Contig 23284.	279
A.11. Contig 33223.	280
A.12. Contig 58390.	281
A.13. Contig 66283.	282

Appendix B. Annotation Tables of Contigs with Aromatic-Inducible Genes 283

 B.1. BLASTp search hits.284

 B.2. BLAST region annotations293

 B.3. BLASTp protein annotations.311

 B.4. InterProScan Annotations.320

References..... 350

List of Tables

Table 2.1.	Differential expression of various genes involved in PAH degradation in <i>Mycobacterium vanbaalenii</i> PYR-1	46
Table 3.1.	Bacterial strains and plasmids used in this thesis	84
Table 3.2.	Plasmid libraries used for SIGEX analysis in this chapter	94
Table 3.3.	Summary of analysis of inducible genes recovered from EXP-1 and EC-C600 library.....	102
Table 3.4.	Summary of genes, identified by tBLASTx ¹ within LMW-aromatic inducible clones, recovered from the PAH-E library	114
Table 3.5.	BLAST results showing matches between SIGEX-recovered clones and scaffolds in IDBA-UD assembled Illumina reads	130
Table 4.1.	Statistics of various <i>de novo</i> assemblies of Illumina-sequenced metagenomic DNA from the Rock Bay PAH-contaminated site	154
Table 4.2.	Statistical characteristics of contigs obtained from PRICE directed assemblies.....	156
Table 4.3.	Matches between SIGEX-recovered clone sequences and contigs assembled <i>de novo</i> from NGS data.....	158
Table 5.1.	Metagenome sequences uploaded to MG-RAST for analysis	184
Table 6.1.	Mutated genes in <i>E. coli</i> used for bioreporter membrane permeability experiments.....	235

List of Figures

Figure 1.1.	Substrate-induced gene expression (SIGEX) screening	13
Figure 2.1.	Chemical structures of various aromatic compounds used throughout this thesis	25
Figure 2.2.	A generalized pathway for aromatic ring breakdown	29
Figure 2.3.	Metabolic funneling of structurally varied aromatic hydrocarbons to the central metabolites catechol and protocatechuate	31
Figure 2.4.	Gene clusters for the degradation of PAHs.....	39
Figure 2.5.	Pyrene metabolism in <i>Mycobacterium vanbaalenii</i> PYR-1	42
Figure 2.6.	The 150 kb “region A” from <i>Mycobacterium vanbaalenii</i> PYR-1, a gene cluster responsible for catabolism of PAHs into TCA cycle intermediates	47
Figure 2.7.	A gene cluster for PAH degradation in <i>Mycobacterium</i> sp. strain SNP11.....	53
Figure 2.8.	Promoter region for <i>nidAB</i> in <i>Mycobacterium</i> sp. strain CH-2 as determined by primer extension and computational analysis	56
Figure 2.9.	The aromatic hydrocarbon (<i>arh</i>) gene cluster from <i>Sphingomonas</i> sp. strain A4	61
Figure 2.10.	The <i>car</i> -I and <i>car</i> -II loci in <i>Sphingomonas</i> sp. strain KA1	64
Figure 2.11.	The <i>car</i> operon in <i>Pseudomonas resinovorans</i> CA10	70
Figure 3.1.	Restriction digestion of metagenomic DNA for library preparation	88
Figure 3.2.	Flow cytometric calibration for single-cell sorting of bacteria	97

Figure 3.3.	Histograms of GFP expression during different rounds of flow cytometric sorting.....	106
Figure 3.4.	Histograms of GFP expression for aromatic-inducible clones isolated from the PAH-E metagenomic library	108
Figure 3.5.	Induction of naphthalene- and salicylate-inducible clones in <i>Bacillus</i> (bottom row) and <i>E. coli</i> (top row) hosts.....	111
Figure 3.6.	Induction of aromatic-inducible clones in microtitre plates using a variety of LMW aromatic compounds	116
Figure 3.7.	Spore stain of <i>Bacillus</i> 6A5 wild type (A) and 6A5 <i>spo0A</i> ⁻ (B) cells plated on dLB.....	119
Figure 3.8.	Sequential rounds of FCM sorting on the <i>Bacillus</i> 6A5 PAH-B SIGEX library	122
Figure 3.9.	Populations of GFP-expressing cells from genetically different clones	125
Figure 3.10.	SIGEX clones mapped to NGS-derived scaffolds that were assembled <i>de novo</i> using IDBA-UD	128
Figure 4.1.	Quality control data from the production of the 447 bp (A) and 255 bp (B) TruSeq gDNA libraries	142
Figure 4.2.	Flowchart illustrating the workflow carried out for the analysis of metagenomic DNA sequences	148
Figure 4.3.	Comparison of SEED subsystems from MG-RAST pipeline analysis	151
Figure 4.4.	Overview of aromatic-inducible SIGEX-recovered clones.....	160

Figure 4.5.	GFP induction of aromatic hydrocarbon-inducible SIGEX-recovered clones in microtitre plates using a variety of LMW aromatic compounds	163
Figure 4.6.	Annotation of aromatic degrading operons on assembled contigs, with catabolic genes found downstream of the mapped aromatic-inducible SIGEX-recovered clones	167
Figure 4.7.	SIGEX-derived clones mapping to the same contig likely originated from different individuals within the microbial community	169
Figure 4.8.	Chimeric sequences found in a subset of SIGEX clones.....	174
Figure 5.1.	Source hits distribution of MG-RAST annotated features.....	188
Figure 5.2.	Phylum-level classification of Illumina sequence reads as determined using the M5NR database in MG-RAST	191
Figure 5.3.	Phylum-level classification of IDBA-UD assembled contigs as determined using the M5NR database in MG-RAST	193
Figure 5.4.	Genus-level classification of Illumina sequence reads as determined using the M5NR database in MG-RAST	195
Figure 5.5.	Genus-level classification of IDBA-UD assembled contigs determined using the M5NR database in MG-RAST	197
Figure 5.6.	Comparison of bacterial rRNA annotations between raw Illumina reads and assembled contigs	199
Figure 5.7.	Functional category breakdown of reads classified as “Metabolism of aromatics” in the SEED Subsystems database using MG-RAST annotations of Illumina reads.....	202

Figure 5.8. Functional category breakdown of reads classified as “Metabolism of aromatics” in the SEED Subsystems database using MG-RAST annotations of assembled contigs	204
Figure 5.9. KEGG pathway analysis showing the enzymes required for metabolism of benzo[a]pyrene.....	206
Figure 5.10. KEGG pathway analysis showing the enzymes required for the metabolism of benzoate and catechol through hydroxylation	208
Figure 5.11. KEGG pathway analysis showing the enzymes required for the metabolism of toluene and xylene	210
Figure 5.12. KEGG pathway analysis showing the enzymes required for the metabolism of naphthalene and anthracene	212
Figure 5.13. A phylogram showing the taxonomy of biodegradation genes found in the Rock Bay PAH-contaminated site metagenome	215
Figure 5.14. Relative proportions of biodegradation genes in each species annotated within NGS contigs from the Rock Bay PAH-contaminated site metagenome	217
Figure 5.15. The species distribution (left) within each gene class in the BDG database.....	219
Figure 6.1. The <i>mer</i> operon and the K9 bioreporter clone used in this chapter	232
Figure 6.2. Mercury induced expression of clone K9 (in <i>E. coli</i> DH10B) isolated from the Petawawa military contaminated site metagenome using SIGEX	238

Figure 6.3. Mercury induction of metagenomic bioreporter clone K9 in <i>E. coli</i> DH10b and GS071.....	240
Figure 6.4. Effect of LPS truncations on Hg uptake	243
Figure 6.5. Effect of Mg ²⁺ on Hg uptake measured by flow cytometric analysis of GFP expression in various bioreporter hosts.....	245

List of Appendices

Appendix A. SIGEX-derived Clones Mapped to NGS Contigs262

A.1. Contig 243.	263
A.2. Contig 3075.	267
A.3. Contig 3148.	268
A.4. Contig 3721.	271
A.5. Contig 6160	274
A.6. Contig 5976.	275
A.7. Contig 9794.	276
A.8. Contig 14785.	277
A.9. Contig 18132.	278
A.10. Contig 23284.	279
A.11. Contig 33223.	280
A.12. Contig 58390.	281
A.13. Contig 66283.	282

Appendix B. Annotation Tables of Contigs with Aromatic-Inducible Genes283

B.1. BLASTp search hits.	284
B.2. BLAST region annotations	293
B.3. BLASTp protein annotations.	311
B.4. InterProScan Annotations.	320

List of Abbreviations

× g	force of gravity
°C	degrees Celsius
μg	micrograms
μL	microlitres
μM	micromolar
μm	micrometers
a.a.	amino acids
AH	aromatic hydrocarbon
Ap	ampicillin
Ap^R	ampicillin resistant
ARDRA	amplified ribosomal DNA restriction analysis
bp	base pairs
BTEX	benzene, toluene, ethylbenzene, xylenes
CIP	calf intestinal phosphatase
Cm	chloramphenicol
Cm^R	chloramphenicol resistant
CoA	coenzyme A
DFI	differential fluorescence induction
DGGE	denaturing gradient gel electrophoresis
dLB	dilute lysogeny broth
DNA	deoxyribonucleic acid
EDTA	ethylenediaminetetraacetic acid
Em	erythromycin
Em^R	erythromycin resistant
EtOH	ethanol
FACS	fluorescence activated cell sorting
FCM	flow cytometry
FISH	fluorescence <i>in situ</i> hybridization
FITC	fluorescein isothiocyanate
FSC	forward scatter
g	grams
Gb	gigabase pairs
GB	gigabytes
GFP	green fluorescent protein
h	hours
H₂O	water
HMW	high molecular weight
kb	kilobase pairs
kg	kilograms

Km	kanamycin
Km^R	kanamycin resistant
L	litres
LB	lysogeny broth
LMW	low molecular weight
M	molar
Mb	megabase pairs
MB	megabytes
MCS	multi cloning site
MDA	multiple displacement amplification
MeOH	methanol
Mg	magnesium
mg	milligrams
MG-RAST	metagenomic rapid annotation using subsystems technology
min	minutes
mL	millilitres
mm	millimeters
NF	non-fluorescent
NGS	next-generation sequencing
OD	optical density
OTU	operational taxonomic unit
PAH	polycyclic aromatic hydrocarbon
PBS	phosphate buffered saline
PCR	polymerase chain reaction
PE	paired-end (<i>i.e.</i> , DNA sequence reads)
PEG	polyethylene glycol
PMT	photomultiplier tube
PNK	polynucleotide kinase
rDNA	ribosomal DNA
RFLP	restriction fragment length polymorphism
RNAP	RNA polymerase
ROS	reactive oxygen species
rpm	rotations per minute
s	seconds
SIGEX	substrate-induced gene-expression
SIP	stable isotope probing
SSC	side scatter
Taq	<i>Thermus aquaticus</i> polymerase
TGGE	temperature gradient gel electrophoresis
Tris	2-amino-2-hydroxymethyl-propane-1,3-diol

Statement of Contribution

Chapter 2

Experimental design and manuscript editing	<i>Matt Meier</i> <i>Iain Lambert</i>
Literature review and writing.....	<i>Matt Meier</i>

Chapter 3

Experimental design and manuscript editing	<i>Matt Meier</i> <i>Suzanne Paterson</i> <i>Iain Lambert</i>
Carried out experiments.....	<i>Matt Meier</i>
Data interpretation and writing	<i>Matt Meier</i>

Chapter 4

Experimental design and manuscript editing	<i>Matt Meier</i> <i>Iain Lambert</i>
Carried out experiments and computational analyses	<i>Matt Meier</i>
Data interpretation and writing	<i>Matt Meier</i>

Chapter 5

Experimental design and manuscript editing	<i>Matt Meier</i> <i>Iain Lambert</i>
Carried out computational analyses	<i>Matt Meier</i>
Data interpretation and writing	<i>Matt Meier</i>

Chapter 6

Experimental design and manuscript editing	<i>Matt Meier</i> <i>Iain Lambert</i> <i>Alexandre Poulain</i>
Carried out experiments.....	<i>Matt Meier</i>
Data interpretation and writing	<i>Matt Meier</i>

Chapter 1.

Bacterial Responses to Pollutants in the Context of Metagenomics

1.1 Metagenomics: Access to Uncultured Microbes in the Environment

Microorganisms are ubiquitous participants in global biogeochemical cycles, and their ability to degrade or transform environmental pollutants that are otherwise persistent in the environment has sparked an interest in the mechanisms underlying these natural processes. Harnessing those biochemical reactions in the practice of bioremediation is an attractive option for eliminating harmful compounds from contaminated sites (Liu & Suflita, 1993), but understanding the microbial physiology underlying biodegradation processes that take place *in situ* can also be helpful for improving the process of biostimulation or co-metabolism (Nzila, 2013). However, our current understanding of bacterial metabolism is limited by our apparent inability to grow the vast majority of environmental bacteria in a pure culture (Epstein, 2013; Pham & Kim, 2012). In recent years, metagenomics, the study of DNA isolated directly from the environment, has become crucial for our understanding of microbial communities (Daniel, 2005). Metagenomic methods can access information about microbial communities, regardless of the propensity for their individual members to be cultivated (Handelsman, 2004). This thesis describes the use of a multitude of metagenomic

approaches to study microbial genomes in the context of xenobiotic-transforming or biodegradation genes in bacteria indigenous to contaminated sites.

The late 20th century saw an increase in literature suggesting that most species of bacteria in the environment have not yet been cultured using traditional laboratory methods (*e.g.*, growth on agar plates). Depending on the exact method used and the environment examined, current estimates suggest that between 90-99% of species are not readily culturable (Epstein, 2013; Handelsman, 2004; Pham & Kim, 2012; Rondon et al., 1999; Vartoukian et al., 2010). Initial clues from the disparity between microscopic cell counts and plate counts (termed “The Great Plate Count Anomaly”; Staley & Konopka, 1985) progressed into studies that examined the 16S rDNA sequences obtained from total isolated DNA and compared them to those of isolates from the same environment, revealing that many phylogenetic branches in bacteria are poorly represented by cultivated species (Handelsman, 2004). More recently, next-generation sequencing (NGS) technologies such as single-molecule real time sequencing (Pac Bio), ion semiconductor sequencing, pyrosequencing (454), sequencing by ligation (SOLiD), Illumina’s sequencing by synthesis technology (formerly Solexa; Reis-Filho & others, 2009), have allowed deep metagenomic sequencing of several environments (Abbai et al., 2012; Delmont et al., 2011a, 2012a; Fang et al., 2013; Hug et al., 2012; Qin et al., 2010; Ross et al., 2012; Yergeau et al., 2012a, b), and are beginning to provide significant insight into their enigmatic community structures.

This chapter introduces the importance of environmental microorganisms and compares the advantages of different culture-independent microbiological

methodologies. The aspects of environmental bacteria discussed in this chapter focus on their contributions to the transformation of xenobiotics, and the mechanisms through which such processes are regulated.

1.2 Bacteria and the Fate of Xenobiotics

Most biogeochemical cycles, notably the nitrogen and carbon cycles, rely on bacterial metabolism for completion of the cycle. For example, entry of atmospheric nitrogen (N_2) into the biosphere, in the biologically consumable form (NH_x), is *only* accomplished through reduction by the bacterial or archaeal enzyme nitrogenase (Kim & Rees, 1994). The other critical function performed by microbes is the mineralization of organic matter from the biosphere, converting almost every known organic compound back into CO_2 and its other mineral constituents (Díaz, 2004). Bacteria and other factors (both biological and physicochemical) can naturally attenuate xenobiotics through mineralization, partial biodegradation / transformation, or immobilization / demineralization. As a corollary of that fact, bioremediation is pursued as a method for the elimination of many xenobiotics. From an ideal bioremediation perspective, any toxic compounds should be eliminated, and in the case of certain chemicals – for example, chlorinated ethanes (Grostern & Edwards, 2006) – this has been achieved. However, in many cases, the reality is that some chemicals (*e.g.*, PAHs) can actually increase in toxicity over the course of their elimination due to the formation of persistent pathway intermediates (Lundstedt et al., 2007). In these cases, it is important to optimize the bioremediation process to minimize increases in toxicity.

1.3 Bacterial Gene Expression in the Environment

For bacteria to exist in inhospitable environments, they must adapt to changing environmental conditions (*e.g.*, temperature changes; osmotic concentration shifts; varying moisture and nutrient availability, etc.) as well as a wide variety of stressors (*e.g.*, toxic or mutagenic organic pollutants present in the environment; UV light or other forms of radiation; heavy metals; oxidative stress). These adaptive responses often arise through alterations in gene expression that are induced by effector molecules. The subsequent physiological changes are typically the result of a transcription factor (TF) protein binding to an RNA polymerase (RNAP) holoenzyme in a concerted action to express proteins that constitute a beneficial response to the inducing effector molecule (Decker & Hinton, 2013; Helmann, 2009). For example, salicylate (a common aromatic organic molecule) strongly induces transcription of the genes responsible for its degradation (Schell, 1985). Bacterial genes such as this are often regulated together and are transcribed as a polycistronic mRNA from a genetic structure known as an operon. Many operons and regulons have been characterized in detail in the era of molecular biology; however, most of these studies are done using model organisms. This is a drawback because a typical lab strain (*e.g.*, *Escherichia coli* K12) does not share the same genes (especially biodegradation genes) as a cultured environmental strain with a more versatile complement of metabolic functions (*e.g.*, *Pseudomonas* spp.). Furthermore, the characterization of transcription factors and regulatory mechanisms of uncultured microorganisms is even less common, due to the difficulty of isolating genes of interest from large and complex metagenomes.

The genomes of microbes isolated from soil and other complex environments contain more transcription factors and regulated promoters relative to those that occupy more stable environments, such as obligate pathogens (Cases & de Lorenzo, 2005). A major factor affecting the regulation of genes in complex environments is the presence of overarching regulons – *i.e.*, large groups of genes or operons that respond to the same regulatory protein, even if they are found in distinct genomic regions (Cases & de Lorenzo, 2005). For example, σ -factors are protein subunits of the RNAP holoenzyme (Helmann, 2009) that help specify which promoter types are activated for transcription: in addition to the “housekeeping” genes regulated by σ^{70} , there are σ -factors that activate genes for sporulation, starvation (stationary phase), flagellar synthesis, heat shock, iron transport, nitrogen starvation, and others (Campbell et al., 2008; Mooney et al., 2005; Wösten, 1998). Additional regulons that are relevant for environmental bacteria include catabolite repression (which alters gene expression depending on carbon source availability; Fischer et al., 2008; Görke & Stülke, 2008; Ramos et al., 1997; Valentini & Lapouge, 2012), integration host factor (involved in growth phase control; Cases & de Lorenzo, 2005), and oxidative stress response (Dempfle, 1996; Park et al., 2006). As a result of these regulons, large portions of the genome can be affected by several inputs at once: this forms multiple layers of regulation that together determine whether a particular gene will be fully “on” or “off”, or somewhere in between.

1.4 Molecular Methods in Metagenomics

The term metagenome was first used by Handelsman et al. (1998), in reference to the collective genomes of microbes living in soil. Less than two decades

later, the term persists as the name of a rapidly growing field studying microbial DNA isolated from the environment, which can range from the ocean (Venter et al., 2004) to animal microbiomes (Dantas et al., 2013; Ross et al., 2012) to various soil environments (Hug et al., 2012; Uroz et al., 2013). Initially, research was aimed at the recovery of novel functions (*i.e.*, a novel enzymatic reaction) that were phenotypically identified from clone libraries expressing random fragments of metagenomic DNA (Henne et al., 1999). Many such studies have been done in the ensuing years, resulting in a body of literature encompassing a wide variety of novel enzymatic functions that are found on metagenome fragments from uncultured organisms. However, the evolution from clone library metagenomics to high-throughput sequencing metagenomics was sudden: today, the MG-RAST database (<http://metagenomics.anl.gov/>) – which offers a widely used annotation pipeline for raw or assembled metagenomic DNA – currently contains over 34.5 terrabases (Tbp) of sequence from more than 30.4×10^{10} individual sequences in over 93,000 different metagenomes (of which over 13,000 are currently available to the public).

The following sections describe the most common approaches available to analyze metagenomic samples, outlining their advantages, disadvantages, and key innovations. It is important to keep in mind that many of the methods discussed do not always fit into a single category, and are often used in combination.

1.4.1 Functional/Phenotypic Screens

Functional screens for enzymatic activity can provide direct access to genes that perform a particular function or confer a specific phenotype. They rely on the correct heterologous expression of proteins in a host organism. Briefly, metagenomic DNA is cloned randomly into a vector: this could include plasmids for small inserts (typically <15 kb), or cosmids/fosmids/bacterial artificial chromosomes (BACs) for larger inserts (around 40 kb for cosmids/fosmids, and up to 350 kb for BACs). Libraries of clones are then screened for a particular phenotype. This method has been used widely (Gillespie et al., 2002; Hårdeman & Sjöling, 2007; Henne et al., 1999; Kim et al., 2007a; Knietsch et al., 2003; Lämmle et al., 2007; Lee et al., 2006; Ono et al., 2007), but there are constant improvements made to such systems including the use of broad-host range vectors (Craig et al., 2010; Wexler & Johnston, 2010). This is intended to increase the available metagenomic space that is available for sampling, based on the fact that promoter sequences originating from disparate taxa are expressed differently depending on the nature of the host organism (Gabor et al., 2004).

1.4.2 SIP

The incorporation of a heavy isotope (*e.g.*, ^{15}N , ^{13}C) into biomolecules from a specially made heavy-isotope growth substrate allows for the separation of DNA (or in some cases other biomolecules, such as proteins) that belongs to bacteria capable of utilizing that substrate. This procedure, known as stable isotope probing (SIP), enables researchers to link the identity of a microbe from a complex community structure to the degradation and consumption of individual compounds (Chen & Murrell, 2010). However, DNA-SIP only leads to enrichment of those organisms' raw genetic material: there is always a need to somehow identify the microbial species or their genes of interest, either through 16S amplicon sequencing, shotgun sequencing, gel-based techniques, or in combination with other metagenomic approaches (Chen et al., 2008). Furthermore, DNA-SIP yields the total genomic DNA of all organisms capable of utilizing the growth substrate, which is advantageous if the goal is to perform whole-genome analyses, but less so if the experiment is screening for specific genes (*i.e.*, those encoding enzymes that function in the degradation of the substrate). Nonetheless, SIP can be used to significantly narrow down the metagenomic sequence being analyzed in a sample (Chen & Murrell, 2010). Other potential caveats of this method include the possibility of cross-feeding and the difficulty in recovering sufficient quantities of labeled DNA.

1.4.3 RFLP, DGGE, and TGGE

One of the most common ways to measure microbial diversity has historically been through the PCR amplification of 16S rDNA genes (Maidak et al., 1999; Winsley et al., 2012). Using the 16S gene, it is possible to classify variants as operational taxonomic units (OTUs) among the amplified sequences. The fastest and cheapest methods for doing so generally involve gel-based techniques. Specifically, a mixture of PCR amplicons can be visualized using denaturing gradient gel electrophoresis (DGGE) or temperature gradient gel electrophoresis (TGGE); alternatively, if individual clones are first picked from a library, then each can be individually amplified and digested with restriction enzymes to be visualized via restriction fragment length polymorphism (RFLP) or amplified ribosomal DNA restriction analysis (ARDRA) (Kirk et al., 2004; Torsvik et al., 1998). These techniques each have the advantage of being rapid, and can be performed in-house with minimal expertise. By themselves, they can be used for community profiling – that is, to determine if one community contains different members than another. However, if desired, it is possible to sequence the individual 16S gene fragments and perform alignments with database entries to associate species with OTUs.

1.4.4 Probe-based Technologies

Oligonucleotide based probes can be used to hybridize functional genes or 16S genes in complex samples (Demanèche et al., 2009). Since these probes are designed to target specific gene families or specific species within a gene family, this method provides a rapid approach for determining the presence of a target population for metagenomic samples (typically a clone library) in a very directed manner. In a similar vein, PCR primers may be used as probes instead of the nucleic acid hybridization approach. Although PCR amplification of metagenomic DNA or the hybridization of probes in clone libraries is a straight forward method for detecting the presence of particular genes or species, it is also possible to use fluorescence microscopy to obtain direct cell counts and quantify different variants using fluorescence *in situ* hybridization (FISH; Hill et al., 2000). The advantage of this method is that a sample can be analyzed for species diversity without culture or the need for DNA isolation or library creation. This is valuable because not all metagenomic DNA isolation methods are equal, and may not always accurately represent the diversity present in an environment (Delmont et al., 2011b, 2012b). Additionally, FISH may be combined with flow cytometry (FCM; a method that is capable of analyzing and sorting individual cells as they pass by a detector in real-time; Shapiro, 2003) to enable the collection of cell counts in a high-throughput manner (Hill et al., 2000).

1.4.5 Gene and Protein Expression Analysis

In contrast to the techniques described in the preceding sections (which all focus on determining the presence of a specific target within metagenomes), many

methods have also been developed that examine how uncultured bacteria alter their gene expression in response to different environmental stimuli. The measurement of gene expression changes can help to identify genes of previously unknown function (through “guilt by association”), and may also give indications as to whether a gene or species is ecologically significant for a particular process.

A growing number of studies are utilizing various proteomic (Siggins et al., 2012) and transcriptomic (Carvalhais et al., 2012) approaches to measure changes in microbial gene or protein expression in microbial communities. Microarrays (Simon & Daniel, 2011) and qPCR (McDavid et al., 2012) have also been used for metagenomics. The types of large-scale datasets generated by these various –omics technologies can enable researchers to make new correlations between environmental factors and the identities of microbes present (Simon & Daniel, 2011).

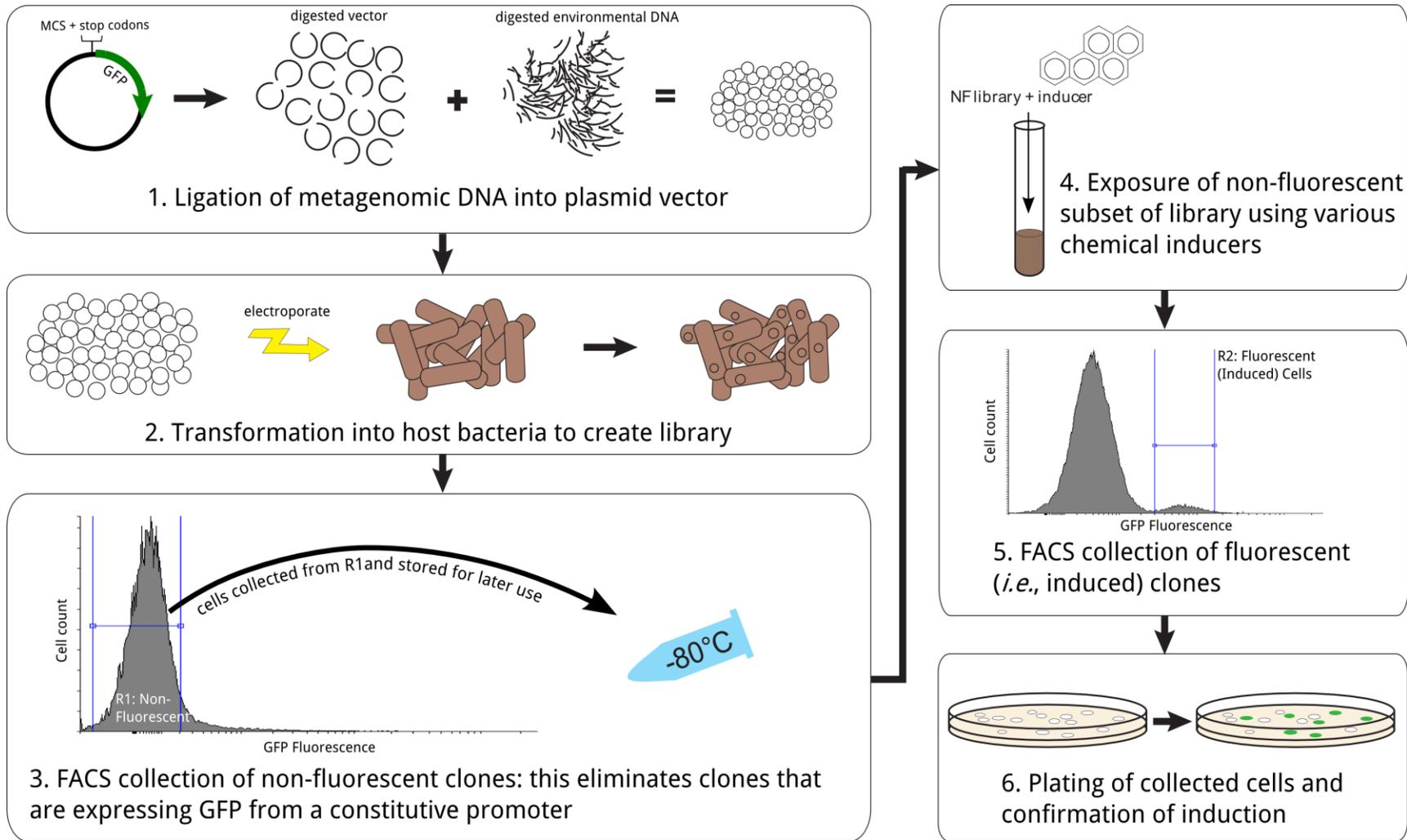
1.4.6 SIGEX

Substrate-Induced Gene Expression (SIGEX) was first described in 2005 by Uchiyama and colleagues (Uchiyama et al., 2005) as a high-throughput method for recovering novel genes from metagenomic samples. The assay is based on the idea that many prokaryotic genes are induced by their substrates, particularly in the case of catabolic operons (Díaz & Prieto, 2000). Uchiyama et al. (2005) reasoned that it is possible to screen a metagenomic library for operons involved in specific processes by shotgun cloning DNA upstream of a reporter gene and then measuring the expression of each clone in the presence versus the absence of an inducing compound. They did this successfully in a groundwater metagenome of ~152,000

clones (with an average insert size of 7 kb) using *E. coli* as a host organism to recover several benzoate and naphthalene inducible genes, including a novel cytochrome P450 enzyme. The reporter gene used was GFP, allowing high-throughput screening of metagenomic libraries containing many clones using fluorescence activated cell sorting (FACS).

In SIGEX, the metagenomic clone library is first propagated in liquid media lacking an inducer, and clones that are *not* expressing GFP are sorted (Figure 1.1). This fraction, which is non-fluorescent (NF) under normal growth conditions, represents clones containing genes that are not constitutively transcribed. The NF cells are then grown in the presence of the desired inducer; subsequently, any rare clones that are upregulated by this compound are sorted based on the expression of GFP. A detailed protocol for SIGEX has been published (Uchiyama & Watanabe, 2008). However, to date, few follow up studies – except for those from the original authors (Uchiyama & Miyazaki, 2010; Uchiyama & Watanabe, 2007) – have used SIGEX for metagenomic analyses (Lee et al., 2011). Nonetheless, it is recurrently cited as a method with excellent potential for recovering novel operons from metagenomic clone libraries (Daniel, 2005; Ekkers et al., 2012; Simon & Daniel, 2011; Taupp et al., 2011; Wexler & Johnston, 2010; Yun & Ryu, 2005).

Figure 1.1. Substrate-induced gene expression (SIGEX) screening. This schematic shows the main steps for cell sorting to obtain inducible clones from a metagenomic library.



In studies of pathogen-host interactions, a method very similar to SIGEX known as differential fluorescence induction (DFI) has been described (Rediers et al., 2005). DFI is used to screen for genes that are upregulated during pathogenesis (rather than catabolic genes). The primary difference is that DFI uses genomic libraries instead of metagenomic libraries, and the “inducer” may consist of exposing the bacteria to a host to induce an infectious response. Furthermore, there is usually an extra round of FACS.: instead of initially creating a NF sub-library, the first round of FACS in DFI is used to sort fluorescent – both induced and constitutively expressed clones – from the total library (in an inducing environment). Two subsequent rounds of FACS are used to sort non-expressing cells (in the absence of induction) and expressing cells (again in the presence of inducing signal). DFI has been used to screen for novel genes involved in biofilm formation by *Salmonella* (Hermans et al., 2011) and plant growth promotion by *Azospirillum* (Pothier et al., 2007).

1.4.7 NGS

The ability to rapidly obtain large quantities of high-quality sequence data is a relatively recent – but extremely important – advance in metagenomic analysis. Sanger (dideoxy chain termination) sequencing was used for several early (but influential) studies, including the Sargasso Sea shotgun sequencing and acid mine drainage community sequencing (Rusch et al., 2007; Tyson et al., 2004; Venter et al., 2004). In the acid mine drainage community (Denef et al., 2010; Tyson et al., 2004), the low complexity community structure (caused by the extreme nature of the environment) enabled a high read-depth (*i.e.*, multiple reads at a given locus

for each species) during sequencing. On the other hand, even though the Sargasso Sea was chosen for its relatively low complexity with regards to macrofauna, it was quickly revealed that a lack of ecosystem diversity with respect to larger life forms has little bearing on the prokaryotic diversity that may be present (Rusch et al., 2007; Venter et al., 2004). Since those publications, Sanger sequencing has fallen out of favor for shotgun sequencing, and NGS has become increasingly affordable and high-throughput. The main technologies responsible for this increase are pyrosequencing (*e.g.*, 454 GS FLX: 700 bp reads, ~0.7 Gbp / run), sequencing by synthesis (*e.g.*, Illumina/Solexa HiSeq: up to 100 bp paired end reads, 600 Gbp / run), and sequencing by ligation (*e.g.*, SOLiDv4: 50 bp paired end reads, 120 Gbp / run). Compared to Sanger sequencing, these technologies lack speed (24 h to 2 weeks per run for NGS; 3 h for Sanger) and accuracy (~98 - 99.94% for NGS technologies; 99.999% for Sanger); however, they make up for this with their ability to produce millions or even billions of individual reads in a single run, at a cheaper cost per base relative to Sanger sequencing (Liu et al., 2012).

1.4.8 Single-Cell Analysis

It is becoming increasingly important to analyze the individual cells within a community or clonal population. As sequencing technologies improve, single-cell genomics is now possible using multiple displacement amplification (MDA); this enables a very fine scale analysis of microbial communities that was not possible until recently (Blainey, 2013; de Souza, 2013; Stepanauskas, 2012). With single-cell genomics, entire draft genomes can be obtained (and numerous inferences about their biochemical pathways made) without the need to culture *or* to

computationally differentiate that organism's genomic DNA from a mixture of others. Although promising, this technology is currently resource intensive, and to characterize an entire microbial community would require many individual genome sequences.

1.5 Hypotheses and Objectives of this Thesis

Industrial activities often result in land that is contaminated by a variety of toxic chemicals; in Canada, over 21,000 locations are listed in the Federal Contaminated Sites Inventory. These sites range from groundwater, sediments, or soil to undeveloped lots within cities. Some of the common types of soil pollution found in these locations include: aromatic hydrocarbon or BTEX contamination originating from petroleum industry byproducts or fuel spills; toxic heavy metals such as mercury, arsenic or lead; chlorinated solvents; and pesticides. Generally, the toxicity, mutagenicity, carcinogenicity and teratogenicity are the primary human health concerns associated with the contaminants. Until remediated, contaminated sites are limited in their land use and productivity, since humans cannot safely access them or use their natural resources; furthermore, certain types of pollution are dangerous to wildlife, and may also biomagnify within food webs. Such sites are not only a major environmental health concern, but also represent a significant financial obligation to parties responsible for their cleanup. Although bioremediation is an appealing method for the detoxification of contaminated environments, we lack a detailed understanding of the degradation of many

xenobiotics in the environment, which can be a factor that limits the use of bioaugmentation during bioremediation of contaminated sites.

A multitude of molecular metagenomic methods were described in section 1.4; however, the weaknesses associated with those methods (*e.g.*, low throughput, reliance on protein expression, requirements for previous sequence characterization, *etc.*) demand the development of new approaches to study metagenomes. Although a wide variety of environments have been already used in metagenomic sequencing projects (Abbai et al., 2012; Delmont et al., 2012a; Hug et al., 2012; Kristiansson et al., 2011; Ross et al., 2012; Rusch et al., 2007; Shi et al., 2013; Suenaga et al., 2009; Uroz et al., 2013; Yergeau et al., 2012b, a; Yu & Zhang, 2012), the knowledge that we can gain from such data is limited by our functional understanding of the genes identified. A weakness of such studies is that the databases used for metagenome annotation are curated using functions that were assigned by biochemical studies of lab-domesticated bacteria; while this is effective for genes or proteins with high sequence similarity, it is not a viable way to predict highly specialized or novel functions unless combined with other independent methods. Therefore, an overarching goal of this thesis is to evaluate whether an approach utilizing contaminant-induced gene expression in combination with massively parallel metagenome sequencing can provide additional data to help us understand the contribution of both cultured and uncultured organisms to biodegradation in contaminated sites. This research may ultimately lead to the

development of new tools to help us evaluate contaminated sites and their remediation.

A focus of our laboratory has been the evaluation of hazards associated with aromatic hydrocarbons. Therefore, we have chosen to analyze a soil derived from a site that is heavily contaminated with low molecular weight (LMW; less than three benzene rings) and high molecular weight (HMW; three or more benzene rings) aromatic hydrocarbons. However, while there is a significant amount of information in the literature regarding the nature of enzymes associated with the degradation of simple aromatic compounds, and their genetic regulation, information regarding HMW compounds is limited and dispersed throughout the literature. Therefore, we have attempted to review what is currently known about the enzymology of HMW aromatic degradation, and the genetic regulation of these elements. This is found in Chapter 2. Chapters 3, 4, and 5 describe the optimization and implementation of SIGEX, a high throughput promoter trap method for identifying inducible elements in a metagenomic sample, and its combination with NGS. These studies provide a methodological framework for examining the genetic and functional components of contaminated soils. In principle, SIGEX also represents a powerful tool for the identification of bioreporters. In Chapter 6 we provide a proof-of-principal for bioreporter development using a Hg inducible element as an example. The thesis is described in the following objectives and hypotheses:

Chapter 2

Rationale Our understanding of the regulation of aromatic catabolizing genes is derived mainly from cultured organisms. Furthermore, current research on PAH-degrading operons has not fully described their mechanisms of transcriptional regulation.

Objectives Evaluate and summarize our current understanding of aromatic hydrocarbon degrading genetic elements, using published data, to determine what aspects in our understanding of aromatic degradation remain uncharacterized. This data will be used as a framework to understand aromatic degrading genes throughout the remainder of the thesis.

Chapter 3

Rationale Since environmental microbes are important for biodegradation processes, and our understanding of the regulatory mechanisms of biodegradation genes is largely based on the study of cultured microbes, this chapter aims to assess the prevalence of different aromatic-inducible functional genes found with culture-independent methods.

Hypothesis Aromatic-degrading elements are carried by individuals within the microbial community of a PAH contaminated site, and we can access a subset of those elements using SIGEX.

Objective To optimize and implement SIGEX for the recovery of physiologically relevant aromatic hydrocarbon-inducible clones from a PAH-contaminated site metagenome.

Chapter 4

Rationale Since metagenomic plasmid libraries (described in Chapter 3) only provide short cloned sequences, it is necessary to obtain the surrounding sequence to understand the genomic context of a library clone.

Hypothesis Massively parallel NGS data can be integrated with Sanger reads from SIGEX to obtain the sequences of the regions surrounding SIGEX-recovered genetic elements.

Objectives To map and annotate the greater genomic context of SIGEX-recovered clones to longer metagenomic contigs assembled *de novo* from NGS sequence data, and describe the genes present in the surrounding sequences as they relate to aromatic metabolism.

Chapter 5

Rationale Clone-based library screening methods, such as SIGEX, that rely on heterologous expression of metagenomic DNA may be unable to completely capture the biodiversity that is present in metagenomic samples. Chemical analysis of the Rock Bay soil bioslurry showed

that HMW-aromatic degradation took place; however, SIGEX screens did not uncover any HMW-aromatic degrading genes.

Hypotheses Genes previously known to be involved in biodegradation that were *not* retrieved using metagenomic library screens may still be detectable within metagenomic NGS data; thus, some PAH-degrading genes may be present in the Rock Bay soil sample despite their absence in SIGEX experiments.

Objectives To explore the functional genes present, and their taxonomic relationships to one another, in a PAH-contaminated site using NGS Illumina sequencing and *in silico* annotation methods. Genes in the biodegradation gene database (BDG) and the aromatic degrading genes described in Chapter 2 will be used to guide this analysis.

Chapter 6

Rationale SIGEX is designed to recover inducible genetic elements, and since the detection of xenobiotics can be achieved using genetic elements that are inducible by the compounds of interest (in combination with an easily measured reporter gene), it may be possible to screen for novel bioreporters using SIGEX. Several bioreporters already exist, but they could be improved via increased uptake of the compound of interest.

Hypothesis SIGEX can be used for the design of novel whole-cell bioreporter constructs. These, and existing constructs, can be improved by the use of bacterial hosts with truncated lipopolysaccharide components.

Objectives To test the SIGEX scheme for its efficacy of use in the characterization of novel whole-cell bioreporters and assess the bioavailability of mercury (a common environmental contaminant) within several *E. coli* strains with variable amounts of lipopolysaccharide coat.

Chapter 2.

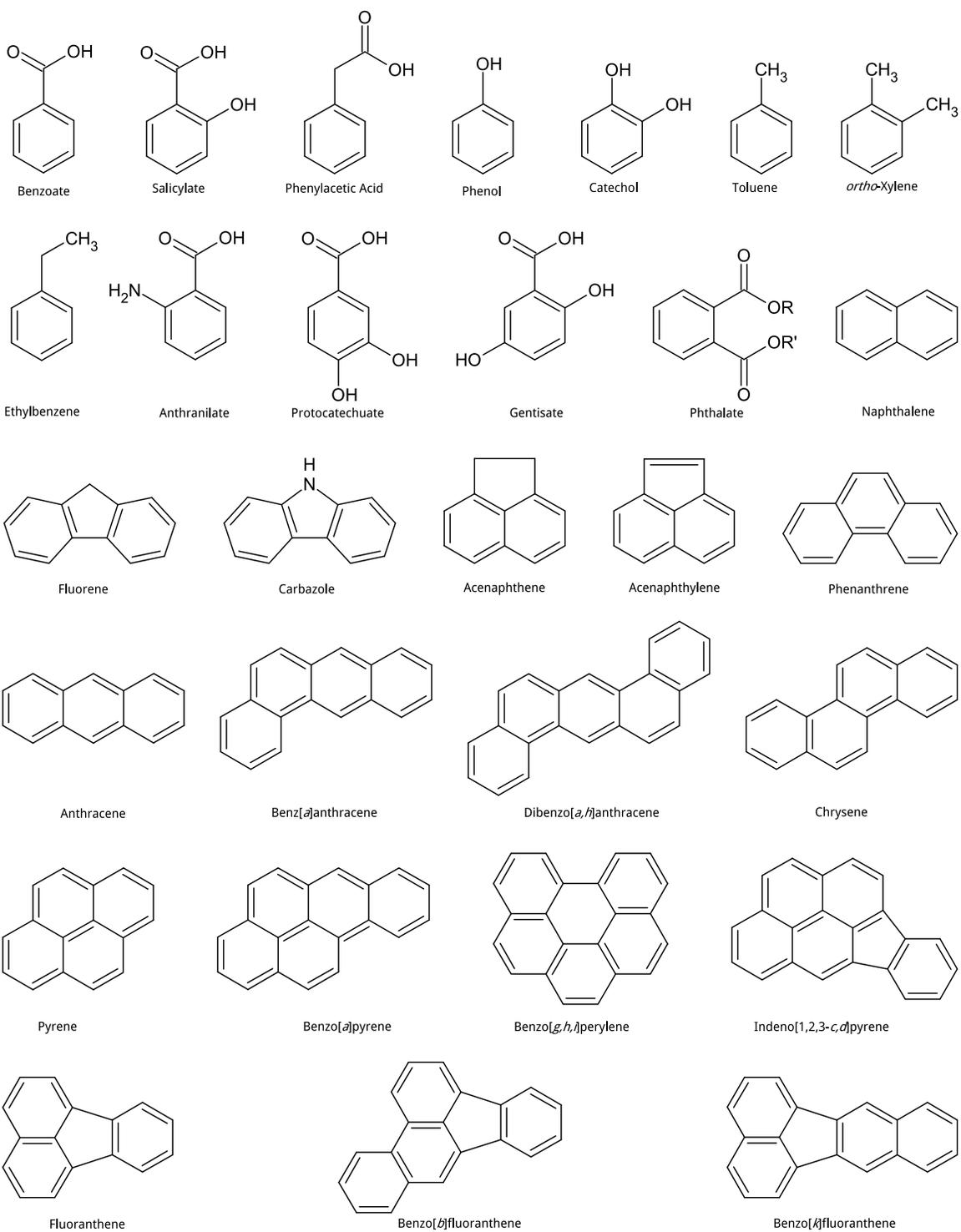
Genetic Regulation of Aromatic Metabolism

2.1 Introduction

2.1.1 AHs and Biodegradation

Aromatic hydrocarbons (AHs) and polycyclic aromatic hydrocarbons (PAHs) are ubiquitous environmental organic pollutants that contain aromatic rings (Figure 2.1; Samanta et al., 2002). They are relatively stable compounds owing to the delocalization of π -electrons in the benzene ring moiety and are consequently somewhat recalcitrant in the environment (Johnsen et al., 2005). PAHs are introduced to the environment through numerous ways including mobile sources (e.g., automobiles), coal, creosote production, manufactured gas plants, oil spills, and several natural phenomena such as forest fires, volcanoes, and petroleum seeps (Johnsen et al., 2007; Urata et al., 2004). Many PAHs are mutagenic, and in some cases, carcinogenic (Ames et al., 1972; IARC, 1983). Although microbial biodegradation is an attractive approach to their removal, bioremediation of sites contaminated with high molecular weight (HMW; 3 or more rings) PAHs is not used on a large scale despite speculation that it could be an economical alternative to more expensive remediation strategies such as excavation or thermal desorption (Labana et al., 2007).

Figure 2.1. Chemical structures of various aromatic compounds used throughout this thesis. Figure created using ACD/ChemSketch.



Soil microbes are highly active participants in global biogeochemical cycles and are responsible for recycling vast quantities and varieties of organic molecules, including PAHs (Kanaly & Harayama, 2000; Peng et al., 2008). Over the course of evolution, bacteria have developed a wide range of catabolic activities such that nearly any organic material can be degraded. However, these potentially valuable organisms are often difficult to analyze, as most estimates suggest that >99% of all prokaryotic species are resistant to culture (Epstein, 2013). Hence, uncultured bacteria contain a great deal of uncharacterized biological diversity, and, as it relates to this thesis, soil microbes may possess a large untapped genetic potential for the degradation of xenobiotics.

Many bacteria have been isolated that are capable of degrading both low molecular weight aromatics (LMW; fewer than 3 rings) and HMW PAHs. The biodegradation pathways of aromatics with a single ring (*e.g.*, benzoate, toluene, catechol, protocatechuate, *etc.*) have been reviewed extensively, as have the regulatory elements controlling transcription of genes in those pathways (Brinkrolf et al., 2006; Díaz & Prieto, 2000; Díaz, 2004; Gallegos et al., 1997; Gerischer, 2002; McFall et al., 1998; Prieto et al., 2004; Ramos et al., 1997; Tropel & van der Meer, 2004). The biodegradation of PAHs has also been reviewed (Haritash & Kaushik, 2009; Kanaly & Harayama, 2000; Mishra et al., 2001; Samanta et al., 2002; Shuttleworth & Cerniglia, 1995), and the evolutionary origin of aromatic-degrading genes has been characterized phylogenetically (Chakraborty et al., 2012). However, to our knowledge, the regulation of these pathways has not been discussed in detail. The subject of this literature review is the genetic regulation of pathways for HMW

PAH degradation in aerobic soil bacteria. This chapter will summarize studies characterizing the molecular biology of PAH-degrading pathways in several taxa (*Mycobacterium* spp., *Pseudomonas* spp., and *Sphingomonas* spp.) and highlight subjects upon which further research is required.

2.1.2 Mechanisms of Aerobic Biodegradation of Aromatic Compounds

Aerobic degradation of aromatics typically occurs through the initial oxidation of the ring by a terminal oxygenase (Peng et al., 2010b). The first catalytic step often adds two hydroxyl groups to the aromatic ring, via activation of molecular oxygen to form a *cis*-dihydrodiol, which is then reduced and re-aromatized to a dihydroxylated compound (Mason & Cammack, 1992). This product can be further metabolized by a ring-cleaving dioxygenase either *ortho*- (between) or *meta*- (adjacent) to the hydroxyl groups (Díaz, 2004; Vaillancourt et al., 2006). The ring-hydroxylating dioxygenase is often a multi-component enzyme comprised of a terminal dioxygenase made up of α - (large) and β - (small) subunits, as well as a ferredoxin and ferredoxin reductase component, which supply electrons to the terminal dioxygenase (Moser & Stahl, 2001; Figure 2.2). Aromatic rings are sequentially broken down and the resulting simple aromatics (*e.g.*, protocatechuate, catechol, *etc.*) are funneled into the tricarboxylic acid (TCA) cycle via the β -keto adipate or gentisate pathways (Díaz, 2004; Mishra et al., 2001). Some examples of aromatic funneling are shown in Figure 2.3.

Figure 2.2. A generalized pathway for aromatic ring breakdown. The first step is the formation of a *cis*-dihydrodiol of an aromatic ring, followed by dehydrogenation to re-aromatize the substrate to a dihydroxylated compound. This is followed by another oxygenation reaction, either *ortho*- or *meta*- to the two hydroxyl groups. Adapted from Takizawa et al. (1994).

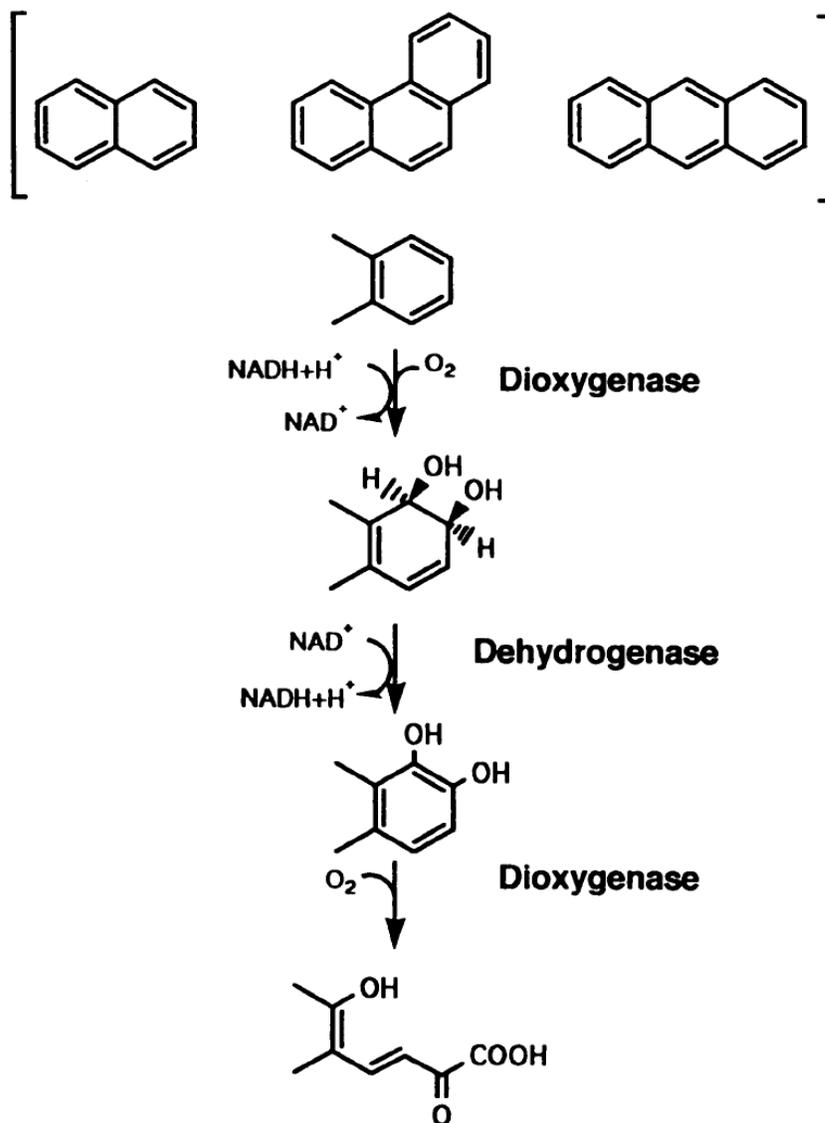
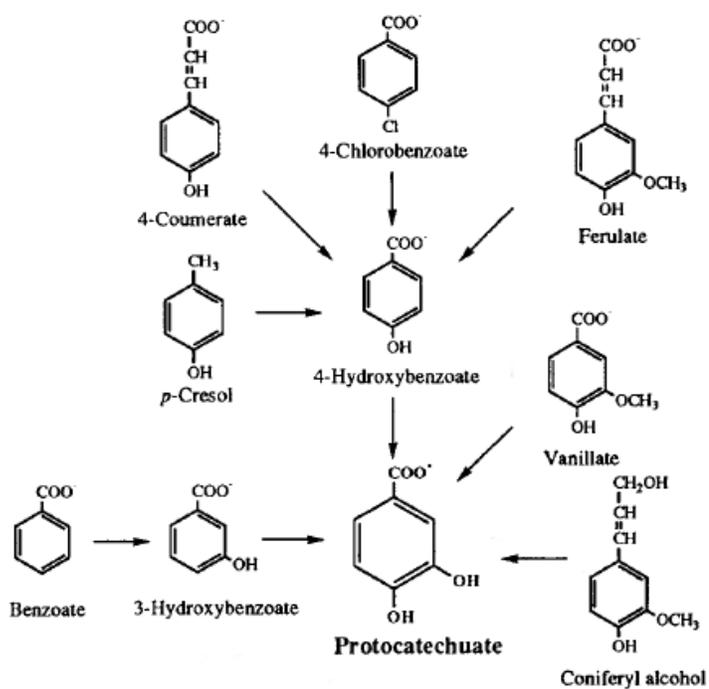
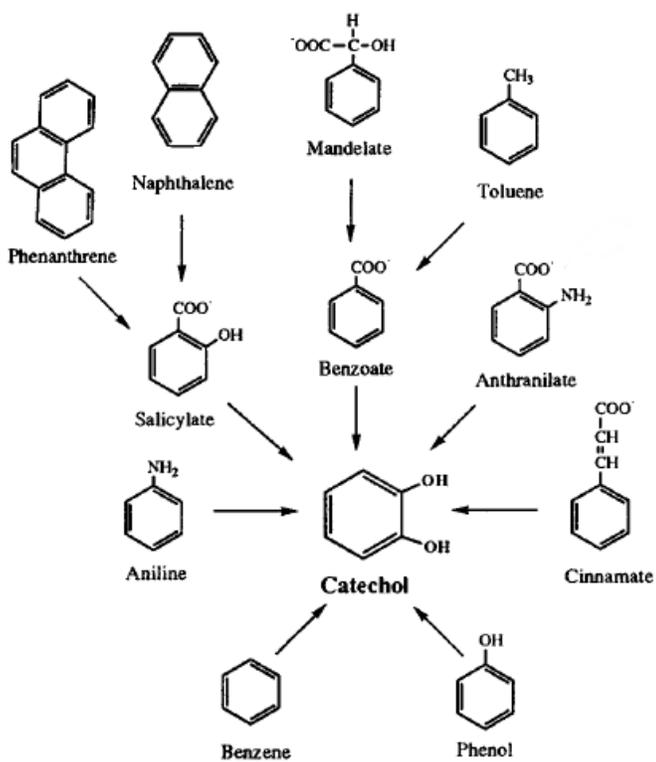


Figure 2.3. Metabolic funneling of structurally varied aromatic hydrocarbons to the central metabolites catechol and protocatechuate. Adapted from Harwood & Parales (1996).



2.2 Bioavailability and uptake of AHs

PAHs are bulky hydrophobic molecules with low aqueous solubility (Johnsen et al., 2005). Thus, PAH bioavailability in contaminated environments is generally low (Johnsen & Karlson, 2007), and one limitation of bioremediation is the difficulty accessing contaminants that are sorbed to material in the environment. Bacteria that are capable of degrading HMW aromatic compounds typically have unique cell wall characteristics that help solubilize PAHs, thereby increasing their bioavailability; for instance, sphingomonads have glycosphingolipids in their outer membrane, which assist in transport of hydrophobic molecules (Pinyakong et al., 2003). However, biodegradation of PAHs, as well as AHs, has been reported for a wide variety of both Gram-negative and Gram-positive species (Haritash & Kaushik, 2009; Kanaly & Harayama, 2000), indicating that the ability to access PAHs is present in many taxa regardless of their outer membrane properties.

Biological molecules that help solubilize hydrocarbons are known as biosurfactants (surface active agents; Li & Chen, 2009; Rahman & Gakpe, 2008). These compounds play an important role in PAH solubilization, and their presence is often associated with the ability to degrade PAHs (Nie et al., 2010). It has been suggested that solubilization and subsequent diffusion across the cell membrane plays the largest part in PAH uptake. One study found that adsorption to substances found in the soil, such as humic acids, may also increase access to PAHs in the environment (Smith et al., 2009). However, evidence in this important area is lacking: no reports were found of enzymes involved in active transport of PAHs across bacterial membranes. Recently, however, it was shown that the mycelia of

certain fungi (in that study, *Pythium ultimum*) can actively mobilize and transport some PAHs in the soil to a greater extent than diffusion alone, and that this can increase their accessibility to nearby bacteria (Schamfuß et al., 2013). This demonstrates the important role that fungi play in environmental degradation, and highlights their symbiotic relationship with bacteria. For some hypotheses regarding proteins that may be involved in xenobiotic transport in bacteria, refer to van den Berg (2005). To further complicate this issue, some strains of PAH-degrading bacteria do not produce biosurfactants (Dagher et al., 1997), leaving it a mystery as to which mechanisms other than diffusion (if any) are involved.

2.3 Regulation of Genes Involved in LMW AH Catabolism

Compared to PAHs, LMW AHs are usually simpler in structure, less hydrophobic, and occur more frequently in nature. For these reasons, biodegradation pathways for these compounds have been more extensively studied than those associated with the degradation of more complex HMW PAHs. There are several comprehensive reviews on various aspects of LMW aromatic metabolism including dioxygenation (Mishra et al., 2001), biodegradation pathways (Díaz, 2004), promoters involved in aromatic catabolic pathways (Díaz & Prieto, 2000), regulation of catechol (a common intermediate in aromatic degradation) operons (McFall et al., 1998), and transcription factors involved in LMW AH catabolism (Tropel & van der Meer, 2004). Furthermore, unique pathways of LMW AH metabolism have been described in detail in several bacteria, including – but not limited to – *Escherichia coli* (Diaz et al., 2001; Fernandez et al., 2006; Ferrandez et al., 1998, 2000; Nogales et al., 2007), *Pseudomonas* spp. (Cowles et al., 2000;

Delgado & Ramos, 1994; Jimenez et al., 2002; Parales & Harwood, 1993; Santos & Sá-Correia, 2007), *Rhodococcus* spp. (Haddad et al., 2001), *Sphingomonas* spp. (Story et al., 2001), *Burkholderia* spp. (Laurie & Lloyd-Jones, 1999), *Streptomyces* spp. (Park & Kim, 2003), *Acinetobacter* spp. (Dal et al., 2005), and *Corynebacterium glutamicum* (Brinkrolf et al., 2006).

Most of the transcription regulators in LMW aromatic catabolic pathways are positive regulators, meaning that in the presence of an inducing compound, the regulatory protein binds to a region of DNA upstream of the operon and promotes transcription. The most common families of transcription factors implicated in aromatic catabolic pathways are the LysR family, IclR family, AraC/XylS family, GntR type, TetR, MarR, or FNR types, XylR, and NtrC types (Díaz & Prieto, 2000; Tropel & van der Meer, 2004). With the exception of GntR type repressors, these transcription factors are activators. Each family of transcription factors has a “typical” gene organization (*i.e.*, location and direction of transcription in an operon), some of which will be discussed in more detail throughout the chapter.

2.4 Regulation of Genes Involved in HMW PAH Catabolism

In spite of the knowledge that many HMW PAH degradation or mineralization pathways exist in various bacterial species (or consortia of species), limited information is available regarding the control of transcription for the specific enzymes involved. The genetics of PAH catabolism was last reviewed by Habe & Omori (2003). This section presents an overview of regulatory mechanisms in

pathways for PAH degradation; discussion will revolve around several specific pathways that have been characterized in aerobic bacteria.

2.4.1 Regulation in *Mycobacterium vanbaalenii* PYR-1

The best characterized species with respect to its molecular biology of PAH degradation is, by far, *Mycobacterium vanbaalenii* PYR-1, a Gram-positive bacterium first isolated in 1986 from petroleum contaminated sediment for its ability to mineralize pyrene (Heitkamp et al., 1988). Remarkably, it is also capable of mineralizing or degrading – in addition to pyrene – biphenyl, naphthalene, anthracene, fluoranthene, 1-nitropyrene, phenanthrene, benzo[*a*]pyrene, benz[*a*]anthracene, and 7,12-dimethylbenz[*a*]anthracene (Kim et al., 2006). This strain has been the focus of numerous studies by Dr. Carl Cerniglia's group, in which multiple PAH inducible proteins and their respective genes were identified. PYR-1 was initially characterized by proteomics; more recently, the genome has been sequenced, revealing the complete set of pathways that are present (Kim et al., 2008; 2009). Strong evidence exists that there are two very separate regulatory pathways that control the genes for pyrene and fluoranthene metabolism, but specific modes of action for putative regulatory proteins remain elusive. Structural and biochemical data has been published regarding substrate specificity of the two terminal PAH oxygenases in PYR-1 (Kweon et al., 2010).

Among the genes that have been identified in PYR-1 are the archetypal *nidAB*, and *nidD* (called *nid* for naphthalene inducible dioxygenase) genes – part of a terminal PAH ring hydroxylating system. In particular, *nidA* and *nidB* encode the large (50 kDa) and small (13 kDa) subunits, respectively, of a PAH dioxygenase

that is divergent (40-50% amino acid identity) from other PAH dioxygenases (Khan et al., 2001). The proteins NidA and NidB were shown to convert pyrene and phenanthrene to *cis*-dihydrodiols (Khan et al., 2001). Brezna et al. (2003) was the first to show, using Southern hybridization, that *nidA*- and *nidB*- like genes are found at least twice in the PYR-1 genome. Furthermore, Kim et al. (2004) established that pyrene and fluoranthene induced different subsets of proteins, suggesting that different terminal dioxygenases were involved in the degradation of different PAHs; this has also been described in other strains of *Mycobacterium* (Krivobok et al., 2003).

Later studies elucidated the enzymes involved in phenanthrene (Stingley et al., 2004a) and phthalate (Stingley et al., 2004b) degradation in PYR-1. The genes involved in phthalate degradation are well-characterized; although it is not a HMW PAH, it is a downstream intermediate of *Mycobacterium* pathways for HMW PAH degradation. The phthalate operon consists of 5 genes transcribed in the same direction with an upstream divergently transcribed regulatory gene, *phtR*. PhtR belongs to the IclR family of transcription regulators; in aromatic catabolic operons, these proteins are typically transcribed upstream and divergently from the promoters they control (Tropel & van der Meer, 2004). However, this regulatory gene location and orientation is unique compared to other phthalate degradative operons in Gram-positive bacteria, suggesting that a different mechanism of regulation is present in PYR-1 (Stingley et al., 2004a, b).

NidA3B3, consisting of α and β subunits and showing a high degree of similarity to NidAB, was identified in PYR-1 (Kim et al., 2006) and expression of

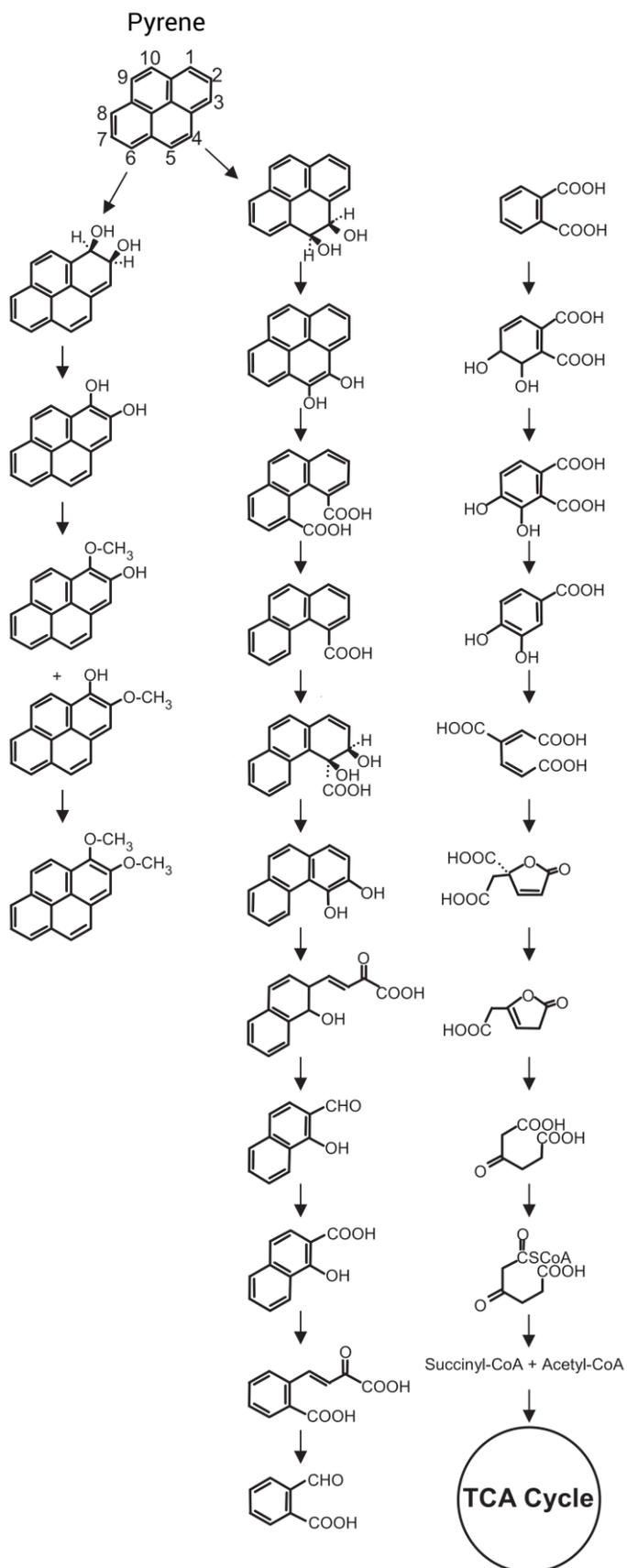
the proteins was measured by 2-dimensional gel electrophoresis (2DGE) in the presence of pyrene, phenanthrene, anthracene, and fluoranthene, relative to a glucose-only control. The NidA3B3 proteins were highly expressed during growth in fluoranthene, while the other PAHs induced minor expression. The genomic region encoding these proteins (Figure 2.4) contained four open reading frames: *nidA3*, *nidB3*, *nidR* (a putative MarR-family transcriptional regulator, transcribed upstream of *nidA3* and in the same direction), and *orf4* (a putative dehydrogenase). Although MarR-family regulatory proteins are typically located upstream of the operons they control, they are usually divergently transcribed (Tropel & van der Meer, 2004). This was the first report of a MarR-family regulatory protein near a *nid* type operon; however, it has not been tested for its ability to regulate the *nid* promoters.

Figure 2.4. Gene clusters for the degradation of PAHs. The organization of the *nidA3B3* terminal dioxygenase genes in *Mycobacterium vanbaalenii* PYR-1. Upstream is *nidR*, a putative MarR-family transcription regulator; downstream is *orf4*, a protein similar to alcohol dehydrogenase. The order of the genes is opposite that of the first discovered dioxygenase genes in this strain. Adapted from Kim et al. (2006).



Since its genome was sequenced, PYR-1 has been subjected to various systems biology studies (Kim et al., 2008, 2007b; Kweon et al., 2007), revealing a the global organization and regulation of genes. The first description of a complete pyrene degradation pathway was based on biochemical, proteomic, and genomic analysis (Kim et al., 2007b; Figure 2.5). Metabolites of pyrene were identified using gas chromatography-MS (GC-MS), proteins involved in metabolism were identified using 1-dimensional sodium dodecyl sulfate polyacrylamide gel electrophoresis (1D SDS-PAGE), and these data were correlated with predicted proteins identified in the PYR-1 genome database. Polypeptides were quantified in pyrene-induced and sorbitol grown cultures using MS for spectral counting. Using normalized peptide counts, 142 proteins were induced at least twofold; of these, 25 were established as aromatic degradation-related proteins. Kim et al. (2007b) described a complement of 27 enzymes that would allow complete degradation of pyrene into β -ketoacetyl-CoA, which would then enter the tricarboxylic acid (TCA) cycle. Several interesting observations were made with respect to regulation of the genes involved. First, the upper pathway (*i.e.*, enzymes involved in conversion of pyrene to phthalate) contained 16 proteins that were upregulated more than twofold; only one enzyme in the upper pathway was not detected as significantly upregulated. Moreover, the terminal dioxygenase, dihydrodiol dehydrogenase, and ring-cleavage dioxygenase were not detected in the control culture, but were abundant in the pyrene induced culture. These results suggest a very stringent regulation of the genes (either by efficient repression, tightly controlled activation, or both), the specifics of which have yet to be determined.

Figure 2.5. Pyrene metabolism in *Mycobacterium vanbaalenii* PYR-1. This was the first reported pathway for pyrene degradation with an enzyme assigned to each step. Modified from Kim et al. (2007b).



The approach used by Kim et al. (2007b) was also used to elucidate the fluoranthene metabolic pathway of PYR-1 (Kweon et al., 2007). That study identified 123 proteins that were upregulated by fluoranthene; of these, 76 were not present in uninduced controls. The other 47 were found in the control, but were at least twofold higher with fluoranthene. Fifty-three of these proteins were re-selected by proteomic analysis as possible fluoranthene metabolic enzymes, many of which showed overlap with the previously determined pyrene pathway (Kim et al., 2007b; Kweon et al., 2007). The most highly induced proteins included NidA3 – the α subunit of the dioxygenase whose fluoranthene activity was previously noted (Kim et al., 2006) – as well as several other oxygenases. The phthalate degradative pathway (*pht* operon) was not upregulated during fluoranthene metabolism (Kweon et al., 2007). This was an unexpected observation, as it was definitively present in previously characterized PAH-induced cells (Kim et al., 2007b); it is also unusual because the predicted metabolic pathway of fluoranthene leads to phthalate degradation followed by funneling to the β -keto adipate pathway.

Kweon et al. (2007) determined three possible pathways for the degradation of fluoranthene to TCA cycle intermediates in PYR-1, plus one dead-end pathway which is the result of a detoxifying *O*-methylation. This study provided evidence for metabolites that had not been previously reported (particularly the formation of 4-hydroxy-6-oxo-6*H*-benzo[*c*]chromene-7-carboxylic acid and 2-hydroxy-biphenyl-2'-carboxylic acid) and therefore filled crucial gaps in the pathway.

However, despite the knowledge of inducible genes, the regulatory proteins governing their transcription remain unknown.

The genome of PYR-1 was analyzed using BLASTP to assemble its PAH degradation pathways based on previous knowledge of PAH metabolic pathways in other organisms (Kim et al., 2008). Several observations have important implications for the regulation and genetic organization of the putatively involved genes. Kim et al. (2008) identified 194 genes that are likely related to aromatic hydrocarbon metabolism. The closest relatives to these genes were found in other *Mycobacterium* species as well as *Terrabacter*, *Arthrobacter*, *Nocardioides*, *Streptomyces*, and *Rhodococcus*. From their earlier proteomic data (Khan et al., 2001; Kim et al., 2004, 2007b; Kweon et al., 2007), Kim et al. (2008) were able to verify that 67 of those genes are expressed in the presence of PAHs. Summarized expression data for several key dioxygenase enzymes is shown in Table 2.1. In PYR-1, most of the dioxygenases are found in a large genomic region around 150 kb in length, containing all the genes required for PAH metabolism and the β -ketoacid pathway (Figure 2.6).

Table 2.1. Differential expression of various genes involved in PAH degradation in *Mycobacterium vanbaalenii* PYR-1 based on summarized proteomic data. Adapted from Kim et al. (2008).

Gene product	Peptide Counts		
	Control	Pyrene	Fluoranthene
NidA	0	13.5	0
NidB2	0	8.1	0
PhtAa	0	9.5	0
PhtAb	0	5.4	0
PhtAc	1.5	2.7	1.9
PhtAd	3.0	17.6	1.9
PhtB	0	17.6	0
NidA3	30.5	35.2	61.7
NidB3	23.5	33.8	18.0

Figure 2.6. The 150 kb “region A” from *Mycobacterium vanbaalenii* PYR-1, a gene cluster responsible for catabolism of PAHs into TCA cycle intermediates. The genetic organization is unusual in that it appears to be haphazardly arranged with respect to functionality. Functional categories are shown here by different colors: green designates PAH catabolic genes; red designates transcription regulation; blue designates genes involved in other functions; yellow designates a role in genetic mobilization and phage proteins; black designates membrane transport; white genes have no predicted function. Figure adapted from Kim et al. (2008).



2.4.2 Regulation in *Mycobacterium* spp.

Although PYR-1 has been examined most extensively, there are several other closely related *Mycobacterium* isolates that are capable of PAH degradation (Miller et al., 2004). At least two separate ring-hydroxylating dioxygenases were identified in *Mycobacterium* sp. strain 6PY1, a species capable of using pyrene or phenanthrene as carbon sources (Krivobok et al., 2003). 2DGE was used to screen for pyrene inducible proteins, and the two dioxygenases identified – Pdo1 and Pdo2, encoded by *pdoA1B1* and *pdoA2B2* – were capable of PAH degradation. Pdo1 dihydroxylated mainly pyrene, while Pdo2 oxidized mainly phenanthrene (Krivobok et al., 2003). These polypeptide sequences were similar to dioxygenases from the phenanthrene degrading *Nocardioides* sp. strain KP7 (Saito et al., 2000); the nucleotide sequences of *pdoA1* and *pdoB1* are 99% and 98% identical to *nidA* and *nidB* from PYR-1.

During heterologous expression in *E. coli*, the co-expression of ferredoxin components from a phenanthrene degrading strain *Nocardioides* sp. KP7 was found to increase the oxidation of both phenanthrene and pyrene. In terms of regulation, Pdo1 was induced by growth in benzoate, phenanthrene, and pyrene, but absent when grown with acetate; Pdo2 was not benzoate inducible, but was detected in the presence of pyrene and phenanthrene. Krivobok et al. (2003) suggest that Pdo1 is inducible by downstream metabolites of PAH degradation, while Pdo2 expression must be controlled by a different regulatory mechanism. The genes *pdoA1* and *pdoB1* were found in the same configuration as *nidAB* in PYR-1, with the small subunit upstream of the large subunit. The *pdoA2* and *pdoB2* genes have the

opposite configuration. No putative transcription factors or promoters have been described.

An isolate of *Mycobacterium* sp. strain S65 was found to grow on pyrene, phenanthrene, and fluoranthene as sole carbon and energy sources; two separate loci were found (Sho et al., 2004) using Southern hybridization to screen for *nidA* homologs. The genetic organization of one cluster was similar to the *nidAB* locus in PYR-1, the other cluster was similar to the *pdoAB* locus. These genes were shown to be inducible by pyrene and phenanthrene. Several novel genes were also described in this study, including the *nidX* and *pdoX* genes (found downstream of the *nid* and *pdo* loci, respectively), which were identified as α -subunits of aromatic dioxygenases.

Liang et al. (2006) used 2DGE to examine proteins expressed during pyrene degradation in *Mycobacterium* sp. strain KMS. This study was the first to identify all the components of a dioxygenase system in a single gel and show inducibility of the enzymes. A total of 17 proteins were upregulated by pyrene itself and two were upregulated by quinone. The pyrene induced proteins included aromatic ring hydroxylating dioxygenase subunits, dihydrodiol dehydrogenase enzymes, and iron-sulfur binding proteins (Liang et al., 2006). Multiple paralogs of dioxygenase enzymes were found, including at least five β -subunits of dioxygenases (Liang et al., 2006). The genome of KMS has since been sequenced (GenBank Accession Number CP000518, along with plasmids CP000519 and CP000520), revealing that the organization of the *nidAB* genes (known as Mkms_1667 and Mkms_1668,

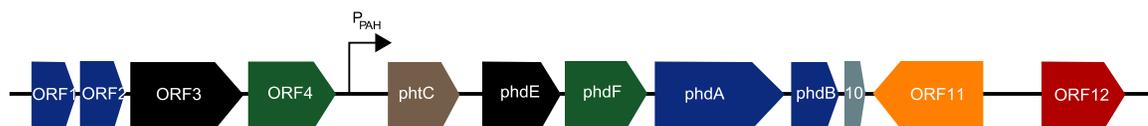
respectively, in KMS) is the same as *nidAB* in PYR-1, and the sequences are 99% identical.

Proteins upregulated during fluoranthene metabolism have been examined in *Mycobacterium* sp. JS14. Lee et al. (2007) used 1D SDS-PAGE and 2DGE and found 23 proteins with increased expression in response to growth in fluoranthene; among these upregulated proteins, several ring hydroxylating enzymes were identified. In addition to catabolic enzymes, the induction of several stress related proteins and a GMC-type oxidoreductase (which shares similarity with genes responsible for degradation of EPTC, a thiocarbamate herbicide) was described (Lee et al., 2007).

A 14 kb genomic fragment from *Mycobacterium* sp. strain SNP11 (capable of using fluoranthene, pyrene, phenanthrene, and fluorene as sole carbon and energy sources) was isolated using a probe derived from *pdoA2* from strain 6PY1 (Pagnout et al., 2007). The recovered fragment was sequenced and found to contain 12 ORFs. The *nidA* and *pdoA2* genes in SNP11 were >99% identical in nucleotide sequence to *nidA* in PYR-1 and *pdoA2* of 6PY1, respectively. Among the 12 ORFs, two distinct gene clusters were present, separated by an intergenic region containing a promoter (Figure 2.7). The upstream cluster contained genes similar to the α - and β - subunits of a dioxygenase, a dehydrogenase, and an extradiol dioxygenase. The second gene cluster contained a putative promoter region (P_{PAH}), followed by a putative decarboxylase, dihydrodiol dehydrogenase, extradiol dioxygenase, the α - and β - subunits of a dioxygenase, an unknown protein, a transposase, and an AraC/XylS family regulatory protein. The role of ORF12, the putative transcriptional

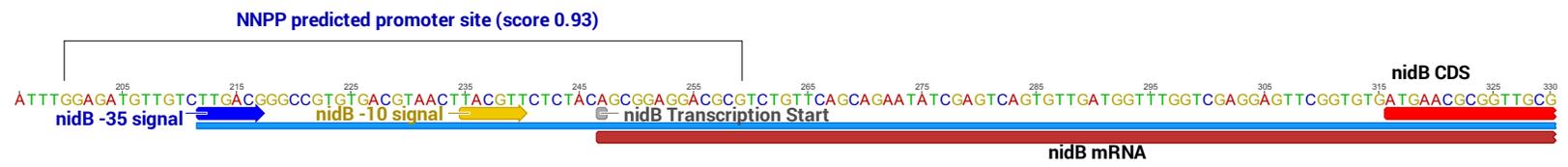
regulator, was not reported. However, the organization of these genes (*i.e.*, the regulatory protein transcribed in the same direction downstream of the operon) has been reported for other AraC/XylS type regulators in aromatic catabolic pathways (Tropel & van der Meer, 2004). The P_{PAH} promoter was tested for transcription activity using a β -galactosidase reporter gene. Expression, measured in *M. smegmatis* (whose genome does not contain any PAH oxygenases), was approximately 110-fold higher than a promoterless control; however, the inducibility of this promoter was not tested.

Figure 2.7. A gene cluster for PAH degradation in *Mycobacterium* sp. strain SNP11. The first cluster (left) encodes a dioxygenase (ORFs 1 and 2), dehydrogenase (ORF3), and extradiol dioxygenase (ORF4). The second cluster, transcribed from the promoter P_{PAH}, encodes a decarboxylase (*phtC*), dihydrodiol dehydrogenase (*phdE*), extradiol dioxygenase (*phdF*), dioxygenase (*phdA* and *phdB*), an unknown protein (ORF10), a transposase (ORF11), and an AraC/XylS family regulatory protein (ORF12). Adapted from Pagnout et al. (2007).



Mycobacterium sp. strain CH-2 can mineralize phenanthrene, pyrene, and fluoranthene, using the first two as sole carbon and energy sources (Churchill et al., 2008). Two dioxygenases were isolated from a genomic DNA library, one of which had high sequence similarity to *nidAB* from PYR-1, and the other was similar to *pdoA2B2* from 6PY1. The organization of the *nidAB* locus was similar to that previously described for PYR-1 (Kim et al., 2006), and the organization of *pdoA2B2* was similar to the homologous locus in 6PY1 (Pagnout et al., 2007). Furthermore, Churchill et al. (2008) predicted a putative promoter for the *nidAB* genes using computational analysis, and determined the transcription start site using primer extension (Figure 2.8). The transcription start site, an adenine (A) nucleotide, is located 68 bp upstream of the *nidB* start codon. The -10 and -35 sites of the putative promoter are consistent with other *Mycobacterium* species. Using a *lacZ* reporter gene, the promoter was found to confer strong transcriptional activity (approximately an 8-fold increase) over controls, which included a promoterless gene and a deletion mutant of the -10 region from the *nid* promoter (Churchill et al., 2008). To assess expression patterns, reverse transcription qPCR was done on acetate controls and phenanthrene induced cultures of CH-2. The genes *pdoA2B2* and *nidAB* were only detected in the phenanthrene induced cultures. Furthermore, the presence of the AraC/XylS type regulator downstream of the *pdo* gene cluster suggests a possible role in regulation; however, this was not directly tested.

Figure 2.8. Promoter region for *nidAB* in *Mycobacterium* sp. strain CH-2 as determined by primer extension and computational analysis. Modified from Churchill et al. (2008).



2.4.3 Regulation in Sphingomonads

The biodegradation of PAHs by *Sphingomonas*, a Gram-negative genus of bacteria, was reviewed by Pinyakong et al. (2003). The authors describe pathways for biphenyl, naphthalene, naphthalenesulfonate, phenanthrene, anthracene, fluoranthene, and fluorene degradation. Therefore, this section will focus on the more recent developments in our understanding of HMW PAH degradation in sphingomonads. Generally, the observations made by Pinyakong et al. (2003) regarding genetic organization of PAH degradative pathways in sphingomonads is that functionally related gene clusters are not always grouped together, raising questions as to how (or whether) the genes are regulated as cohesive units.

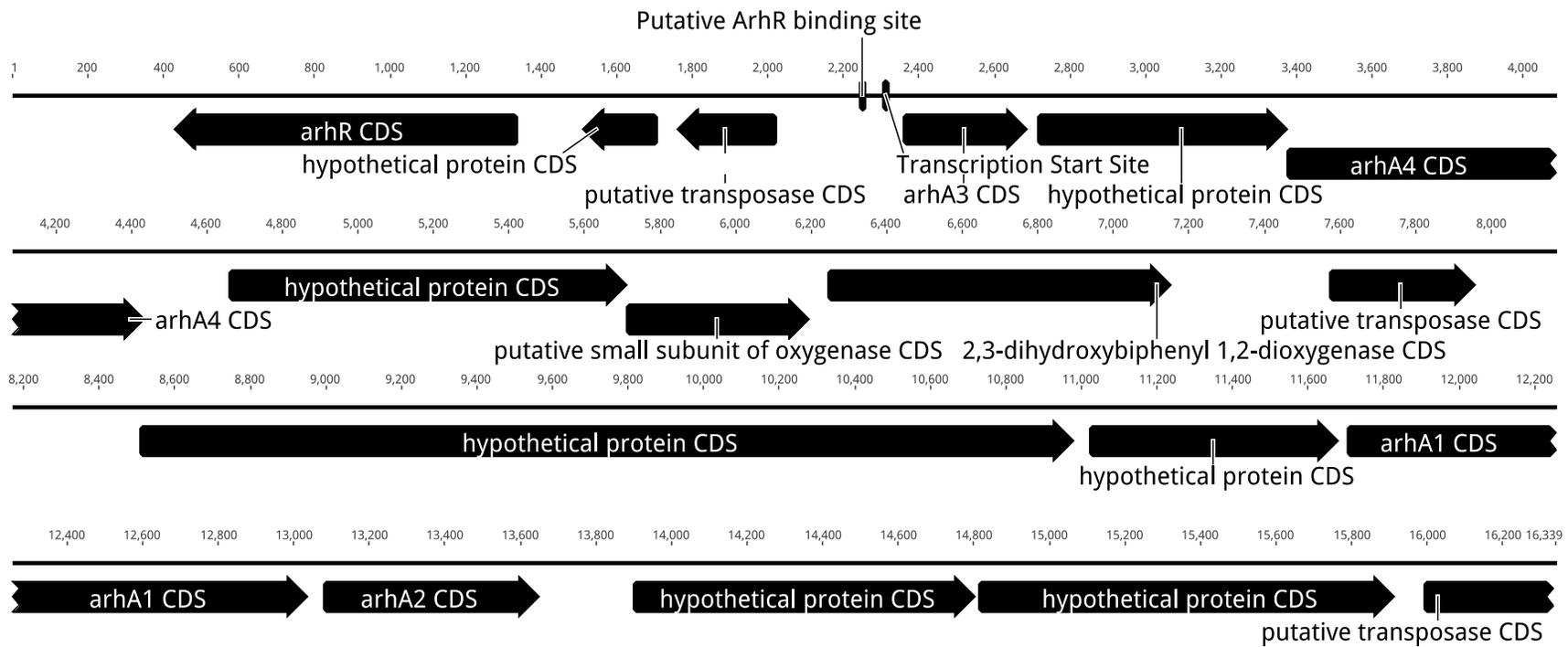
Sphingomonas sp. strain CHY-1 was isolated for its ability to grow on chrysene as the sole carbon and energy source (Demanèche et al., 2004). CHY-1 is also able to grow on naphthalene, phenanthrene and anthracene – but not acenaphthene, fluorene, fluoranthene, pyrene, benz[a]anthracene, or benzo[a]pyrene. Naphthalene and phenanthrene induced faster initial rates of chrysene mineralization compared to succinate and glucose. To identify proteins involved in the catabolism of chrysene, total proteins were analyzed using *in vivo* ³⁵S labeling and 1D SDS-PAGE. 2DGE was also used to identify PAH induced proteins. Two of the upregulated polypeptide sequences were used to design degenerate PCR primers; the resulting amplicons, a β -subunit of a dioxygenase and an extradiol ring-cleavage dioxygenase, were used to design primers for cosmid library screening. Two unique gene clusters were identified from the library – one contained *phnA1_aA2_aB*, encoding a terminal dioxygenase (designated PhnI) and a

putative alcohol dehydrogenase. The other locus contained a meta-cleavage dioxygenase, a ferredoxin, a putative ring-hydroxylating oxygenase (designated PhnII), and an isomerase, encoded by the genes *phnCA3A2_bA1_b*, respectively. These loci are arranged in a similar manner as catabolic genes from other sphingomonads. Isolation and heterologous expression of PhnI showed dihydroxylation of various PAHs; this was the first report of a PAH dihydroxylating enzyme in a *Sphingomonas* strain. PhnI was most active on naphthalene, but also acted on biphenyl, phenanthrene, and anthracene. Some degradation of benz[a]anthracene and chrysene was also observed – around 1% of that observed for naphthalene (Demanèche et al., 2004). This accounts for slow growth rates during growth with chrysene as the sole carbon and energy source. For PhnII, the small subunit is the first gene in the transcriptional unit, having a similar arrangement to the *Mycobacterium* spp. *nid* genes (Kim et al., 2008). A later study by the same group (Jouanneau & Meyer, 2006) characterized a dihydrodiol dehydrogenase (cloned separately from the other gene clusters; since the genome has not been sequenced, it is unknown which genes are located nearby) with specificity for PAH dihydrodiols, catalyzing the conversion to PAH-catechols for a variety of compounds including chrysene, benz[a]anthracene, and benzo[a]pyrene.

Degradation of acenaphthene and acenaphthylene was not extensively characterized until a study by Pinyakong et al. (2004). In that report, *Sphingomonas* sp. strain A4 was able to use both compounds as sole carbon and energy sources, but no other PAHs were reported as growth substrates. A 5 kb fragment of genomic DNA containing four ORFs was obtained by screening a library for

dioxygenase activity. ORFs 3 and 4 were similar to ring-hydroxylating dioxygenases, and therefore termed *arhA1A2* for “aromatic hydrocarbons” (Pinyakong et al., 2004). GC-MS was used to identify metabolites of PAH degradation in *E. coli* strain JM109 expressing the ArhA1A2 proteins from *Sphingomonas* sp. strain A4, as well as a ferredoxin / ferredoxin reductase (from *Sphingobium* sp. strain P2), from the *lac* promoter. In this heterologous expression system, ring hydroxylation of acenaphthene, acenaphthylene, naphthalene, phenanthrene, anthracene, and fluoranthene was observed (even though some of these are not growth substrates for A4); however, no pyrene degradation was detected. Phylogenetic analysis showed that the *arhA1A2* genes are closely related to other characterized PAH dioxygenases (including *phn*, *bph*, *nah*, *ndo*, *dxn*, *dnt*, and *car* family genes from a wide variety of Gram-negative species). Kouzuma et al. (2006) showed that ArhR (a LysR-type transcriptional activator) was required for acenaphthene degradation, and that induction by acenaphthene increased expression of the *arhA3* and *arhA1* genes by 2.4 and 6.2 fold, respectively (as measured by RT-PCR). Kouzuma et al. (2006) also identified a putative binding site for ArhR as well as the transcription start site. This constitutes one of few studies in which a transcription factor for a PAH-degrading gene cluster is described; in this case, the regulatory gene was located upstream and divergent from the promoter (Figure 2.9). Furthermore, the transcriptional unit for these genes was unusually long (over 11 kb), and separated from the transcriptional regulator by several remnant transposase genes (Figure 2.9), perhaps indicating a recent evolutionary origin. However, it remains unknown whether HMW PAHs can activate transcription via ArhR.

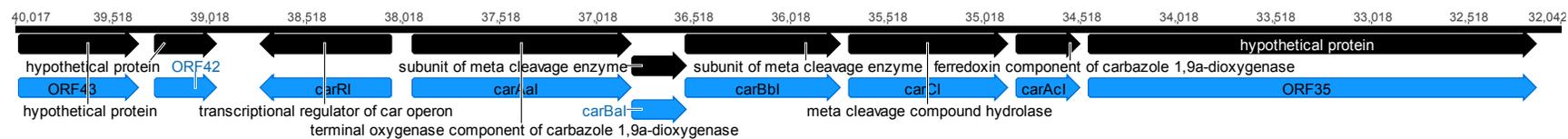
Figure 2.9. The aromatic hydrocarbon (*arh*) gene cluster from *Sphingomonas* sp. strain A4.



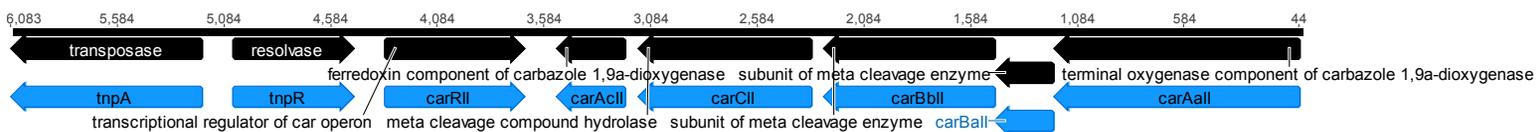
The plasmid pCAR3 found in *Sphingomonas* sp. strain KA1 contains genes that encode the mineralization of carbazole (Shintani et al., 2007). This system is unique in that it uses angular dioxygenation; additionally, carbazole is an N-heterocyclic PAH, so some aspects of its degradation are slightly different than those discussed up to this point. Angular dioxygenation is a mechanism by which oxygen is added to a carbon bonded to a heteroatom, as well as the carbon *ortho*- to that one, in an aromatic ring (Kasuga et al., 2001; Nojiri et al., 2001). The pCAR3 plasmid is around 250 kb in size and contains 263 ORFs. When classified according to their predicted functions, numerous degradative genes were found, some of which were previously known. There are two clusters for carbazole degradation on pCAR3, *car-I* and *car-II*, each containing genes for dioxygenation (Urata et al., 2006; Figure 2.10). In terms of function, CarAaI and CarAaII were both found to perform angular dioxygenation of carbazole to 2'-aminobiphenyl-2,3-diol; a ferredoxin protein (CarAcI or CarAcII) was required for this reaction, and addition of a ferredoxin reductase (FdrI or FdrII) dramatically increased the amount of degradation (Urata et al., 2006). Although various components of each locus could be interchanged with little consequence on efficiency of carbazole transformation, it appears that two separate systems exist, at least based on their genetic organization. Relating to gene expression, it was reported that reverse transcriptase quantitative (RT-q) PCR successfully amplified fragments from polycistronic messenger RNA (mRNA) in carbazole induced KA1. The genes *carAaIBaIBbICIAcI* (*car-I* locus) and *carAaIIBaIIBbIICIIAcII* (*car-II* locus) were each found to be transcribed in a single RNA molecule for each locus (Urata et al., 2006).

Figure 2.10. The *car*-I and *car*-II loci in *Sphingomonas* sp. strain KA1. The *carRI* GntR-family regulator is divergently transcribed from the *car* metabolic genes, while *carRII* is found downstream of the metabolic genes in its locus. Adapted from Urata et al. (2006).

pCAR3 Locus A (car-I gene cluster)



pCAR3 Locus B (car-II gene cluster)



A number of putative regulatory proteins were identified on the pCAR3 plasmid (Shintani et al., 2007). The CarRI protein regulates the *car-1* locus and CarRII regulates *car-2* (Figure 2.10); both are GntR-type regulators (Urata et al., 2006). GntR-type regulators are typically repressors that bind DNA adjacent to the promoter and physically block access to RNA polymerase (RNAP); they are released in the presence of an inducing compound (Tropel & van der Meer, 2004). Also found on pCAR3 is a group of genes, the *and* cluster, responsible for anthranilate dioxygenation. This group is regulated by *andR*, an AraC family regulator (Shintani et al., 2007). Another set of regulators, *catRI* and *catRII*, are activators for the *catBCAD* and *catFJI* genes, respectively (Shintani et al., 2007). This cluster is responsible for the degradation of catechol to acetyl-CoA. While *catRI* is similar to a LysR-type regulator, *catRII* is an IclR family regulator. Also identified were transcription regulators controlling genes for degradation of fluorene and dibenzofuran, the *dhb/fln* operon, under control of the LysR-type regulator *flnR*; another cluster, *lig*, involved in protocatechuate metabolism, is controlled by the LysR-type regulator *ligR*. Several other putative regulators were annotated, but were not functionally characterized.

Multiple levels of control for the carbazole degradative pathway are present (Shintani et al., 2007). This allows the upper pathway to be transcribed only when required, and it allows different growth substrates to induce the pathway best suited for their degradation. Also worth highlighting is the presence of at least 51 transposon and integration related predicted proteins: this suggests that the plasmid evolved by the piecing together of segments containing catabolic gene

clusters from various sources. It also points to the possibility of integration of these traits into the genomes of strains in which it resides. The putative transposons from which pCAR3 may have originated are described in the supplemental material of Shintani et al. (2007). A carbazole degradation gene cluster was also analyzed in *Sphingomonas* sp. strain CB3 (Shepherd & Lloyd-Jones, 1998).

The mineralization of benzo[*a*]pyrene by *Sphingomonas yanoikuyae* JAR02 has also been studied (Rentz et al., 2008). The notable feature of benzo[*a*]pyrene degradation is the requirement of a biostimulant, namely salicylate. Interestingly, strain JAR02 did not accumulate dihydrodiols following benzo[*a*]pyrene degradation. This is significant because the dihydrodiol intermediates are those implicated in mutagenesis and carcinogenesis (Ames et al., 1972). Since increases in mutagenicity sometimes occur following bioremediation (Lemieux et al., 2008; Lundstedt et al., 2003), it is important to determine factors that may affect the formation of toxic compounds during the process of biodegradation.

2.4.4 Regulation in Pseudomonads

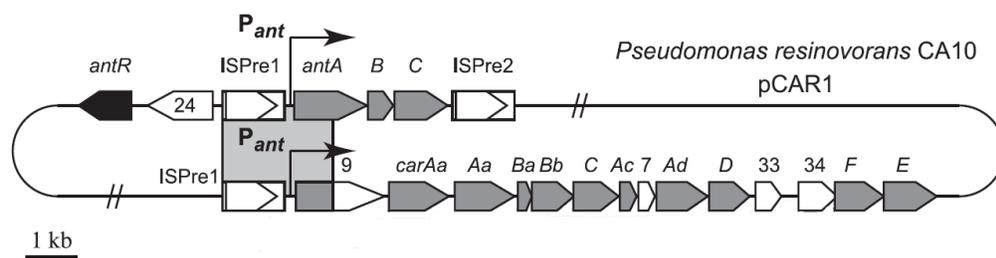
In *Pseudomonas putida*, several aromatic degrading pathways are very well known; for example, the toluene and xylene degradation pathways found on the TOL plasmid (Gallegos et al., 1997), the *ortho*-cleavage of catechol (McFall et al., 1998), and breakdown of naphthalene by the *nah* operon (Dennis & Zylstra, 2004; Park et al., 2005a, b, 2002; Simon et al., 1993). In these pathways, the enzymes, genes, regulatory features, and pathway intermediates are well studied. The metabolic pathway for pyrene degradation is known for *Pseudomonas* XPW-2 (Zylstra et al., 1994), a strain that can grow on naphthalene, biphenyl, anthracene,

chrysene, and pyrene. Genes encoding a carbazole dioxygenase were characterized from *Pseudomonas* sp. strain CA10 (Sato et al., 1997). From *Pseudomonas putida* OUS82 (Takizawa et al., 1994), a carbazole dioxygenase and a dihydrodiol dehydrogenase have been found. Unfortunately, gene expression for terminal dioxygenases acting on HMW PAHs is not well characterized. However, one study (Kamath et al., 2004) examined how plant matter (*i.e.*, carbohydrates, amino acids, aromatics, organic acids, volatile compounds, vitamins, and proteins) from the roots of various plants affects induction of PAH-degrading genes in *Pseudomonas fluorescens* HK44. That study was based on the observation that a higher rate of microbial PAH degradation is found in the rhizosphere than in unplanted soils. They report several plant-based inducers of the *nah* operon, as well as several repressors of the operon; the results of such induction studies could inform bioremediation strategies.

The operons for quinoline (Carl & Fetzner, 2005) and carbazole (Miyakoshi et al., 2006) degradation have been studied in *Pseudomonas putida* 86 and *Pseudomonas resinovorans* CA10, respectively. The carbazole operon has been extensively characterized with respect to genetic regulation; much like the *Sphingomonas* spp. carbazole degradation operon found on pCAR3, this carbazole operon is also found on a plasmid, pCAR1 (Urata et al., 2004). It contains the *car*_{CA10} gene cluster (Figure 2.10), which encodes enzymes for the upper pathway (*i.e.*, the breakdown of carbazole into anthranilate and 2-hydroxypenta-2,4-dienoic acid), and the *meta* pathway (which degrades 2-hydroxypenta-2,4-dienoic acid to TCA cycle intermediates; Miyakoshi et al., 2006). The upper pathway consists of

the genes *carAaAaBaBbCAcAd*. The *carAa* gene is in fact present twice – perhaps as part of an adaptation to increase its expression. The enzymes encoded by this operon are CarA, a multicomponent dioxygenase (consisting of a terminal dioxygenase [*carAa*], ferredoxin [*carAc*], and a ferredoxin reductase [*carAd*]); CarB (*carBaBb*), a *meta*-cleavage enzyme; and CarC (*carC*), a *meta*-cleavage compound hydrolase. The lower pathway is encoded by *carDEF*, which comprises the *meta*-cleavage pathway (Urata et al., 2004).

Figure 2.11. The *car* operon in *Pseudomonas resinovorans* CA10. Genes *carA* to *carC* are part of the upper pathway, and *carD* to *carF* are part of the lower pathway for meta-cleavage. The *ant* genes are responsible for anthranilate degradation. This schematic shows the σ^{70} promoter P_{ant} present in two copies, likely duplicated by a transposon at some point during evolution. It is regulated by the *antR* gene, an AraC/XylS type positive regulator. Adapted from Miyakoshi et al. (2006).



Anthranilate is cleaved by a 1,2-dioxygenase encoded by the *antABC* operon of strain CA10. The *ant* operon is transcribed in a single polycistronic message (based on qPCR), regulated by the promoter P_{ant} , from a transcription start site 53 bp upstream from the first start codon (Urata et al., 2004). The -10 and -35 elements of P_{ant} correspond to a σ^{70} promoter sequence, and at least 70 bp upstream of the promoter is required for anthranilate inducible expression. Induction by catechol (the pathway product) did not give rise to transcription of the reporter gene luciferase. The transcriptional regulator AntR, an AraC/XylS type regulator, is responsible for the inducibility of P_{ant} (Urata et al., 2004). The *antR* gene is divergently transcribed upstream of the *ant* operon. Most AraC/XylS type regulators are located upstream but transcribed in the same direction as the genes they regulate; however, as always, there are known exceptions (Tropel & van der Meer, 2004). AraC/XylS-type regulators are also often under regulation of another regulatory protein (Tropel & van der Meer, 2004), but this has not yet been investigated, to our knowledge, for AntR. The *car* gene cluster is also under control of the P_{ant} promoter, but a separate copy of it (sharing identical nucleotide sequence). It is located approximately 1.9 kb upstream of the *carAa* start codon, just upstream of ORF9, a remnant of transposition (Urata et al., 2004). This gene cluster is also inducible with anthranilate or carbazole. Expression of the *car* genes is constitutive (though low), and AntR positively regulates transcription from the P_{ant} promoter. The constitutive promoter, P_{carAa} , was also characterized in this strain (Miyakoshi et al., 2006).

2.5 Community-based Approaches to Characterize PAH-Degrading Genes

Genes that are upregulated by PAHs in cultured isolates are not necessarily the most ecologically significant factors for degradation, since microbial communities may be shaped to a greater extent by uncultured microbes that have yet to be characterized (Schloss & Handelsman, 2005). It is possible that species of bacteria possessing highly efficient PAH-degrading genes are not cultivable, or are selected against by isolation methods because they do not use the substrate of interest as a sole carbon or energy source. Studies that have examined the phylogenetics of PAH-degrading communities using 16S rDNA sequencing have found in several cases that the most common phylum is Proteobacteria (Andreoni et al., 2004; Lafortune et al., 2009; Lors et al., 2010; Molina et al., 2009), while many of the frequently isolated PAH degraders are Actinobacteria. A study using qPCR quantification of *nid* genes and *nah* genes from Mycobacteria and Proteobacteria found that both gene classes were present in a coal tar contaminated sediment, indicating that several known types of PAH-degrading pathways may act simultaneously in microbial communities (Debruyne et al., 2007). Furthermore, phylogeny based approaches (Chakraborty et al., 2012) show that most ring hydroxylating oxygenases (*i.e.*, of all homologs for each subgroup of RHO_alpha_C, CDD: cd00680, within the non-redundant NCBI database) derive from Proteobacteria (65%), while Actinobacteria comprise the next greatest proportion (20%). The remaining 15% are found in Firmicutes, Cyanobacteria, and Archaea. Cébron et al. (2011) used SIP on contaminated soil microcosms incubated with

phenanthrene as either the sole carbon source added, or along with root exudates. The population of active PAH degraders shifted substantially when root exudates were present. Furthermore, Uyttebroek et al. (2007) showed that Proteobacterial populations were dominant compared to *Mycobacterium* in PAH-amended microcosms. Taken together, these culture-independent studies indicate that the diversity of bacteria capable of such processes is high.

Although the proteins isolated by conventional methods might have important implications for fundamental research or interesting biotechnological applications, it may also be beneficial to use novel metagenomic techniques, such as promoter-trap systems including SIGEX (Uchiyama et al., 2005), for the recovery of novel PAH-degrading operons. Screens based on the transcriptional activation of metagenomic clones could be more useful in the context of uncultured species, due to the reduced reliance on complete protein expression and total mineralization of the test compound. This approach would produce a better community-centric vision of the genetics of biodegradation, although it suffers from the bias of requiring a compatible host organism. At this time, most culture-independent studies have examined variables such as dioxygenase copy number from various taxa during bioremediation, or the composition of 16S rDNA genes present during PAH degradation (Kahng & Oh, 2005; Ni Chadhain et al., 2006; Peng et al., 2010a).

2.6 Other Genetic Factors Influencing PAH Degradation

When considering the studies discussed in this chapter, it may be prudent to consider that a sampling bias likely plays a role when screening for proteins involved in PAH degradation. In the context of a single species growing on defined

media in the presence of a single PAH, a very different set of genes may be expressed when compared to typical environmental growth (Liu et al., 2005). Recent studies indicating the role of carbon catabolite repression also add a layer of regulation that may be relevant in a biotechnology context (Zhang & Anderson, 2013). Regulatory mechanisms, which are not well understood for PAH-degrading genes (Singleton et al., 2009), might behave differently when bacteria are living in a natural environmental context. Many different growth states are possible (and therefore different σ -factors will be used) and a highly complex environment full of possible effector molecules will alter gene expression drastically (Kamath et al., 2004). Regulation is also influenced by the selective pressures that gave rise to their function: while a particular catabolic gene is selected to perform a novel catalytic function, the regulatory elements must co-evolve mechanisms to keep the genes efficiently expressed – in other words, to increase transcription when the substrate is available, but *not* when a simpler alternate source of carbon or energy is present (Galvão et al., 2007). As with most regulatory systems, there ought to be a sensing mechanism to determine the concentration of the substrate (or cognate inducer) in the environment or inside the cell. Because of the complex evolution of many HMW PAH or xenobiotic degrading pathways, the genes may not be as well conserved or ubiquitously distributed as is the case for ancient pathways such as glucose metabolism. Consequently, it is difficult to identify trends in a coherent

manner, a problem that is exacerbated by the gaps in our knowledge of the diversity of microbial genetics.

2.7 Conclusions

A wide range of HMW PAH-degrading strains have been isolated from the environment, and many of those have been characterized with respect to the genes they carry that are responsible for HMW PAH ring metabolism (primarily the terminal multicomponent dioxygenases). A large amount of proteomic data has been generated with respect to the expression of PAH-degrading proteins in the presence of a variety of PAH inducers. In some cases, promoters have been characterized in upstream sequences, and occasionally, regulatory proteins have been found proximal to these clusters. However, except in a few cases (Kouzuma et al., 2006; Pinyakong et al., 2004; Tecon et al., 2006), the role of these regulatory proteins – and their induction via PAHs – remains undetermined.

A common theme in PAH degradative operons is that the genes for degradation of a single PAH are often haphazardly arranged, and found in atypical genetic structures (Figure 2.6); this suggests that the induction of a single pathway is complex and not necessarily affected by only one inducer or regulatory protein. It would appear that while certain genes are well conserved and widely distributed (e.g., *nid*- or *pdo*- like genes; see Peng et al. (2010b) for a review), their overall structure is relatively disorganized. This would explain why it is sometimes necessary to induce a particular strain with multiple aromatic compounds to achieve the highest possible induction (Gottfried et al., 2010; Pinyakong et al., 2003), and might also provide clues about their evolutionary origins (Shintani et

al., 2007). In PAH-degrading genetic elements, genes that are functionally related do not always appear to be co-transcribed; this diverges from a typical operon structure, in which a regulatory gene is often located proximally, and all the genes necessary for all or a few steps of the pathway are transcribed polycistronically. Contrast this with the coherent organization in aromatic degradation pathways found in *Pseudomonas* spp. degradative plasmids (Harayama *et al.*, 1987), wherein the same organization is also well conserved for isofunctional genes. This difference may be related to the possibility that PAH-degrading operons have complex evolutionary origins (Chakraborty *et al.*, 2012). Furthermore, an implication for bioremediation strategies is that conditions for maximal gene expression must be determined for different communities; this is rarely (if ever) achieved with a single HMW PAH inducer, and this is part of the reason that biostimulation can be so effective (Kamath *et al.*, 2004; Rentz *et al.*, 2008). Another observation common to the genomes of strains containing PAH oxygenases is that multiple copies of highly similar genes are often found (Urata *et al.*, 2004). Gene duplication can be an adaptation resulting in greater cytoplasmic enzyme concentrations, raising the degradation rate of toxic compounds (Urata *et al.*, 2006); alternately, a duplicated gene may branch off on a divergent evolutionary pathway to a novel function or may be regulated differently (Galvão *et al.*, 2007).

Several key issues regarding PAH degradation have yet to be addressed in the literature. First and foremost is the question of what transcription factors are responsible for the regulation of terminal oxygenase genes. It is clear that induction of PAH oxygenases takes place, both in response to the presence *vs.* absence of

PAHs, as well as different substrates such as pyrene vs. phenanthrene (Kim et al., 2004; Sho et al., 2004). However, there are evidently no reports of transcription factors that activate or repress genes in a manner that is preferential to specific PAHs. Whether such transcription factors exist, or whether the enzymes are regulated at a different level (such as translation), remains unclear. Insight into the regulation of these genes could have practical implications for bioremediation (Kamath et al., 2004), such that artificially creating conditions to induce the most environmentally relevant oxygenases might drastically improve the efficiency of degradation without competitive inhibition of the enzymes (McLellan et al., 2002). An objective of Chapters 3 and 4 is to identify and characterize novel genetic elements that are inducible by aromatic hydrocarbons, which may help us to understand the regulation of biodegradative enzymes. Furthermore, although this chapter has discussed several examples of metagenomic studies on PAH degradation, there is no clear consensus on which – if any – taxa are the most active PAH degraders in the environment. A better understanding of this problem might facilitate researchers to target the growth of specific populations of microbes during bioremediation.

Our understanding of PAH degradation will be enhanced by the characterization of newly cultivated species, aided by the exponentially increasing quantity of metagenomic NGS data. The evolutionary origins of PAH-degrading pathways can be better determined through examination of a wider variety of species. Metagenomics can help us determine the extent to which particular uncultured organisms are relevant to environmental PAH degradation. To obtain

a better understanding of PAH-degrading gene clusters, it is necessary to determine the cognate inducers of each pathway, as well as the mechanisms of their regulation; principally, this means characterizing the regulatory proteins responsible for driving transcription of PAH-degrading genes (which currently constitutes a large gap in the literature). Forthcoming genomic and metagenomic studies should help address many of these unanswered questions.

Chapter 3.

Exploration of an Aromatic Hydrocarbon Contaminated Soil Metagenome using Substrate-Induced Gene Expression

3.1 Introduction

Aromatic hydrocarbons (AHs), including polycyclic aromatic hydrocarbons (PAHs; aromatic hydrocarbons with 2 or more fused benzene rings), are a diverse class of chemicals that include a variety of toxic and carcinogenic substances that are widespread environmental pollutants (Lemieux et al., 2008; Lundstedt et al., 2007; Mumtaz & George, 1995). They are produced during manufacturing processes involving hydrocarbons, are by-products of combustion (including anthropogenic sources as well as natural sources such as forest fires or volcanoes), and are found in a wide array of industrial effluents (Johnsen & Karlson, 2007). Although the structure of PAHs imparts a relatively long environmental half-life (Kanaly & Harayama, 2000), most can be mineralized under both aerobic and anaerobic conditions (Haritash & Kaushik, 2009) by a variety of microbial species (DeBruyn et al., 2012; Labana et al., 2007; Peng et al., 2008).

At the community level, our understanding of AH biodegradation is constantly evolving amid the massive influx of novel sequence data – both in the context of individual genomes (Brunet-Galmés et al., 2012) and metagenomes (Martin et al., 2012). Progress has been made in understanding the specific enzymes and pathways (Kim et al., 2009) involved in aromatic degradation at contaminated sites, particularly regarding the terminal dioxygenases that perform the first step in aerobic aromatic metabolism (Kweon et al., 2010; Singleton et al., 2012); however, detailed characterization of enzymes on a case-by-case basis simply cannot keep up with the demand imposed by metagenomic discovery. Furthermore, the mechanisms through which these genes are regulated are still relatively unexplored.

While some controversy exists regarding the best way to mine metagenomes (Ekkers et al., 2012), a multi-disciplinary approach is generally needed to link microbial genes or species to biodegradative functions within contaminated sites. Techniques such as qPCR analysis (Yergeau et al., 2012a), microarray (Yergeau et al., 2009), expression libraries (Lämmle et al., 2007; Singleton et al., 2009) and stable isotope probing (SIP) (Jones et al., 2011) have been used successfully to identify organisms that may be involved in biodegradation, but, with the exception of expression libraries and SIP, do not always target functional genes. Substrate Induced Gene Expression (SIGEX) was initially proposed as a method for uncovering novel catabolic operons from metagenomes (Uchiyama & Miyazaki, 2010; Uchiyama & Watanabe, 2007, 2008; Uchiyama et al., 2005). It is based on single-cell sorting of clones from a plasmid library using flow cytometry (FCM):

metagenomic clones of interest are identified by the increased expression of a downstream fluorescent reporter gene in the presence – but not in the absence – of an inducing compound.

SIGEX was initially perceived as having great potential for mining genes from metagenomic samples in a high-throughput manner, without requiring prior knowledge of the sequences being screened for (Handelsman, 2005; Taupp et al., 2011; Yun & Ryu, 2005). However, SIGEX, and metagenomic promoter traps in general, have not lived up to this potential. In this chapter, four important factors are addressed that have limited the use of SIGEX: 1) the difficulty inherent to obtaining high quality metagenomic DNA for cloning, 2) the potential lack of compatibility between host transcriptional machinery and metagenomic DNA, 3) the difficulty associated with measurement of differences between populations with unusual distributions of gene expression within each population, and 4) the challenge of obtaining upstream and downstream sequences not contained on the SIGEX-cloned sequence. The utility of this SIGEX protocol is demonstrated in recovering differentially regulated genes from a metagenome derived from AH contaminated soil, and these genes are mapped to large metagenomic contigs derived from next-generation sequencing (NGS).

3.2 Methods

3.2.1 Plasmids, Strains and Growth of Bacteria

Strains and plasmids used in this chapter are listed in Table 3.1. Antibiotics were added from stocks to their appropriate final concentrations: ampicillin to 100 $\mu\text{g}/\text{mL}$, chloramphenicol to 10 $\mu\text{g}/\text{mL}$, kanamycin to 10 $\mu\text{g}/\text{mL}$ and erythromycin to 0.3 $\mu\text{g}/\text{mL}$. Dilute LB (dLB) was made at 1:10 strength relative to LB (Lennox L broth, Invitrogen), and if indicated, maltose was added to 2% (dLB/M). Dilute M9 media (dM9) contained 1X M9 salts, 1% pyruvate, 0.1% casamino acids (Difco), and trace elements (Sambrook & Russell, 2001).

Table 3.1. Bacterial strains and plasmids used in this thesis.

Strain	Genotype	Use in this thesis
<i>E. coli</i> DH5α	F ⁻ Δ(<i>argF-lac</i>)169 φ80 <i>lacZ</i> 58(M15) Δ <i>phoA8 glnV44</i> (AS) λ ⁻ <i>deoR481 rfbC1 gyrA96</i> (NalR) <i>recA1 endA1 thiE1 hsdR17</i>	Routine cloning strain
<i>E. coli</i> DH10b	F ⁻ <i>mcrA</i> Δ(<i>mrr-hsdRMS-mcrBC</i>) φ80 <i>lacZ</i> ΔM15 Δ <i>lacX74 recA1 endA1 araD139</i> Δ (<i>ara, leu</i>)7697 <i>galU galK</i> λ ⁻ <i>rpsL nupG</i> /pMON14272 / pMON7124	High-efficiency electroporation for ligated plasmid libraries, host for several SIGEX plasmid libraries
<i>E. coli</i> GS071	F ⁻ , [<i>araD139</i>]B/r, Del(<i>argF-lac</i>)169, lambda ⁻ , e14 ⁻ , <i>flhD5301</i> , Δ(<i>fruK-yeiR</i>)725(<i>fruA25</i>), <i>relA1, rpsL150</i> (<i>strR</i>), <i>rbsR22</i> , Del(<i>fimB-fimE</i>)632(::IS1), <i>deoC1 soxRS</i> ⁻	Derived from MC4100; used for testing for endogenous SoxRS regulators in paraquat inducible clones
<i>E. coli</i> NR6112	F ⁻ <i>lacpro</i> Δ(<i>lacpro</i>) <i>ara thi rfa</i>	Deep rough mutant; used as a library host when membrane permeability to large hydrophobic molecules was required (Lambert et al., 2001)
<i>Bacillus</i> 6A5 (<i>spo0A</i> ⁻)	<i>spo0A</i> ⁻	<i>Bacillus</i> host for SIGEX libraries; contains Em ^R marker in sporulation genes for ease of use in flow cytometry
Plasmid	Origin(s) of replication and selectable markers	Genes
pAD123 (5938 bp)	pBR322 (<i>E. coli</i>), pTA1060 (G ⁺ rolling circle replication for <i>Bacillus</i>); Cm ^R and Ap ^R	Promoterless <i>gfpmut3a</i> with 3 downstream stop codons (Dunn & Handelsman, 1999)
pMUTIN4 (8610 bp)	Em ^R and Ap ^R	Contains a cloning site where DNA fragments are inserted for creating the corresponding chromosomal gene knockouts in <i>Bacillus</i> species (Vagner et al., 1998)
pMMeb (5954 bp)	See pAD123	Promoterless <i>gfpmut3a</i> with 3 stop codons; also contains a novel MCS for cloning low-, mid-, and high-GC digested DNA

3.2.2 DNA Manipulations and Molecular Methods

General molecular methods were performed as described by Sambrook & Russell (2001). For routine plasmid isolation from *E. coli*, the Wizard® Plus SV kit or the Wizard® Midiprep kit was used according to the manufacturer's instructions (Promega). Restriction enzymes were purchased from New England Biolabs (NEB).

3.2.3 Soil Samples and Treatments

Rock Bay (Victoria Harbour, British Columbia, Canada; samples donated by BC Hydro and Transport Canada) was the location of various industrial activities since ca. 1862. Runoff from a coal gasification plant, tannery, propane tank farm, concrete batch plant, and an asphalt plant resulted in soil contamination consisting of significant quantities of coal gasification by-products, with PAH concentrations ranging from 475 to 12,600 $\mu\text{g/g}$ of soil (Whynot, 2009). Soil was homogenized and 5 L slurries (20% w/v in Milli-Q water) were prepared in duplicate BioFlo 110 bioreactors. Reactors were aerated and agitated continuously, incubated in the dark at a constant temperature of 25 °C, and kept at a pH between 6.5 – 8.0. Days 0-60 from this bioslurry showed a significant decrease or elimination of aromatic compounds, including complete elimination of naphthalene, acenaphthene, fluorene and phenanthrene, and an overall 60% decrease, by mass, of priority PAHs (Whynot, 2009). Samples were taken from each reactor at 15 day intervals over a 90 day period, starting on day 0; glycerol was added to samples at a final concentration of 15% and they were stored at -80 °C until use.

Canadian Forces Base (CFB) Petawawa was the source of an explosive-contaminated sandy soil, from the anti-tank firing position of a firing range,

donated by Dr. Sylvie Brochu (Defence Scientist, Life Cycle of Munitions Group, Energetic Materials Section, Defence Research and Development Canada [DRDC], Valcartier, Québec). The 2.5 cm of topsoil was collected with an acetone-rinsed stainless steel scoop; samples were stored immediately in polyethylene bags, and kept in the dark at 4 °C until use in bioslurry experiments done as described for Rock Bay soil, above.

3.2.4 Metagenomic DNA Isolation

From the Rock Bay bioslurry glycerol stocks, 5 time points (days 0, 15, 30, 45, and 60) were combined from each reactor for a total of 10 samples. Combined samples were centrifuged at 4500 x g for 10 minutes and 3 g sediment was collected and washed with an equal volume of wash buffer (10 mM EDTA, 50 mM Tris, 50 mM phosphate buffer, pH 8.0). DNA was extracted using the MO-BIO PowerMax® Soil DNA Isolation Kit, EtOH precipitated, and resuspended in 1 mL of 2 mM Tris (pH 8.0).

The CFB Petawawa sample used for metagenomic DNA isolation was taken from day 90 of the bioslurry time course.

3.2.5 Preparation of Vector DNA for Library Creation

The plasmid pMMeb (Table 3.1) was the vector for library creation. It carries the promoterless *gfpmut3a* gene preceded by a multicloning site (MCS) for the insertion of metagenomic DNA fragments, and has origins of replication for both *Escherichia coli* and *Bacillus cereus*. Three in-frame stop codons just upstream of *gfpmut3a* prevent translational read-through. Twenty µg pMMeb was digested to completion with BamHI for insertion of metagenomic Sau3AI fragments. DNA was

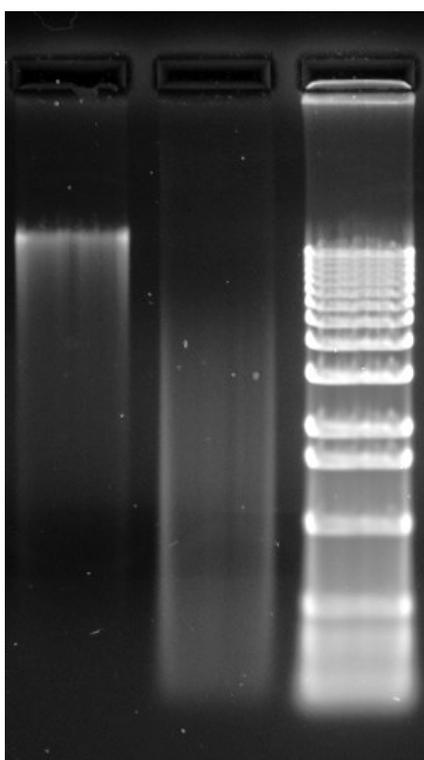
extracted with phenol-chloroform-isoamyl alcohol (PCI), then chloroform-isoamyl alcohol (CI), and precipitated with EtOH. DNA was dephosphorylated to completion using calf intestinal phosphatase (CIP; purchased from NEB), and purified from an agarose gel by GeneClean. The eluent was purified further by phenol extraction and EtOH precipitation, resuspended in 10 mM Tris (pH 8.0) and quantified by Nanodrop.

3.2.6 Preparation of Metagenomic Insert DNA

Metagenomic insert DNA was prepared by partial digestion of 5 μ g of DNA with Sau3AI. Purified DNA was run on an agarose gel, and showed elimination of the high-molecular-weight fraction and production of fragments from 0-12 kb with approximately even density (Figure 3.1). The desired size range was isolated from the gel using GeneClean, then purified by phenol/chloroform extraction and EtOH precipitation. Remaining contaminants – such as humic acids and any remaining small DNA fragments that could interfere with ligation – were eliminated by precipitation with polyethylene glycol (PEG) as follows. DNA dissolved in 100 μ L of 10 mM Tris pH 8.0 was mixed with 35% PEG8000 / 30 mM MgCl₂, vortexed, and centrifuged at room temperature at 21000 x g for 45 min. Pelleted DNA was rinsed with 70% EtOH. Size-fractionated digested metagenomic DNA was resuspended in 10 μ L of 10 mM Tris (pH 8.0).

Figure 3.1. 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 (Invitrogen).

1 2 3



3.2.7 DNA Purification by Gene Cleaning

Three volumes of cold 6 M NaI were added to the DNA sample. In the case of gel extraction, a volume of 6 M NaI equal to 3 times the mass of the gel slice was used, and the mixture was incubated at 60 °C for at least 10 min to completely melt the agarose. Melted gel samples were cooled slowly to prevent the formation of single-stranded DNA. Ten μL of glass milk silica beads were added to bind DNA; glass milk bound to DNA was pelleted by centrifugation at 10,000 X g for 10 seconds. The supernatant was removed and the pellet was resuspended by pipetting in 500 μL of cold NEW Wash buffer (50% EtOH, 100 mM NaCl, 10 mM Tris pH 7.5, 1 mM EDTA). This process was repeated twice and the pellet was dried at 37 °C for 10 min. DNA was eluted with autoclaved Milli-Q water or Tris-HCl buffered water by heating at 60 °C for 1 min. The DNA was recovered in solution following a final centrifugation at 10,000 X g for 10 seconds.

3.2.8 Phenol/Chloroform Extraction of DNA

Phenol extraction was performed as described by Sambrook & Russell (2001). Briefly, an equal volume of phenol (Invitrogen): chloroform: isoamyl alcohol (Sigma) in a 25:24:1 ratio was vortexed thoroughly with the sample. The emulsion was centrifuged for 2 min at 21,000 X g, and the aqueous phase was removed without disrupting the interphase. This process was repeated with one volume of chloroform: isoamyl alcohol (24:1), and the resulting aqueous phase was EtOH precipitated to desalt and concentrate the DNA.

3.2.9 EtOH Precipitation

To concentrate and purify DNA by EtOH precipitation, 1 μg of glycogen was added as a carrier. Subsequently, 0.1 volumes of 3 M sodium acetate (pH 5.2) was added to the sample, and gently mixed with 2.5 volumes of cold 100% EtOH. Following incubation on ice for at least 20 min, the sample was centrifuged at 21,000 X g for 30 min. The supernatant was discarded and the pellet of DNA was rinsed with 500 μL of cold 70% EtOH and centrifuged at 21,000 X g for another 2 min. The 70% EtOH was removed and excess EtOH in the pellet was allowed to evaporate at room temperature for 5 min before resuspension of the DNA in the required buffer.

3.2.10 Plasmid Library Construction

To create a plasmid library, size-fractionated DNA, digested with Sau3AI, was ligated into vector DNA digested by BamHI (5:1 insert:vector molar ratio). Typically, 10 – 100 ng of vector DNA at a concentration of 1-10 ng/ μL was required to obtain sufficient CFUs. T4 DNA ligase and buffer (Invitrogen) were added and the ligation reaction was cooled slowly to 4 °C (to encourage proper annealing of complementary ends) and incubated for 18 h. Ligated DNA was column purified using the PureLink™ PCR kit (Invitrogen) and eluted with 50 μL of nuclease-free water.

3.2.11 Transformation of Plasmid Libraries into Hosts

Transformation of *E. coli* cells with exogenous plasmid DNA was carried out in salt-free LB (1% tryptone, 0.5% yeast extract, no NaCl) using an *E. coli* Gene Pulser set at 2.5 kV using a 2 mm gap electroporation cuvette (Fisher). Thirty

individual electrotransformations using 1.5 μ L column-purified ligation mixture were performed, each containing 0.5 – 1 ng of ligated DNA. Incubation with 1 mL SOC was done for 1 h at 37 °C, with shaking at 220 rpm. Following recovery, 4 mL LB/Amp was added, and the cultures were incubated at 37 °C, with shaking at 220 rpm, for 16 h. The transformation cultures were pooled and harvested by centrifugation at 4500 \times g for 10 min at 4 °C. The pooled samples were resuspended in LB/Amp and 25% glycerol in a volume 1/10th the original culture volume, and stored at -80°C until use; this comprised the metagenomic library stock.

Electrocompetent *Bacillus* 6A5 (*spo0A*-) was transformed as described by Turgeon et al. (2006). Cells were prepared from cultures grown to OD 0.4 in LB supplemented with 250 mM sucrose; they were concentrated 150-fold by washing in 5 progressively smaller volumes of electroporation buffer (250 mM sucrose, 1 mM HEPES, 1 mM MgCl₂, and 10% glycerol), and stored at -80 °C until use. Transformations were done using the same protocol as for *E. coli*.

3.2.12 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 (Vagner et al., 1998) 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 conferring 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.

3.2.13 Libraries Constructed for this Thesis

The libraries interrogated in this chapter are described in Table 3.2. EC-C600 was made from *E. coli* DNA isolated via miniprep of bacterial genomic DNA (Wilson, 1997). PAH-E and PAH-B were constructed using Rock Bay soil DNA. The PAH-E library contains DNA from 10 combined samples comprising days 0, 15, 30, 45, and 60 from duplicate bioslurry experiments. The PAH-B library was generated using plasmid DNA isolated from the PAH-E library with subsequent transformation into a *B. cereus* host rather than *E. coli*.

Table 3.2. Plasmid libraries used for SIGEX analysis in this chapter.

Library	DNA Source	Characteristics	Host Organism	Insert Sizes	Total # Clones (Total Gb)
EC-C600	<i>E. coli</i> genomic DNA	DNA derived from <i>lac+</i> organism	<i>E. coli</i> DH10b (N.B. <i>lac-</i>)	4-10 kb	90,800 (0.6)
EXP-1	CFB Petawawa (On., CA)	Metagenomic DNA from an explosive contaminated site	<i>E. coli</i> DH10b	1.5-10 kb	601,000 (3.5)
PAH-E	Rock Bay (Victoria Harbour, B.C., CA)	Metagenomic DNA from a PAH contaminated site	<i>E. coli</i> DH10b and <i>E. coli</i> NR6112 (<i>rfa-</i>)	2-7 kb	1,600,000 (7.2)
PAH-B	Rock Bay (Victoria Harbour, B.C., CA)	Metagenomic DNA from a PAH contaminated site	<i>Bacillus</i> 6A5 (<i>spo0A-</i>)	2-7 kb	1,600,000 (7.2)

3.2.14 Flow Cytometry (FCM) Analysis

The BD Biosciences FACSaria II flow cytometer was used for cell sorting and measurement of fluorescence. This cytometer was fitted with a quartz cuvette flow cell operated at 70 PSI using the 70 μm nozzle, resulting in sheath flow velocities ~ 6 m/s at the point of interrogation. The light source was a non-polarized 488 nm laser (Coherent Sapphire solid state) operated at a power of 13 mW (Shapiro, 2003). Scattered light was not filtered in the FSC direction, but SSC light was filtered by a 530/20 band-pass filter for analysis of GFP fluorescence. Parameters were acquired using logarithmic amplification over 5 decades, with FSC and SSC thresholds of 200. Diagrams were created using Flowing Software (Perttu Terho, <http://www.flowingsoftware.com>).

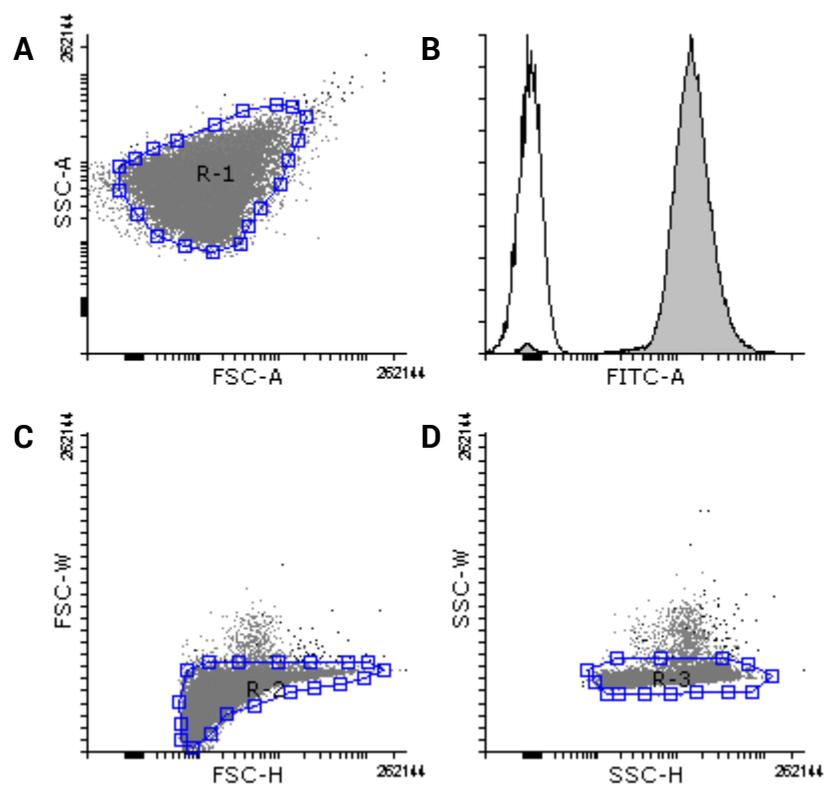
3.2.15 Identification of Bacterial Populations and Sorting Based on GFP Fluorescence

The FACSaria was calibrated for quality control on each use by running cytometer setup and tracking (CST) 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 (Figure 3.2). 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 ~ 2000 by adjusting cell density rather than flow rate; flow rate was kept at 1.0 to ensure optimal resolution of populations.

Figure 3.2. Flow cytometric calibration for single-cell sorting of bacteria. Blue dots represent an identified bacterial population. Cells outside of this region were excluded from analysis. **(A)** Dot plot showing FSC and SSC clustering of a population of *E. coli* (R1) **(B)** Histogram plots of GFP expression for pMMeb (promoterless GFP; left, unfilled) and pMMpos1 (constitutive GFP; right, grey fill) control plasmids. **(C)** Doublet gating on FSC channel (singlets are in R2). **(D)** Doublet gating on SSC channel (singlets are in R3).



3.2.16 SIGEX Induction Protocol

Prior to FCM analysis, a 100 μ L aliquot of the library was inoculated into 5 mL of media supplemented with appropriate antibiotics. When using frozen stocks, a 1 h recovery period in 1 mL SOC was used. Recovery cultures were rinsed in PBS 3 times before resuspension in media. For samples requiring induction, the inducing compound was added from a stock solution. Negative controls used a culture of the strain containing an empty vector (*i.e.*, no promoter upstream of GFP) treated with media (and the solvent of each chemical, if applicable). The culture was grown at 37 °C with shaking for 18 h unless otherwise stated. The cells were centrifuged at 10000 \times g for 3 min, washed 3 times in an equal volume of PBS and resuspended in 1 mL of sheath fluid (PBS). This suspension was then used to make dilutions at densities appropriate for analysis on the flow cytometer.

3.2.17 DNA Sequence Analysis

Initial Sanger reads of the metagenomic inserts were obtained using the primers GfpSeq (5'-GTTGCATCACCTTCACCCTCTCCACTGACAG-3') and pADLeft (5'-ACCTGACGTCTAAGAACCCATTATT-3'), which anneal to the vector upstream of the GFP and downstream of the MCS, respectively; subsequent reads were obtained by primer walking. Sequencing reactions were carried out by BioBasic (Mississauga, Canada). Sequences were aligned manually using BioEdit and analyzed using ORF finder (with bacterial codon usage) and subsequent BLASTp queries, or by simply searching the NCBI BLAST database using tblastx. An E-value cut-off of 1e-5 was used (Altschul et al., 1997).

Metagenomic shotgun sequencing of the matched DNA sample used for PAH-E library creation was performed by Genome Québec (Montréal, Canada) using the Illumina HiSeq platform, and reads were assembled using IDBA-UD (Peng et al., 2012). Compiled Sanger reads for each SIGEX clone were mapped to their respective NGS-derived contigs using Geneious v. 6.1 after initially determining the best hits to contigs using a local BLAST database of the assembled contigs.

3.3 Results

3.3.1 Library Construction

The construction of a metagenomic library containing a high proportion of clones with large inserts is critical to the success of SIGEX. While a variety of methods are published for purifying soil DNA (Schmeisser et al., 2007), it was found to be essential to eliminate short contaminating fragments of DNA and small molecules that can interfere with ligation. In our plasmid- and restriction enzyme-based cloning process, precipitation of metagenomic DNA with PEG 8000 following restriction digestion was determined to be the best way to eliminate a high proportion of small (<1kb) inserts. Clones with $\geq 90\%$ inserts > 1 kb were found in the libraries. For example, the PAH-E library, used for most of the experiments in this chapter, contained inserts between 1 and 10 kb in 36 of 40 clones tested by restriction digestion with HincII.

3.3.2 Proof of Principle: Induction of EXP-1 Metagenomic Library with Paraquat and *E. coli* Genomic Library with Paraquat and IPTG

Several experiments were undertaken to ensure that our modified SIGEX system, using the shuttle vector pMMeb, would function in the same manner as originally described (Uchiyama et al., 2005). A genomic library derived from *E. coli* C600 (*lac*⁺) genomic DNA was screened using IPTG and paraquat as inducers; the host organism, DH10b, was *lac*⁻. In both cases, biologically relevant inducible genes were recovered: for example, *lacZ* was recovered using IPTG induction, and *pqiB*, a known paraquat-inducible gene, was recovered using paraquat (Table 3.3).

Table 3.3. Summary of analysis of inducible genes recovered from EXP-1 and EC-C600 library. Induction was measured by flow cytometry unless noted.

Inducer	Library	Clone	Fold Induction	Genes/Proteins (Accession)	% A.A. Identity	Organism	Taxonomy (Phylum: Order)	Comments
IPTG	EC-C600	IPTGH4	5.42 ¹	ND	ND	ND	ND	Cleaves Xgal only in presence of IPTG
		IPTGH8	6.69 ¹	beta-galactosidase (CP002890) transcriptionally fused to GFP	97	<i>E. coli</i> UMN18	Proteobacteria: Gammaproteobacteria	Cleaves Xgal only in presence of IPTG
		IPTGH10	10.19 ¹	beta galactosidase small chain <i>lacZ</i> (CP002729) transcriptionally fused to GFP	100	<i>E. coli</i> UMNK88	Proteobacteria: Gammaproteobacteria	Cleaves Xgal only in presence of IPTG
PQ	EC-C600	PQD5	5.88	ND	ND	ND	ND	ND
		PQD9	4.01	<i>pqiB</i> (YP_003228134.1), transcriptionally fused to GFP	98	<i>E. coli</i> O26:H11 str. 11368	Proteobacteria: Gammaproteobacteria	Paraquat-inducible protein B; known to be upregulated by superoxide stress.
		PQC1	3.40	ND	ND	ND	ND	ND
		PQE1	3.29	ND	ND	ND	ND	ND
		PQE2	2.79	ND	ND	ND	ND	ND
PQ	EXP-1 Meta-genome	PQG11	70.24	peptidase M16 domain-containing protein (ACU58203.1), transcriptionally fused to GFP, <i>pqqL</i> domain	49	<i>Chitinophaga pinensis</i> DSM 2588	Proteobacteria: Deltaproteobacteria	Inducibility abolished in SoxRS- <i>E. coli</i> ; <i>pqqL</i> known to be upregulated during <i>E. coli</i> superoxide stress.
		PQA20	92.15	hypothetical protein MXAN_1454 (ABF93145.1), transcriptionally fused to GFP	38	<i>Myxococcus xanthus</i> DK 1622	Proteobacteria: Deltaproteobacteria	Inducibility abolished in SoxRS- <i>E. coli</i> ; BLAST results indicate that MXAN_1454 shares some similarity with an ABC-type transport system involved in Fe-S cluster assembly (CBK97053.1 from <i>Eubacterium siraeum</i> DSM 15702), which indicates a role in oxidative stress response
		PQM17	29.35	hypothetical protein Npun_F1891	29	<i>Nostoc punctiforme</i> PCC 73102	Cyanobacteria: Nostocales	Inducibility abolished in SoxRS- <i>E. coli</i> ; Unable to determine a function for genes

Inducer	Library	Clone	Fold Induction	Genes/Proteins (Accession)	% A.A. Identity	Organism	Taxonomy (Phylum: Order)	Comments
		PQA4	30.40	(ACC80546.1), transcriptionally fused to GFP Fjo21 (AEI68364.1) upstream of and toward GFP	52	<i>Myxococcus fulvus</i> HW-1	Proteobacteria: Deltaproteobacteria	Inducibility abolished in SoxRS- <i>E. coli</i> ; 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 matches to the A4 clone include DoxX family proteins, an obscure putative oxidoreductase.
		PQG8	33.66	oxidoreductase molybdopterin binding protein (GI:269785296)	33	<i>Sphaerobacter thermophilus</i> DSM 20745	Chloroflexi: Sphaerobacteridae	Inducibility abolished in SoxRS- <i>E. coli</i> ; In addition to the oxidoreductase, this clone shares some similarity with a putative glutathione S-transferase from <i>Acaryochloris marina</i> MBIC11017 (plasmid pREB1).
		PQG13	55.11	2-deoxy-D-gluconate 3-dehydrogenase (EEQ93510.1), transcribed divergently from GFP	58	<i>Ochrobactrum intermedium</i> LMG 3301	Proteobacteria: Alphaproteobacteria	Two-fold inducible in SoxRS- <i>E. coli</i>
		PQK19	42.97	Npun_F1891 (ACC80546.1), transcriptionally fused to GFP	30	<i>Nostoc punctiforme</i> PCC 73102	Cyanobacteria: Nostocales	Nearly entire ORF for Npun_F1891 is present; this gene is similar 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	<i>Haliscomenobacter hydrossis</i> DSM 1100	Bacteroidetes: Sphingobacteria	Divergently transcribed ORF contains a PRK10572 conserved domain (DNA-binding transcriptional regulator AraC); in <i>Haliscomenobacter hydrossis</i> , 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.

Inducer	Library	Clone	Fold Induction	Genes/Proteins (Accession)	% A.A. Identity	Organism	Taxonomy (Phylum: Order)	Comments
MERC		PQL10	8.99	signal peptide protein thioredoxin (YP_411745), ORF divergent from GFP; TonB-like protein upstream from thioredoxin (ABB74352.1)	66	<i>Nitrosospira multififormis</i> 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	<i>Methylotenera mobilis</i> JLW8	Proteobacteria: Betaproteobacteria	PAS/PAC sensor is a membrane bound protein that likely senses oxygen-related stress.
		PQD7	18.20	putative ATP-dependent RNA helicase (YP-001658093.1), transcriptionally fused to GFP; upstream, transcribed in the same direction, is hypothetical protein MAE_30780 (YP-001658092.1)	29	<i>Microcystis aeruginosa</i> NIES-843	Cyanobacteria: Chroococcales	ATP-dependent RNA helicase is similar to CRISPR-associated Cas3 helicase; the upstream hypothetical protein, 4 rounds of PSI BLAST, appears to be related to putative transposases.
	EXP-1	PQK9	313.77	<i>merR</i> (divergent) and <i>merT</i> (transcriptionally fused to GFP)	99	<i>Nitrosomonas europaea</i>	Proteobacteria: Betaproteobacteria	Inducible by mercury, but discovered by paraquat induction, where paraquat contained trace amounts of Hg ²⁺ as determined by ICP-MS. Potential for novel bioreporter system.

¹ Induction measured by FLUOStar Optima
ND = no data available

3.3.3 Induction of PAH-E and PAH-B Libraries with Aromatic Compounds

A mixture of LMW aromatic compounds was initially used to interrogate the PAH-E library. A scheme similar to differential fluorescence induction (DFI) (Pothier et al., 2007; Rediers et al., 2005) was employed for cell sorting in which the top 1% of GFP-expressing cells was collected from the total library in the presence of inducers (see Figure 3.3A illustrating each round of FACS sorting for aromatic-inducible clones). This sub-library was then applied to the cytometer in the absence of inducer, and cells exhibiting the least GFP expression (bottom 10%) were collected (Figure 3.3B). Finally, this sub-library was again induced and the cells expressing GFP at the highest levels (top 0.1% and 1%, separately) were sorted (Figure 3.3C) and plated for analysis. Clones were confirmed as inducible using FCM analysis (Figure 3.5).

Figure 3.3. Histograms of GFP expression during different rounds of flow cytometric sorting of aromatic-inducible clones from the PAH-E library. Arrows depict which population was collected by cell sorting for use in the next round of SIGEX. **(A)** Sorting of inducible and constitutive mixed populations: the total library applied to the FCM in the presence of LMW aromatics. Rare GFP expressing cells (within arrow) were sorted and grown for use in the next step. **(B)** Elimination of constitutive clones: cells sorted in A were applied to the FCM with no inducers added. The significant GFP-fluorescent population (right peak) represents constitutively GFP-expressing cells from the original library. The population below the arrow putatively included mainly inducible cells, although in practice there were still many false positives that required removal in the final step. **(C)** Collection of inducible clones: the non-GFP expressing population that was sorted from B was used in the final sort, in which the inducible clones were enriched relative to non-inducible clones by sorting the population indicated by the arrow in the presence of LMW aromatics.

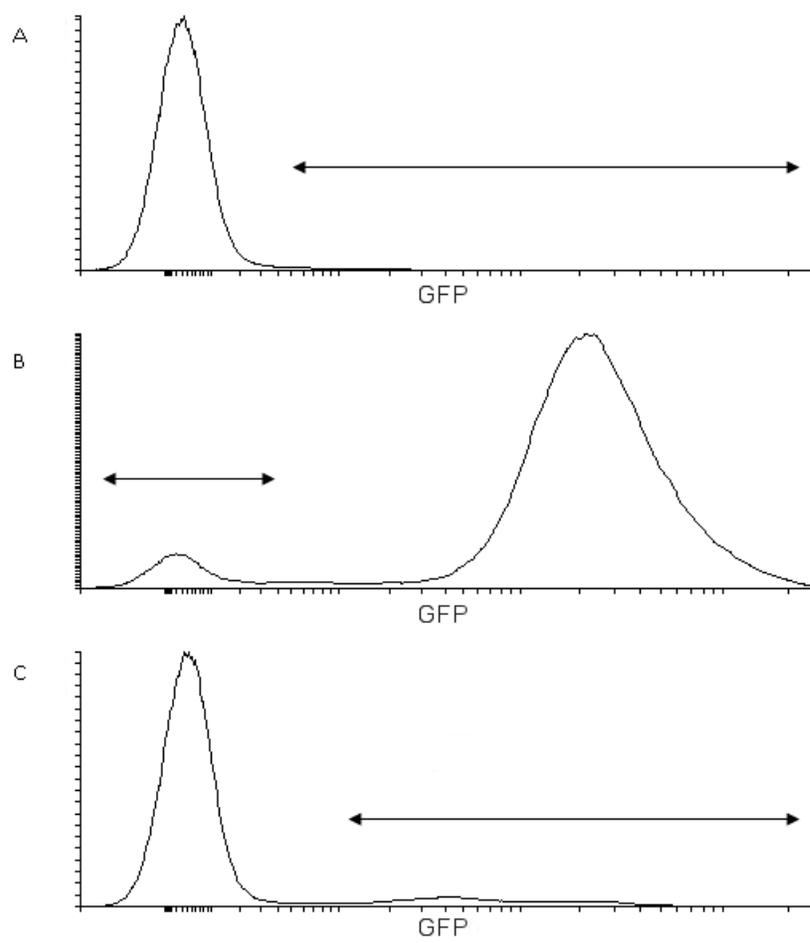
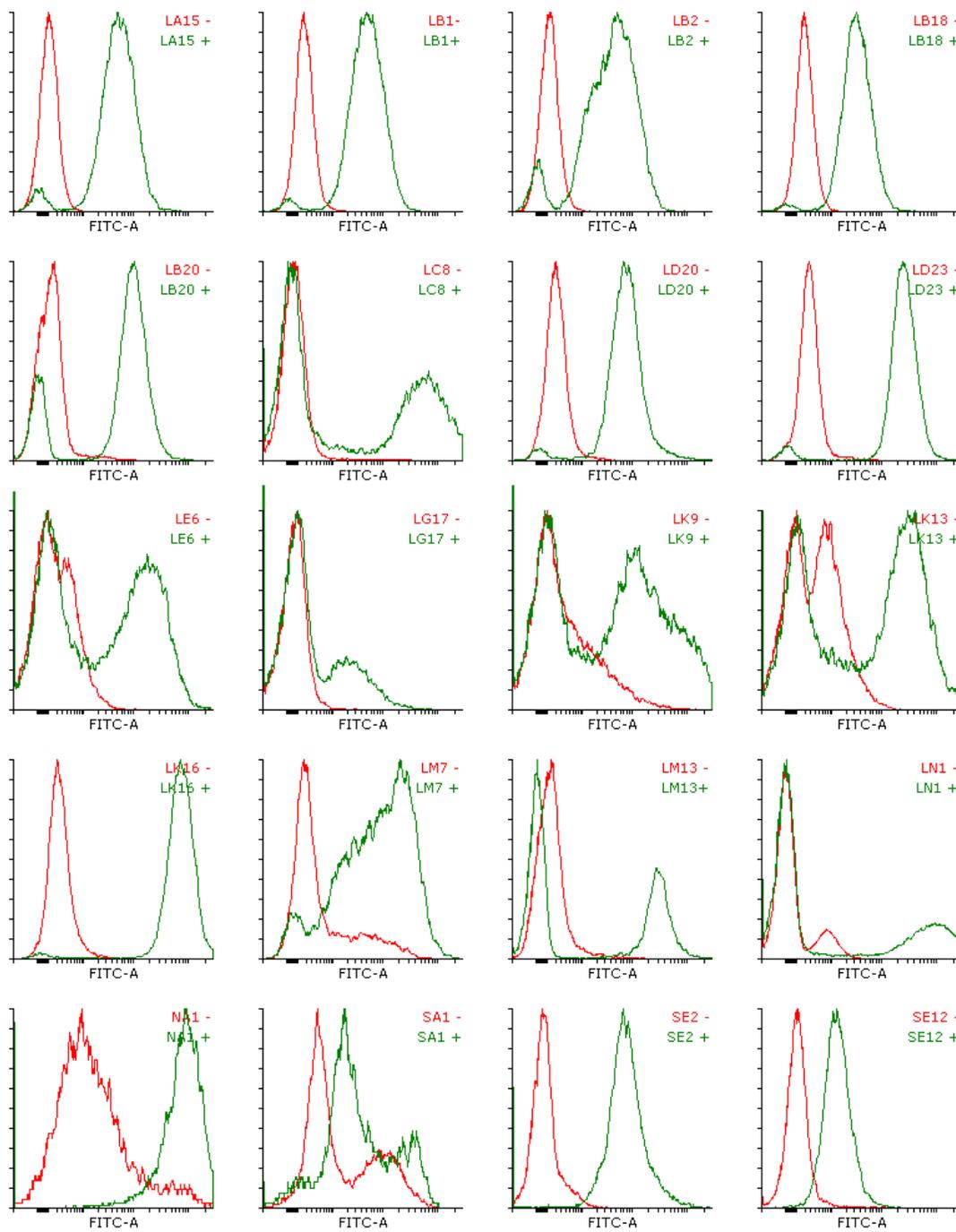
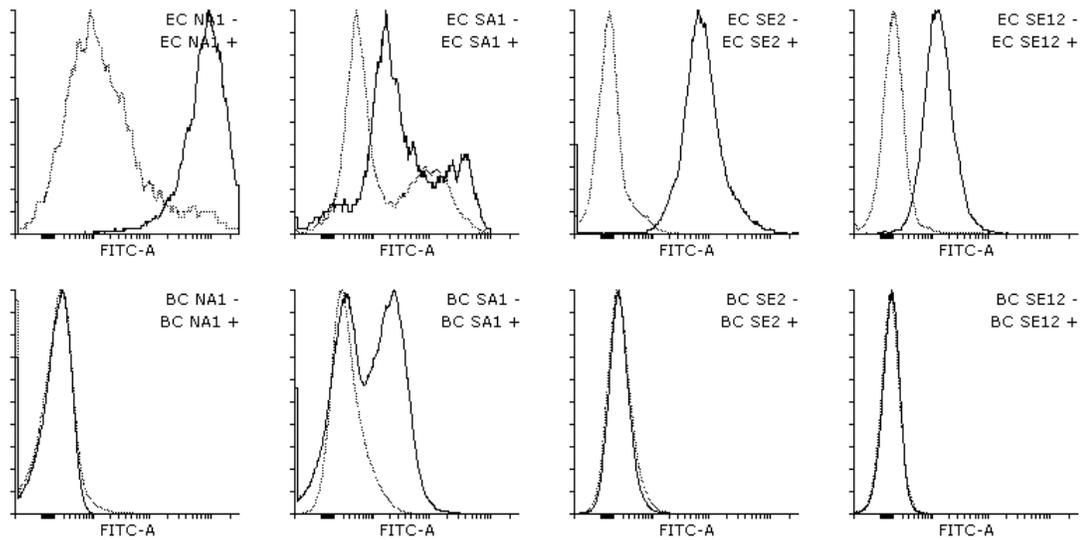


Figure 3.4. Histograms of GFP expression for aromatic-inducible clones isolated from the PAH-E metagenomic library. Red lines indicate cultures grown with no inducer while green lines indicate cultures grown in the presence of 100 μ M LMW aromatic mixture (containing equimolar quantities of benzoate, salicylate, catechol, phenol, phenylacetic acid and naphthalene). Cultures were grown in *E. coli* at 37 °C for 18 h in dM9 before the addition of inducer; inductions were carried out for 3 h under identical growth conditions. Clone names are shown in the top right of each panel.



The use of a Gram-positive host organism may be beneficial for screening genes that do not function in *E. coli* (Uchiyama & Miyazaki, 2010; Yun & Ryu, 2005). However, no inducible clones were recovered from the PAH-B library using the Gram-positive host *B. cereus*. In lieu of this, a *B. cereus* host was transformed directly with several LMW-aromatic-inducible clones recovered using SIGEX from the PAH-E library in *E. coli*, and subsequently measured their inducibility in *B. cereus* using FCM (Figure 3.5). Three of 4 clones recovered using the *E. coli* host did not show any induction in the presence of LMW aromatic compounds in *B. cereus*. However, one clone, SA1, was also inducible in the *B. cereus* host (4.13-fold in *B. cereus* compared to just 1.8-fold in *E. coli* under the same growth conditions).

Figure 3.5. Induction of naphthalene- and salicylate-inducible clones in *Bacillus* (bottom row) and *E. coli* (top row) hosts. The clones were recovered using SIGEX with an *E. coli* host; only clone S-A1 appears to be inducible in *Bacillus*.



3.3.4 Analysis of Aromatic-Inducible Genes Recovered from the PAH-E Library

Three-hundred eighty-four putative LMW aromatic-inducible clones recovered from the PAH-E metagenomic library using SIGEX were tested for inducibility in microtitre plates. Of these, the 96 most inducible were subjected to restriction digestion of plasmid DNA to determine uniqueness. Twenty distinct clones, ranging in size from 1.1 to ~7 kb were end-sequenced (analysis is shown in Table 3.4). The inducibility of these clones was examined using FCM (Figure 3.4) and the microtitre plate assay (Figure 3.6). The sequences revealed that most clones contain high sequence similarity to aromatic-degrading genes or operons, often derived from plasmid sequences, from the genus *Pseudomonas* (14 of the 20 clones). No aromatic-inducible clones were recovered that aligned to sequences from phyla outside of Proteobacteria, and 16 of the 20 clones aligned to Gammaproteobacteria sequences.

Table 3.4. Summary of genes, identified by tBLASTx¹ within LMW-aromatic inducible clones, recovered from the PAH-E library and their induction by 100 μ M LMW aromatics as measured by FCM. Genes putatively regulating the promoters driving GFP expression are indicated.

Clone	Fold Induction	Similar Proteins (Accession)	E-value (Bit Score)	Organism	Class	Comments
NA1	11.86	XylU gene (CAC86830.1) transcriptionally fused to GFP; putative transposases (AF043544.2, AF043544.2) upstream.	0.0 (692)	<i>Pseudomonas putida</i> plasmid pWWO	Pseudomonadales	Partial Toluene Degradation operon (James & Williams, 1998).
SA1	1.81	NahT chloroplast ferredoxin-like protein (AAP44221.1) transcriptionally fused to GFP; NahG (AAP44222.1) salicylate hydroxylase and OprD (NZ_AJMR01000068.1) upstream.	0.0 (719)	<i>Pseudomonas</i> sp. ND6 plasmid pND6-1	Pseudomonadales	Aligns to plasmid pND6-1, a naphthalene degradation plasmid (Li et al., 2004). Contains NahG gene and degrades salicylate to catechol in <i>E. coli</i> .
SE2	60.15	SgpA (ACO92374.1) first 13 bp of gene are transcriptionally fused to GFP; SgpR (ACO92380.1) divergently transcribed.	0.0 (941)	<i>Pseudomonas putida</i> plasmid pAK5	Pseudomonadales	SgpA is a salicylate 5-hydroxylase ferredoxin reductase, the first gene in a salicylate-gentisate pathway; SgpR, an LTTR, is the putative regulator of the operon.
SE12	9.80	PSF113_1991 (AEV62003.1) transcriptionally fused to GFP; upstream is a MarR-family protein (AEV62002.1).	1e-122 (215)	<i>Pseudomonas fluorescens</i> F113	Pseudomonadales	MarR regulates multiple antibiotic resistance through upregulation of efflux pumps, etc., and is often activated by AHs (Martin et al., 1996; Roldan et al., 2008).
LA15	52.36	acriflavin resistance protein (ADQ85831.1) transcriptionally fused to GFP	3e-102 (359)	<i>Methylovorus</i> sp. MP688	Methylophilales	Efflux pump involved in antibiotic resistance.
LB1	29.50	acriflavin resistance protein (ADQ85831.1) transcriptionally fused to GFP	9e-126 (457)	<i>Methylovorus</i> sp. MP688	Methylophilales	Efflux pump involved in antibiotic resistance.
LB2	41.59	acriflavin resistance protein (ADI29216.1) transcriptionally fused to GFP	3e-174 (617)	<i>Methylothenera versatilis</i> 301	Methylophilales	Efflux pump involved in antibiotic resistance.
LB18	17.89	acriflavin resistance protein (AEM51786.1) transcriptionally fused to GFP	1e-127 (365)	<i>Stenotrophomonas maltophilia</i> JV3	Xanthomonadales	Efflux pump involved in antibiotic resistance.
LB20	47.16	NdsR (BAC53588.1) transcribed divergently from GFP; promoter driving GFP expression is from a salicylate hydroxylase (BAC53589.1)	7e-126 (457)	<i>Pigmentiphaga</i> sp. NDS-2	Burkholderiales	NdsR is a putative LTTR; part of a salicylate degradation operon.
LC8	69.24	DDE-type transposase (AFM32558.1) transcriptionally fused to GFP; upstream is a LTTR transcribed toward GFP (AFM32556.1); divergently transcribed is NahG.	0.0 (690)	<i>Pseudomonas stutzeri</i> CCUG 29243	Pseudomonadales	Similar to a transposon involved in naphthalene degradation.
LD20	37.91	NahG (AAD02146.1) transcriptionally fused to GFP; divergently transcribed is NahR (AAD02145.1)	0.0 (549)	<i>Pseudomonas stutzeri</i> (GI:4104761)	Pseudomonadales	GFP is fused to the first gene in this salicylate degradation operon (Bosch et al., 1999).

Clone	Fold Induction	Similar Proteins (Accession)	E-value (Bit Score)	Organism	Class	Comments
LD23	80.28	Salicylate hydroxylase (AAY21679.2) transcriptionally fused to GFP; divergently transcribed NahR (AAY21678.2)	0.0 (899)	<i>Pseudomonas fluorescens</i> strain PC20 plasmid pNAH20	Pseudomonadales	GFP is fused to the first gene in this salicylate degradation operon; similar plasmid pNAH20 (Heinaru et al., 2009).
LE6	37.34	SalA (AAZ08064.1) transcriptionally fused to GFP; divergent BphR2 (AAZ08063.1).	0.0 (650)	<i>Pseudomonas pseudoalcaligenes</i> KF707	Pseudomonadales	Naphthalene degradation operon; cross regulated <i>in vivo</i> by BphR1 and BphR2 (Watanabe et al., 2003; Fujihara et al., 2006).
LG17	17.39	Acyl carrier protein phosphodiesterase (ABA73175.1) fused to GFP; membrane protein (ABA73174.1) upstream; partial LTTR protein (ABA73173.1) divergently transcribed.	4e-173 (355)	<i>Pseudomonas fluorescens</i> Pf0-1	Pseudomonadales	Putative antibiotic resistance genes. The acyl carrier protein contains a conserved azoreductase domain (PRK00170; E-value 7.06e-98) and a flavodoxin-like fold (pfam02525; E-value 3.95e-57); the LTTR contains a MarR domain.
LK9	16.00	LTTR (AEY00105.1) transcribed divergently from GFP; Further downstream and also divergent is an IclR type regulator (YP_004931645.1).	6e-148 (531)	<i>Oceanimonas sp.</i> GK1	Aeromonadales	Inducible by phenol. LTTR shows similarity to BenR from <i>Marinobacter hydrocarbonoclasticus</i> ATCC 49840 (YP_005428200.1). In <i>Oceanimonas</i> , a muconate and chloromuconate cycloisomerase (AEY00106.1) is downstream of the promoter (Harwood & Parales, 1996).
LK13	28.93	Salicylate hydroxylase NahG (AAD02146.1) transcriptionally fused to GFP; divergently transcribed NahR (AAD02145.1)	0.0 (200)	<i>Pseudomonas stutzeri</i> (GI:4104761)	Pseudomonadales	Aligns to the same sequence as clone LD20 but represents a unique restriction fragment (Bosch et al., 1999).
LK16	255.48	Salicylate hydroxylase NahG (AAQ89673.1) transcriptionally fused to GFP; NahR (AAQ89672.1) divergently transcribed.	0.0 (708)	<i>Pseudomonas putida</i> (GI:37220701)	Pseudomonadales	First gene in salicylate degradation gene cluster and NahR; this element is widely distributed among biphenyl-utilizing bacteria (Nishi et al., 2000).
LM7	7.19	Proteins NarD (AAG34371.1), NarK (AAG34372.1), and NarG (AAG34373.1) are divergently transcribed from GFP; only the promoter region of NarX (AAG34370.1) is present.	0.0 (697)	<i>Pseudomonas fluorescens</i> (GI:11344596)	Pseudomonadales	Putative nitrate/nitrite transporter (<i>narD</i>), putative nitrate/nitrite transporter (<i>narK</i>), and respiratory nitrate reductase alpha subunit (<i>narG</i>). NarX is a nitrate/nitrite sensor protein (Stewart & Parales, 1988; Rabin & Stewart, 1992).
LM13	61.69	NahG salicylate hydroxylase (ACV05012.1) transcriptionally fused to GFP and NahR (ACV05020.1) is divergent.	0.0 (928)	<i>Pseudomonas aeruginosa</i> strain CGMCC 1.860 plasmid	Pseudomonadales	Nearly complete NahG gene is present. Naphthalene degradation operon; aligns to same sequences as LN1 but at a different site.
LN1	211.36	NahG salicylate hydroxylase (ACV05012.1) transcriptionally fused to GFP and NahR (ACV05020.1) is divergent.	0.0 (878)	<i>Pseudomonas aeruginosa</i> strain CGMCC 1.860 plasmid	Pseudomonadales	Naphthalene degradation operon; aligns to same sequences as LM13 but at a different site. Higher fold induction may be due to the closer proximity of GFP to the promoter.

¹ cutoff E-value of 1e-5 was used; the E-values and bit scores reported correspond to the best match for each set of BLAST hits.

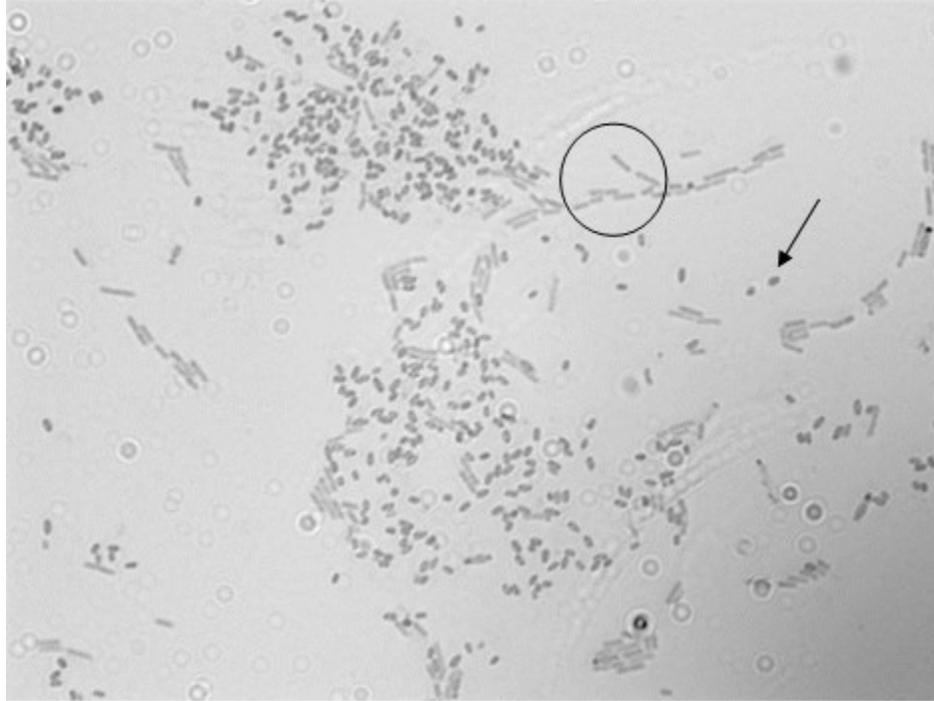
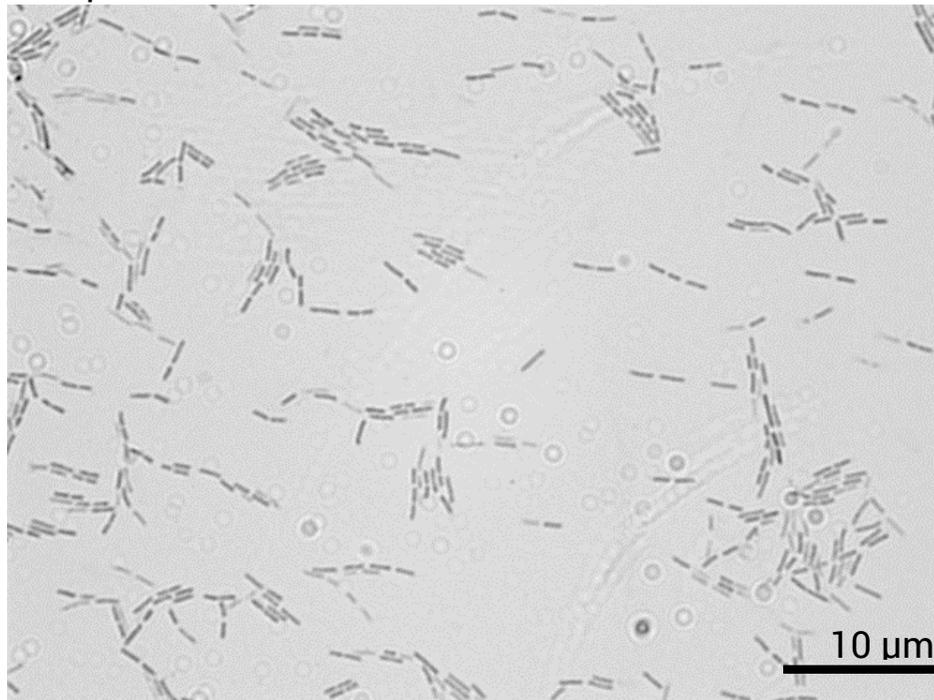
Figure 3.6. Induction of aromatic-inducible clones in microtitre plates using a variety of LMW aromatic compounds. Clones are inducible to different extents relative to one another for each chemical tested here. The y-axes are scaled to an appropriate height depending on the level of induction, which is different for each compound. Error bars indicate standard deviation of four replicates.

Based on tBLASTx searches of the Sanger reads of SIGEX clones, the most common sequences – found in the clones SA1, LD20, LD23, LK13, LK16, LM13, and LN1 – show similarity to *nahG* (salicylate hydroxylase) coupled with *nahR* (an aromatic-inducible LysR-type transcriptional regulator [LTTR]). Multiple efflux transporters (LA15, LB1, LB2, LB18, and SE12), a transposase (LC8), and a variety of other genes that encode proteins with putative metabolic functions (NA1, SE2, SE12, LB20, LE6, LG17, LK9 and LM7) were also found. A detailed analysis of these sequences is displayed in Table 3.4.

3.3.5 Creation of a Non-Sporulating *spo0A* Mutant in *Bacillus* 6A5

Transformation of *Bacillus* 6A5 with pMUTIN4 vector carrying the *spo0A* gene gave rise to colonies resistant to Em. Images of the strain before and after gene interruption are shown in Figure 3.7, demonstrating that sporulation was eliminated after inactivation of the *spo0A* gene.

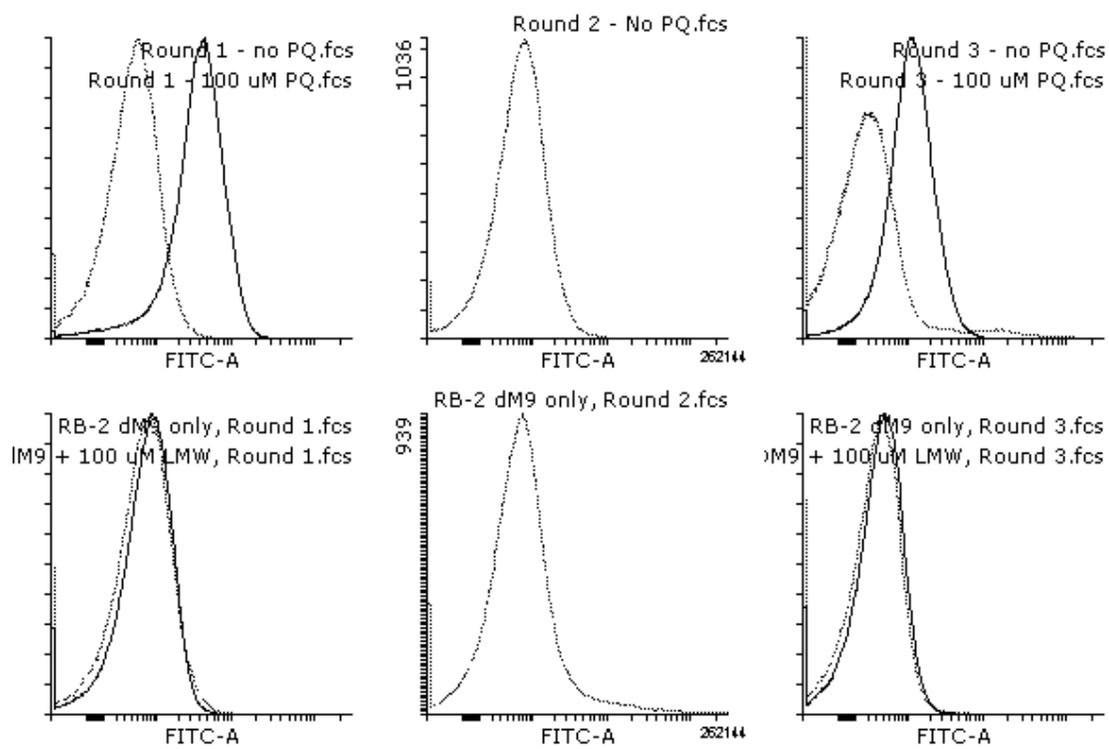
Figure 3.7. Spore stain of *Bacillus* 6A5 wild type (A) and 6A5 *spo0A*⁻ (B) cells plated on dLB (with Em for *spo0A*⁻ strain, to maintain selection on the gene knockout cassette) and incubated for 4 d at 37 °C. In this black and white image, the malachite-green stained spores are seen as small, dark, oval-shaped cells (arrow) relative to the elongated rod-shaped vegetative cells (circle).

A. Wild-type *Bacillus***B. Asporulant *Bacillus***

3.3.6 Inductions of *Bacillus* 6A5 *spo0A*⁻ Rock Bay Metagenomic Library

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 (Figure 3.8).

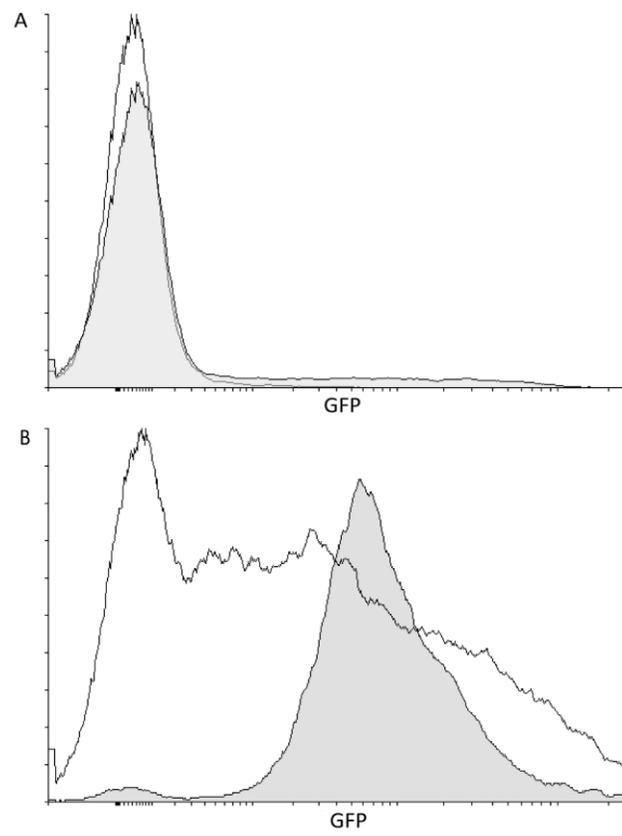
Figure 3.8. Sequential rounds of FCM sorting on the *Bacillus* 6A5 PAH-B SIGEX library produced no apparent enrichment of inducible clones using paraquat (top row) or LMW-aromatics (bottom row). Contrast this with the significant increase in LMW-aromatic inducible clones visible in Figure 3.3 during *E. coli* sorting. Middle panels (column 2) should be highly fluorescent but do not show this characteristic. As shown in columns 1 and 3 (top pane), the addition of paraquat seems to increase auto-fluorescent characteristics of *Bacillus* (solid lines). Solid vs. dotted lines indicate the presence vs. absence of inducer, respectively.



3.3.7 Trends in GFP Expression

FCM analysis of cultures demonstrated that GFP expression of individual cells within a clonal population can vary dramatically. Figure 3.9 shows histograms of GFP expression for four different clones isolated from the PAH-E metagenomic library (all shown in an uninduced state); the only difference between them is the metagenomic promoter driving GFP expression. As shown in Figure 3.9A, a typical population (white) will express GFP in a Gaussian distribution. However, the pattern of GFP expression depends strongly on the genetic nature of the metagenomic sequence: with some clones, nearly-identical expression patterns may exhibit vastly different means (Figure 3.9A). The population shown in grey (with a mean of 2158) has a higher mean than the population shown in white (with a mean of 80), the difference arising from the former exhibiting a long tail of uniformly GFP-expressing cells over a large dynamic range. Conversely, highly variable expression patterns can yield nearly identical means (Figure 3.9B, light grey and white have means of 16114 and 16204, respectively). Figure 3.9B demonstrates that the variation of GFP expression between cells within a population is a function of some unknown genetic determinants.

Figure 3.9. Populations of GFP-expressing cells from genetically different clones (measured using FCM) can show varied patterns of expression that would be undetectable with single-measurement techniques. **(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).



3.3.8 HMW Aromatic Inducers

We attempted to use HMW PAH inducers (individually and in a mixture, at concentrations of 10 μ M) in SIGEX experiments with the PAH-E and PAH-B libraries. None of these experiments yielded clones that were inducible by the PAHs. In total, analysis of >1152 clones recovered with SIGEX using HMW PAH inducers was performed, but only false positives (clones expressing GFP unconditionally) were recovered.

3.3.9 Mapping SIGEX Clones to NGS-Derived Scaffolds

Each of the Sanger reads from SIGEX clones aligned to contigs in the *de novo* assembled NGS metagenome sequence with an E-value of 0.0 in the initial BLAST search (Table 3.5). Since there were multiple hits for each Sanger read, the contig with the longest high-scoring pair (HSP) to the SIGEX clone was used as the reference sequence for mapping in Geneious v. 6.1. A graphical overview of SIGEX clones mapped to NGS-derived contigs is shown in Figure 3.10. The predicted biological roles determined for the contigs using ORF prediction (MetaGeneMark; Zhu et al., 2010) and subsequent BLAST searches, are shown in Table 3.5. Detailed analysis of the contigs is shown in Appendix A.

Figure 3.10. SIGEX clones mapped to NGS-derived scaffolds that were assembled *de novo* using IDBA-UD. This figure demonstrates the ability to use NGS for obtaining a context for SIGEX clones that includes upstream and downstream metagenomic sequence.

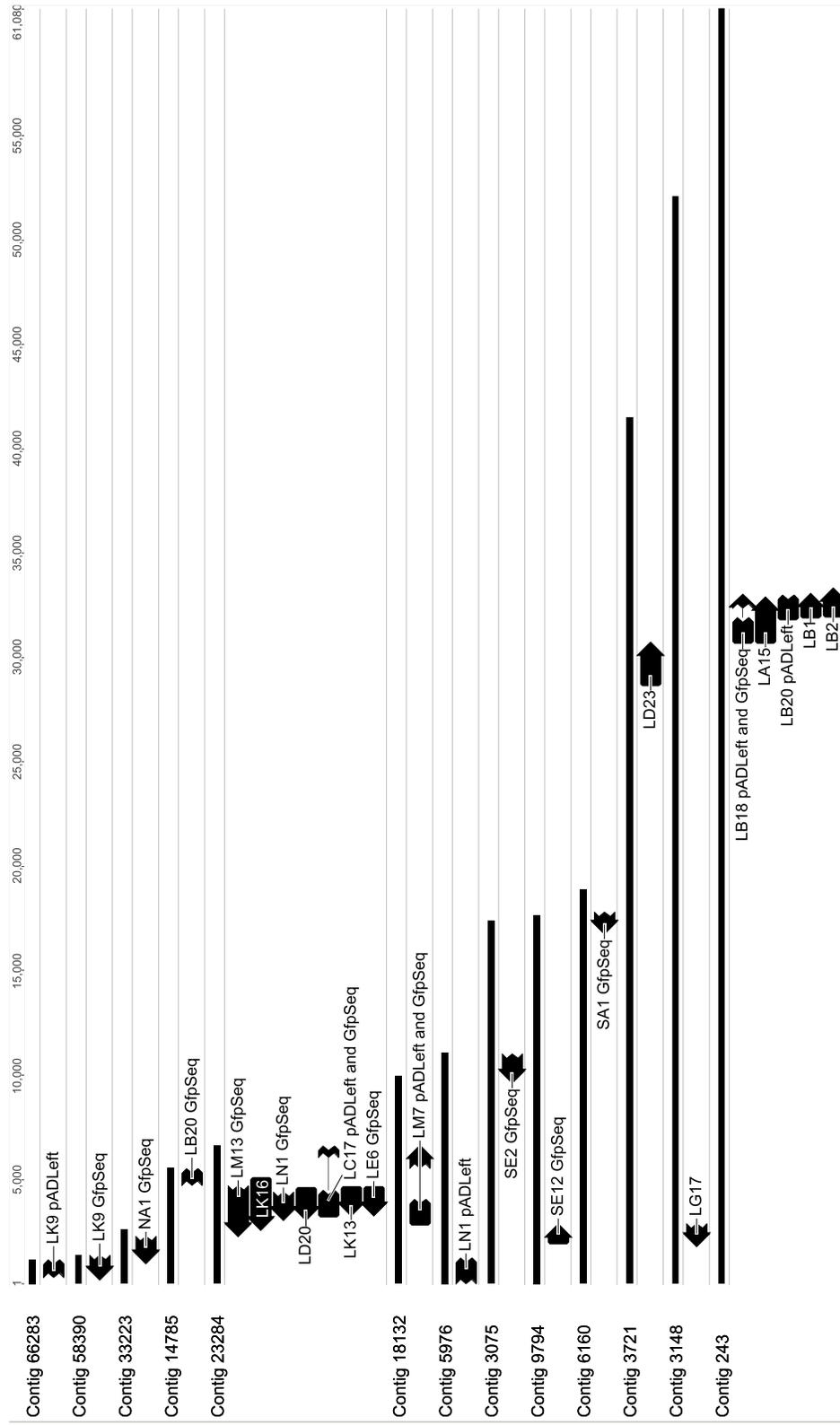


Table 3.5. BLAST results showing matches between SIGEX-recovered clones and scaffolds in IDBA-UD assembled Illumina reads.

Clone	Best Match ¹ to Contig (length in bp)	Pairwise %ID	Predicted Features on Matched Scaffold (Species)
LA15	Contig 243 (61080)	95.6	Operon containing several efflux pumps (<i>Methylothera versatilis</i> 301)
LB1	Contig 243 (61080)	97.2	See LA15
LB18 (GfpSeq)	Contig 243 (61080)	100.0	See LA15
LB18 (pADLeft)	Contig 243 (61080)	87.0	See LA15
LB2	Contig 243 (61080)	97.8	See LA15
LB20 (GfpSeq)	Contig 14785 (5488)	97.9	Type VI secretion protein, nitroreductase, and several hypothetical proteins downstream from DntR/NahR/LinR regulator (<i>Pigmentiphaga</i> sp.)
LB20 (pADLeft)	Contig 243 (61080)	95.5	See LA15
LC17 (GfpSeq)	Contig 23284 (6565)	87.2	Salicylate degradation gene cluster (<i>Pseudomonas stutzeri</i> CCUG 29243 chromosome)
LC17 (pADLeft)	Contig 23284 (6565)	89.9	See LC17 (GfpSeq)
LC8	Contig 23284 (6565)	ND	See LC17 (GfpSeq)
LD20	Contig 23284 (6565)	89.9	See LC17 (GfpSeq)
LD23	Contig 3721 (41449)	81.2	Operon containing sequence from <i>Pseudomonas putida</i> G7 plasmid pNAH7; benzoate transport proteins (<i>Pseudomonas</i> sp.), and partial xyl operon containing various aromatic oxygenases (<i>Azotobacter vinelandii</i> strain DJ)
LE6	Contig 23284 (6565)	89.3	See LC17 (GfpSeq)
LG17	Contig 3148 (52066)	97.0	Contains a histidine kinase protein downstream from putative MarR regulator and ACP phosphodiesterase / azoreductase (<i>Pseudomonas brassicacearum</i> NFM421 and <i>Pseudomonas mandelii</i>)
LK13	Contig 23284 (6565)	89.4	See LC17 (GfpSeq)
LK16	Contig 23284 (6565)	91.8	See LC17 (GfpSeq)
LK9 (GfpSeq)	Contig 58390 (1336)	100.0	LTTR (<i>Oceanimonas</i> sp.) and partial transposon (<i>Pseudomonas pseudoalcaligenes</i>)

LK9 (pADLeft)	Contig 66283 (1142)	95.1	Similar to a Fis family regulator (<i>Pseudoxanthomonas spadix</i>) and transposase Tra8 (<i>Pseudomonas pseudoalcaligenes</i>)
LM13	Contig 23284 (6565)	83.8	See LC17 (GfpSeq)
LM7 (GfpSeq)	Contig 18132 (9903)	96.7	Nitrite and nitrate sensor, transporter, and reductase proteins NarG, NarU, NarK, NarX/Q, NarL (<i>Pseudomonas mandelii</i> and <i>Pseudomonas</i> sp.)
LM7 (pADLeft)	Contig 18132 (9903)	83.4	See LM7 (GfpSeq)
LN1 (GfpSeq)	Contig 23284 (6565)	91.0	See LC17 (GfpSeq)
LN1 (pALeft)	Contig 5976 (10627)	86.9	N/A; apparent chimeric sequence
NA1 (GfpSeq)	Contig 33223 (2562)	95.1	Nitrotoluene catabolic gene cluster; similar to xyl genes from TOL plasmid (<i>Pseudomonas</i> sp. Strain TW3)
SA1 (GfpSeq)	Contig 6160 (18816)	95.0	Lower naphthalene-degrading pathway for catabolism of salicylate to acetyl CoA and pyruvate (<i>Pseudomonas putida</i> strain NCIB 9816-4, plasmid pDTG1, bases 31917 to 51950 except for one gene)
SE2 (GfpSeq)	Contig 3075 (17374)	97.9	Salicylate/gentisate degradation gene cluster (<i>Pseudomonas putida</i> AK5, plasmid pAK5)
SE12 (GfpSeq)	Contig 9794 (17614)	99.3	Operon containing efflux transporter and multidrug resistance proteins (<i>Pseudomonas</i> sp. GM78)
Average		92.4	

¹ BLAST hits were sorted by E-value, and hits with longest query coverage were used, since all hits had at least one match with an E-value of 0.0.

3.4 Discussion

Several key technical modifications from the original description of SIGEX (Uchiyama & Watanabe, 2008; Uchiyama et al., 2005) were applied in the current study. Uchiyama et al. (2005) used an inducible *lac* promoter to screen out clones that were self-ligated, and therefore expressed GFP in the presence of IPTG; however, the *lac* promoter exhibits leaky transcription and therefore their process also removed clones weakly expressing GFP in the absence of chemical inducer, which is a characteristic observed in many aromatic degradation operons (de Lorenzo, 2005). Our promoterless vector achieves a lower background level of GFP expression, and enhances the dynamic range available for the detection of GFP-expressing cells.

The creation of a non-fluorescent library as described originally (Uchiyama et al., 2005) is not always the best approach. It was found that a scheme similar to DFI gave the best results in these experiments, where the first stage of cell sorting includes the collection of induced as well as constitutively expressed clones from the complete metagenomic library; constitutive clones are removed in subsequent rounds of sorting. It is also advantageous to collect fractions expressing high vs. low levels of GFP (Dunn et al., 2003) because they yield genes with different levels of induction. The DFI-type scheme, which incorporates a third round of cell sorting to enrich the inducible clones, resulted in the best purity of recovered inducible clones.

Several advantages are gained by using FCM in metagenomic screens. Primarily, the literature has focused on the high-throughput nature of FCM and its

ability to analyze millions of clones very rapidly (Handelsman, 2005; Taupp et al., 2011). Another advantage – single cell analysis of gene expression (Hermans et al., 2011; Rediers et al., 2005) – relates to the measurement of differences between populations with unusual distributions. As shown in Figure 3.9, populations with similar mean values can have extraordinarily different population shapes and different median values. This is especially relevant in metagenomic analysis because promoters in an environmental context are sensitive to alterations in effector molecules, as well as overarching global regulator proteins; in fact, most environmental genes are regulated as part of complex circuits involving several regulators (Cases & de Lorenzo, 2005). This means that individual cells in the population, which might, for instance, be in different growth phases (*i.e.*, expressing different σ -factors), can activate transcription at the same promoter to different degrees (de Las Heras et al., 2012). This concept is exemplified by Newman & Shapiro (1999), where it was shown that differential gene expression can be modulated by variations in overarching regulatory dynamics within clonal populations of *E. coli*. These differences can be detected using FCM analysis of transcriptional activation, but would not be evident using qPCR or other types of average-measurement techniques.

We have found that SIGEX is limited in several ways, many of which were previously addressed by de Lorenzo (2005). Generally, it is possible that the TFs or promoters necessary for transcription do not function in *E. coli*. This could be due to improper protein folding or a lack of mRNA expression of the TF. Furthermore, as pointed out by de Lorenzo (2005), the substrate of a given pathway

is not always the cognate inducer, and this is manifested in our results via the observation that salicylate – a common intermediate and known inducer of aromatic degrading pathways (Park et al., 2005a) – was the most common and most potent inducer among aromatic inducible clones. Library sizes can also impose significant constraints, given that hundreds of species are likely present but it is only possible to analyze a fraction of their genomes. Although our library sizes are larger (by approximately 10-fold) than the initial SIGEX report (Uchiyama et al., 2005), they still represent only a small fraction of the total metagenomic sequence present in our soil samples. Although a range of inducible clones was detected using a variety of AHs, HMW compounds such as fluoranthene, pyrene, and phenanthrene did not give rise to inducible clones in any of our screens. The mechanistic reason for this could be that the transcription factors (TFs) involved in the activation of genes encoding HMW-metabolizing enzymes have a distal location relative to the promoters they regulate. This is likely, given the recent insights into the genetic arrangements of PAH-degrading islands such as the *phn* island (Hickey et al., 2012), where transcriptional factors implicated in the regulation of terminal dioxygenases are located several kb downstream of the promoters they putatively act upon. Other factors that may limit the recovery of certain clones using the SIGEX scheme include the necessity of proper directionality (*i.e.*, the inducible promoter must be oriented toward GFP) and the possibility that different levels of

regulation may be responsible for expression of the genes of interest (e.g., post-transcriptional regulation such as antisense RNA).

Another limitation imposed by plasmid-based metagenomic screens is that the insert size is often insufficient to determine the original genomic context. In this chapter, it was shown that the SIGEX clones can be effectively mapped to contigs derived from shotgun-sequenced metagenomic DNA (Figure 3.10). The genes analyzed on these contigs often align to entire operon structures (e.g., contig 3075 aligns to the *Pseudomonas* plasmid pAK5 salicylate/gentisate degrading operon) in the GenBank database, demonstrating that it is possible to use NGS to obtain relevant information about upstream and downstream sequences that are not retrieved using SIGEX by itself. The in-depth analysis of these contigs, and the characterization of their features encompassed by the sequence found upstream and downstream of the loci where SIGEX clones were mapped, is the subject of Chapter 4.

Aromatic-inducible genes recovered from the Rock Bay metagenome were 1) induced by at least one LMW aromatic compound (Figure 3.6), and 2) derived from *Pseudomonas* or closely related genera in Proteobacteria (Table 3.4). Each inducible clone recovered showed high similarity to genes that are known components of aromatic metabolic processes, including: various oxygenases (NA1, SA1, SE2, LB20, LD20, LD23, LE6, LK9, LK13, LK16, LM13, and LN1), antibiotic resistance or efflux mechanisms (SE12, LA15, LB1, LB18), transposons carrying genes associated with aromatic degradation (LC8, LM13), and miscellaneous genes (LG17, LM7). Clones encoding metabolic functions, such as salicylate or

naphthalene oxygenases, typically possessed only partial genes – often the first or second gene downstream from the putative promoter. Upstream of the promoter, oriented in the opposite direction, ORFs were often found to share high similarity to TFs that likely respond to inducers and regulate the promoters (*i.e.*, *nahR*). However, only one clone, SA1, was observed to produce metabolic byproducts *in vivo*. This was inferred from the production of a dark brown chemical (most likely catechol) following the addition of 1 mM salicylate to a culture of SA1. Therefore, in most cases, the actual metabolic genes would remain undetected in classical enzyme function screens.

The notion that many aromatic catabolic genes are induced strongly by intermediates such as salicylate is consistent with existing studies (Gottfried et al., 2010; Lönneborg & Brzezinski, 2011; Park et al., 2005a). TFs that respond to salicylate and other LMW aromatics have been described extensively in both Gram-negative and Gram-positive organisms (*i.e.*, the LTTR *nahR* for *Pseudomonas putida* G7 (Park et al., 2002); the GntR-type regulator *narR1* in *Rhodococcus opacus* R7 (Di Gennaro et al., 2010)). There is a notable absence of sequences derived from Gram-positive organisms in the PAH-E metagenomic clones. Even after utilizing the shuttle vector pMMeb, derived from pAD123 (Dunn & Handelsman, 1999), to create the PAH-B library in *B. cereus*, no Gram-positive derived inducible sequences were found. However, taxonomic analysis derived both from the NGS data (Chapter 5) and from 16S rDNA/DGGE studies (Rose, 2010) suggests that Rock Bay soil

contains a very high proportion of Proteobacteria and very low proportions of Gram-positive bacteria.

In this chapter it has been shown that, for simple inducing compounds (*i.e.*, LMW-aromatics) and Gram-negative hosts, SIGEX is an effective way to recover biologically relevant DNA sequences in a high throughput manner. The variety of gene classes identified (*e.g.*, related to metabolism, cellular efflux, transcriptional regulation, and antibiotic resistance) suggests that SIGEX may be a useful phenotypic screen for a wide range of metagenomic targets. We have made several key technical modifications, including improved library construction from fragmented metagenomic DNA, and improved FCM sorting procedures. Moreover we show that the distribution of fluorescence of single cells within a population may provide a mechanism for evaluating increases in gene expression that are not possible using average-measurement techniques. However, we have found that there are certain circumstances, for example in the analysis of HMW aromatics, where SIGEX is ineffective. Finally, SIGEX may be particularly effective when used in concert with NGS analyses that provide further insight into the function of surrounding metagenomic sequences (Chapter 4).

Chapter 4.

Characterization of Aromatic-Inducible Operons in a Contaminated Soil Metagenome using Next-Generation Sequencing

4.1 Introduction

Aromatic hydrocarbons (AHs) are widespread soil contaminants that can be degraded, albeit slowly, by a variety of microorganisms. Many studies have examined aromatic-degrading microbes and their functional genes through culture-based techniques (Kanaly & Harayama, 2000; Kweon et al., 2007; Pinyakong et al., 2003); however, since most strains are resistant to culture, these studies describe a relatively small proportion of environmentally relevant microbial species (Vartoukian et al., 2010). Next-generation sequencing (NGS) provides new opportunities for microbial ecologists to examine the entire metagenome (Xia et al., 2013), and has already been used to investigate a range of soils, including several hydrocarbon-contaminated sites (Bell et al., 2011; Delmont et al., 2012a; Yergeau et al., 2012b). Results from such studies suggest that extremely high sequencing depth is required to capture the diversity of microorganisms present in soil, especially in environments where species richness is high.

An impressive number of computational tools are now available to analyze metagenomic sequences. The objectives of these algorithms include *de novo*

assembly (e.g., IDBA-UD; Peng et al., 2012), targeted *de novo* assembly (e.g., PRICE; Ruby et al., 2013; Stenglein et al., 2012), phylogenetic analysis (e.g., RDP; <http://rdp.cme.msu.edu>), and gene identification/classification (e.g., MG-RAST, <http://metagenomics.anl.gov>; Meyer et al., 2008) (see Wackett, 2012 for a comprehensive list of available resources). While the capacity of computational methodology is expanding rapidly, traditional genetic and biochemical characterization of microbial communities cannot provide *in vitro* or *in vivo* corroboration fast enough to keep pace with *in silico* predictions derived from sequence data. Thus, there is a requirement for high-throughput methods that can identify functionally relevant genetic elements from a metagenome and provide biological support for computational analyses.

The degree to which results from *in silico* predictions overlap with functionally relevant genes in a particular sample remains relatively unexplored in real metagenomic samples. To this end, we have studied a PAH-contaminated soil metagenome using a combined approach. Initially, we used a phenotypic screen (SIGEX; Uchiyama et al., 2005) to search for DNA fragments that were upregulated by various aromatic compounds as determined by expression of a GFP reporter gene (Chapter 3). Next, we used Illumina HiSeq platform NGS data to identify assembled contigs containing the cloned sequences, allowing us to characterize the genomic context surrounding functionally relevant elements. We discuss here the degree of congruence within these results (phenotypic *vs.* *in silico* methods), and show that the majority of clones recovered using SIGEX map with very high similarity to contigs assembled *de novo* from the whole-metagenome shotgun

sequence. Moreover, we show that the surrounding sequence on these contigs reveals complete operons that contain genes that are likely involved in aromatic hydrocarbon degradation.

4.2 Methods

4.2.1 Contaminated Soil Treatment

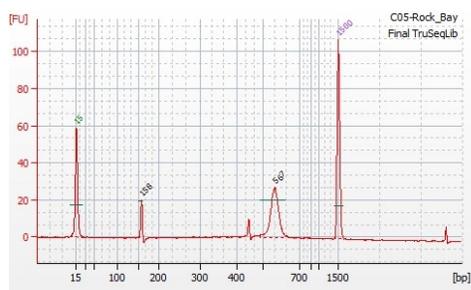
Soil that was heavily contaminated with PAHs (475 – 12,600 $\mu\text{g/g}$ soil; Whynot, 2009) and related AHs was collected from Rock Bay (Victoria Harbour, British Columbia, Canada; samples provided by BC Hydro and Transport Canada). The soil was homogenized before incubation in a 20% w/v bioslurry in duplicate BioFlo 110 bioreactors; the slurry was agitated and aerated constantly for 90 days at a pH of 6.5 – 8.0, at a constant temperature of 25 °C. Samples of the bioslurry, taken every 15 days beginning on day 0, were stored at -80 °C following the addition of glycerol to 15%.

4.2.2 Illumina Sequencing

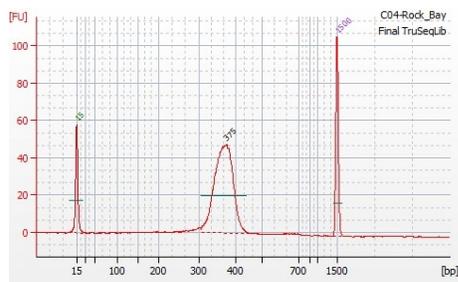
To obtain metagenomic DNA for sequencing, 10 slurry samples (stored as glycerol stocks) were combined and centrifuged to collect a total of 3 g soil. Soil was washed in buffer (10 mM EDTA pH 8.0, 50 mM Tris pH 8.0, 50 mM Na_2HPO_4) and DNA was isolated using the MoBio PowerMax Soil DNA Isolation kit. DNA was further purified by phenol/chloroform extraction and polyethylene glycol precipitation (Chapter 3). Approximately 7 μg of metagenomic DNA was used to create two TruSeq gDNA libraries (performed by Genome Québec, Montréal, Québec). The libraries had average insert sizes of 255 bp and 447 bp (Figure 4.1). Two lanes of paired-end Illumina sequencing (using the HiSeq platform) were obtained from the 255 bp insert library, and one lane from the 447 bp library, using 100 nucleotide paired-end reads.

Figure 4.1. Quality control data from the production of the 447 bp (A) and 255 bp (B) TruSeq gDNA libraries by Génome Canada. Histograms show counts of the number of fragments at various nucleotide lengths (in base pairs). Insert sizes were calculated by removing the length of the adapter sequences.

A. 447 bp inserts



B. 255 bp inserts



4.2.3 SIGEX Experiments

In Chapter 3, we describe the recovery of aromatic-inducible genetic elements from a PAH-contaminated soil using a promoter trap method. Briefly, metagenomic DNA was cloned upstream of a promoterless GFP on the plasmid pMMeb. The metagenomic library was transformed into an *E. coli* host and clones expressing GFP in the presence, but *not* absence, of inducing compound (benzoate, salicylate, naphthalene, phenol, phenylacetic acid, and catechol) were sorted using a FACSaria II flow cytometer. We tested these clones for cross-inducibility and expression under different conditions. GFP fluorescence was measured using a FLUOStar Optima (BMG Labtech) fluorescent plate reader fitted with filters for excitation at 485 nm and emission at 520 nm; gain was set at 750. Clear microtitre plates with 96 (Corning) or 384 (Fisher) wells containing dLB supplemented with the appropriate antibiotic (200 μ L or 100 μ L, respectively) were inoculated by 1:100 dilution from a log-phase culture.

Clones recovered using SIGEX were end-sequenced (Chapter 3). The Sanger reads were aligned manually using BioEdit and vector sequence was removed. Any regions of overlap between reads were used to build a consensus sequence for each SIGEX-recovered clone. These sequences were used to query the nr protein NCBI database using tBLASTx with a cutoff of 1e-5. They were also analyzed with MetaGeneMark (Zhu et al., 2010) and ORF Finder, and the translated ORFs were used to query the nr protein NCBI database using BLASTp with a cutoff of 1e-5.

BLAST results were analyzed with Epos BlastViewer and Geneious (Biomatters Ltd.).

4.2.4 Assembly and Analysis of Illumina Sequence Data

4.2.4.1 *De Novo* Assembly

IDBA-UD (Peng et al., 2012) was run on an Intel dual i7 6-core processor (total of 24 logical cores at 2.0 GHz each) with 192 GB of RAM (BioLinux 7). The kmer sizes were iterated from 20 to 100 with a step size of 1 or 10, following a pre-correction step using a kmer size of 60. CLC Assembly Cell was run on an Intel i7 6-core processor (total of 12 logical cores each at 3.2 GHz, BioLinux 6) and 64 GB of RAM with a kmer size of 31.

4.2.4.2 Annotation of Metagenomic Sequences

The locations of genes on Rock Bay metagenome contigs were determined with MetaGeneMark (<http://exon.gatech.edu/metagenome/Prediction/>; Zhu et al., 2010) using the “mixture of bacteria and archaea” kingdom and the 6-LBA codon model. MetaGeneMark-predicted genes were used to annotate contigs in Geneious. Sequences of predicted genes were tested for similarity to the nr and conserved domain database (CDD) using BLASTx with an E-value cutoff of 1e-1; annotations were transferred to the contigs using Geneious. InterProScan (Quevillon et al., 2005) was used to find protein domains (using the translated ORFs from MetaGeneMark) within all available protein databases. Figures for annotated contigs are shown in Appendix A. Annotation tables are available in Appendix B.

The cBar plasmid annotation software was used to determine if each contig was more likely derived from a plasmid or chromosome sequence based on nucleotide pentamer frequencies (Zhou & Xu, 2010).

4.2.4.3 MG-RAST Analysis

MG-RAST (Meyer et al., 2008) was used to determine functional and taxonomic relationships between metagenomic sequence reads and database sequences (<http://metagenomics.anl.gov/>; metagenomic rapid annotation by subsystems technology). MG-RAST is an online annotation service that uses the SEED algorithm in a standardized pipeline analysis of metagenomic DNA sequences. We uploaded all six FASTQ files obtained from Illumina sequencing to the MG-RAST server, as well as several metagenomic assemblies in FASTA format. FASTQ reads were processed using the default MG-RAST parameters. Coverage of the assembled sequence, which enhances the MG-RAST analysis, was computed for each contig from the IDBA-UD contig output file using the formula: coverage = (number of reads on contig \times 100 bp) / (contig length in bp). SEED subsystem annotations were used to identify any features annotated with 'Metabolism of aromatics'; these reads were assembled using IDBA-UD to obtain contigs putatively containing only aromatic-metabolizing genes.

4.2.4.4 Targeted Assembly using PRICE

PRICE was used to assemble metagenomic sequences directly flanking a known subset of the metagenome. This software uses paired-end information to extend an initial set of contigs or sequences using the raw NGS reads (Ruby et al.,

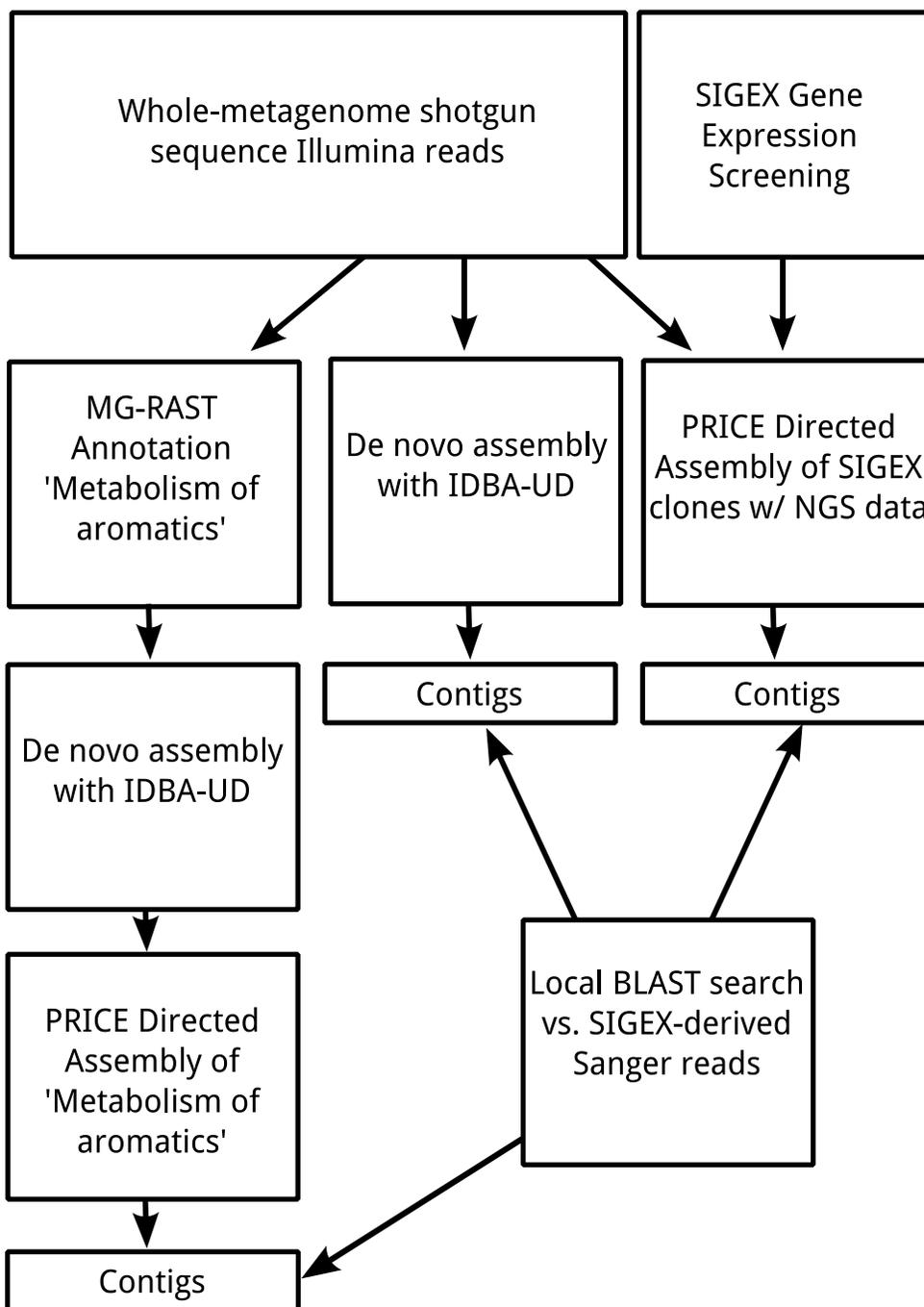
2013). We used 20 cycles of PRICE, run using default parameters, to expand the SIGEX clone sequences (Sanger reads). A second subset of the metagenome, the ‘Metabolism of aromatics’ reads (as annotated by MG-RAST and pre-assembled using IDBA-UD), was also extended using 20 cycles of PRICE.

4.2.5 Mapping SIGEX Clones to Metagenomic Contigs

A local BLAST database was created from the FASTA file of IDBA-UD-assembled contigs, enabling rapid queries to identify contigs of interest within the NGS data. We used this BLAST database to determine the amount of similarity and coverage the SIGEX clones shared with NGS-derived sequences, using the SIGEX clones as queries with an E-value cutoff of $1e-5$ in the BLASTn program. After sorting hits by E-value, the hit with the highest coverage was used to determine which contig was used for mapping that clone.

The contigs derived from IDBA-UD assembly were loaded into Geneious and, based on search results of the local BLAST database, each SIGEX clone Sanger sequence was aligned to its respective best-match contig. The “map to reference” feature of Geneious was used to align the Sanger sequences to the contigs, with up to 5 iterations of the highest sensitivity. Where multiple SIGEX clones mapped to the same contig, alignment with MUSCLE (Edgar, 2004) was performed and Geneious was used to create dendograms and visualize the alignments to examine locations of overlap and polymorphism. This workflow of analysis (Figure 4.2) was repeated for contigs derived from PRICE assembly directed by either ‘Metabolism of aromatics’ or SIGEX Sanger read sequences.

Figure 4.2. Flowchart illustrating the workflow carried out for the analysis of metagenomic DNA sequences.



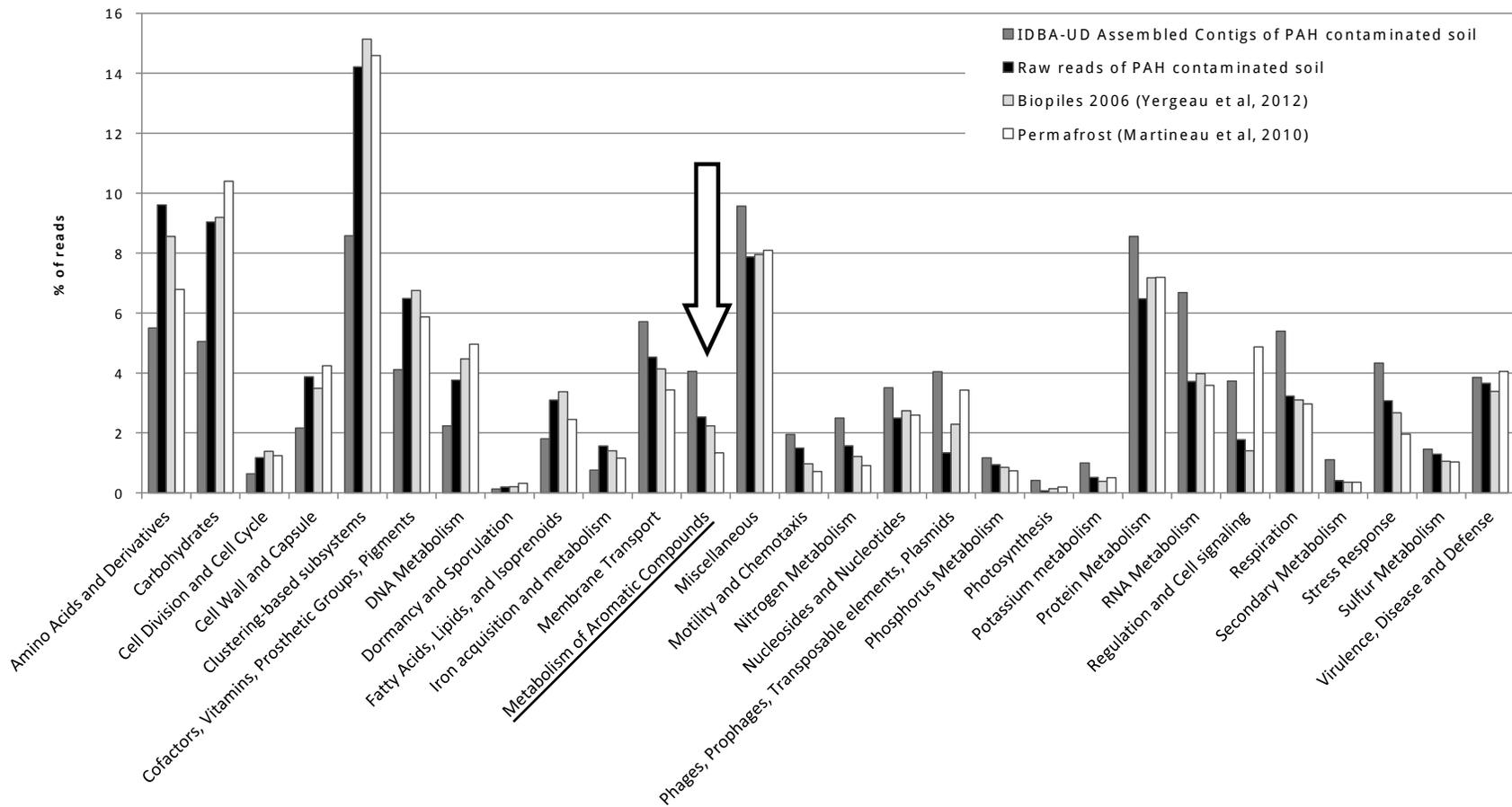
4.3 Results

4.3.1 MG-RAST Analysis

Raw Illumina sequence data was uploaded to the MG-RAST server and is available under the accession numbers 4494328.3, 4494329.3, 4494326.3, 4494327.3, 4494330.3, and 4494331.3. IDBA-UD assembled contigs with coverage information, but without length QC, are available under the accession 4514941.3; scaffolded contigs (also known as supercontigs), which lack coverage information, are available under 4514465.3.

In our MG-RAST datasets, genes annotated as “Aromatic metabolism” by the SEED Subsystems analysis represent approximately 2.5% of the total features identified using raw reads; however, when analysis of IDBA-UD-assembled contigs was done using the same algorithm, we find that this proportion increases to 4.1% (Figure 4.3, arrow). This demonstrates enrichment of aromatic metabolism subsystems when compared to a pristine soil (1.3%; Martineau et al., 2010), and is comparable to another hydrocarbon-contaminated metagenome (2.2%; Yergeau et al., 2012a).

Figure 4.3. Comparison of SEED subsystems from MG-RAST pipeline analysis demonstrates that the PAH-contaminated soil metagenome NGS reads (black bars) are enriched for features annotated as ‘Metabolism of aromatics’ relative to a pristine soil (white bars). This is accentuated even further when assembled contigs derived from the NGS reads are used in the pipeline (dark grey bars). Even relative to other hydrocarbon contaminated sites (light grey bars), the Rock Bay site shows significant enrichment of aromatic hydrocarbon-degrading genes.



4.3.2 *De Novo* Assembly of Metagenomic Sequences

We attempted *de novo* assembly with several programs using various parameters to compare the effectiveness of each. The statistics associated with the contigs generated by each program are shown in Table 4.1. We found that IDBA-UD provided the highest N50, with a value of 9.1 kb when scaffolds were included. The longest contig was 608 kb, and the dataset contains 143 contigs ≥ 100 kb. 25% of the assembled sequence length is contained in the 1998 contigs ≥ 26804 bp.

Table 4.1. Statistics of various *de novo* assemblies of Illumina-sequenced metagenomic DNA from the Rock Bay PAH-contaminated site.

Data Source	Assembler and Parameters	Starting Bases (Gbp)	N50 (bp)	Longest Contig (kb)	Total Assembled Sequence in Mbp (# 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	406,973,927 (117,116)
	IDBA-UD, kmer 20 to 100, step size 1, %ID 99	125	6,638 (9,121 scaffolded)	608	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)

4.3.3 Directed Assembly of Aromatic-Degrading Features using PRICE

Using PRICE, we directed the partial *de novo* assembly of NGS data to metagenomic regions surrounding either: a) SIGEX-derived clones; or b) the contigs assembled from 'Metabolism of aromatics' annotations in the MG-RAST SEED subsystem classification (Table 4.2). This provided two separate ways to direct the assembly, using either the phenotypically characterized sequences garnered from the SIGEX clones, or a completely *in silico* approach based on annotations in MG-RAST. We found that, after 20 cycles of PRICE assembly, a comparable amount of total sequence was obtained (13.7 Mb for the MG-RAST annotated aromatic assembly and 15.2 Mb for the SIGEX clone directed assembly). However, the MG-RAST annotated reads provided a higher quality assembly, with an N50 of 9.6 kb compared to an N50 of 4.6 kb obtained with the PRICE expansion of SIGEX clones.

Table 4.2. Statistical characteristics of contigs obtained from PRICE directed assemblies.

Data Source for Initial 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 4.1)	1,835 (2,128,001, 1269)	20	1649 (9,587)	63	13.7
SIGEX-recovered Clones Inducible by Aromatic Compounds (Sanger-sequenced)	40 (44,365, 1135)	20	3773 (4,662)	33	15.2

4.3.4 Mapping Aromatic-Inducible Clones to Assembled Metagenomic Contigs

We used end-sequenced Sanger reads from each aromatic-inducible SIGEX clone as queries in BLAST searches of a database derived from the IDBA-UD *de novo* assembled contigs (Table 4.3), and then used Geneious to map the best hits onto each contig, which enabled visualization of where the SIGEX-derived clones fit into a larger context (Figure 4.4). Results presented in the body of this paper are a summary of the findings; the reader is referred to Appendix A for the entire complement of maps. We examined contigs corresponding to only the top hit for each clone. In addition to the IDBA-UD assembly, the SIGEX clones were also mapped to contigs assembled by PRICE (Table 4.3). Overall, the SIGEX clones mapping with highest confidence to NGS contigs were from the unbiased and complete *de novo* assembly performed by IDBA-UD (with an average contig size of 18.4 kb and nucleotide identity of 95.0% for mapped clones). However, in certain individual cases, we found that some SIGEX clones mapped to PRICE-assembled contigs that were longer than those from the IDBA-UD assembly (*e.g.*, clone NA1maps to a 2.6 kb contig from the IDBA-UD *de novo* assembly with 95.1% identity compared to a 12.6 kb contig with 94.3% identity from the PRICE assembly; similarly, LK9-GfpSeq maps to a 1.3 kb IDBA-UD contig with 100% identity, but an 8.8 kb contig in the PRICE assembly with 99.3% identity).

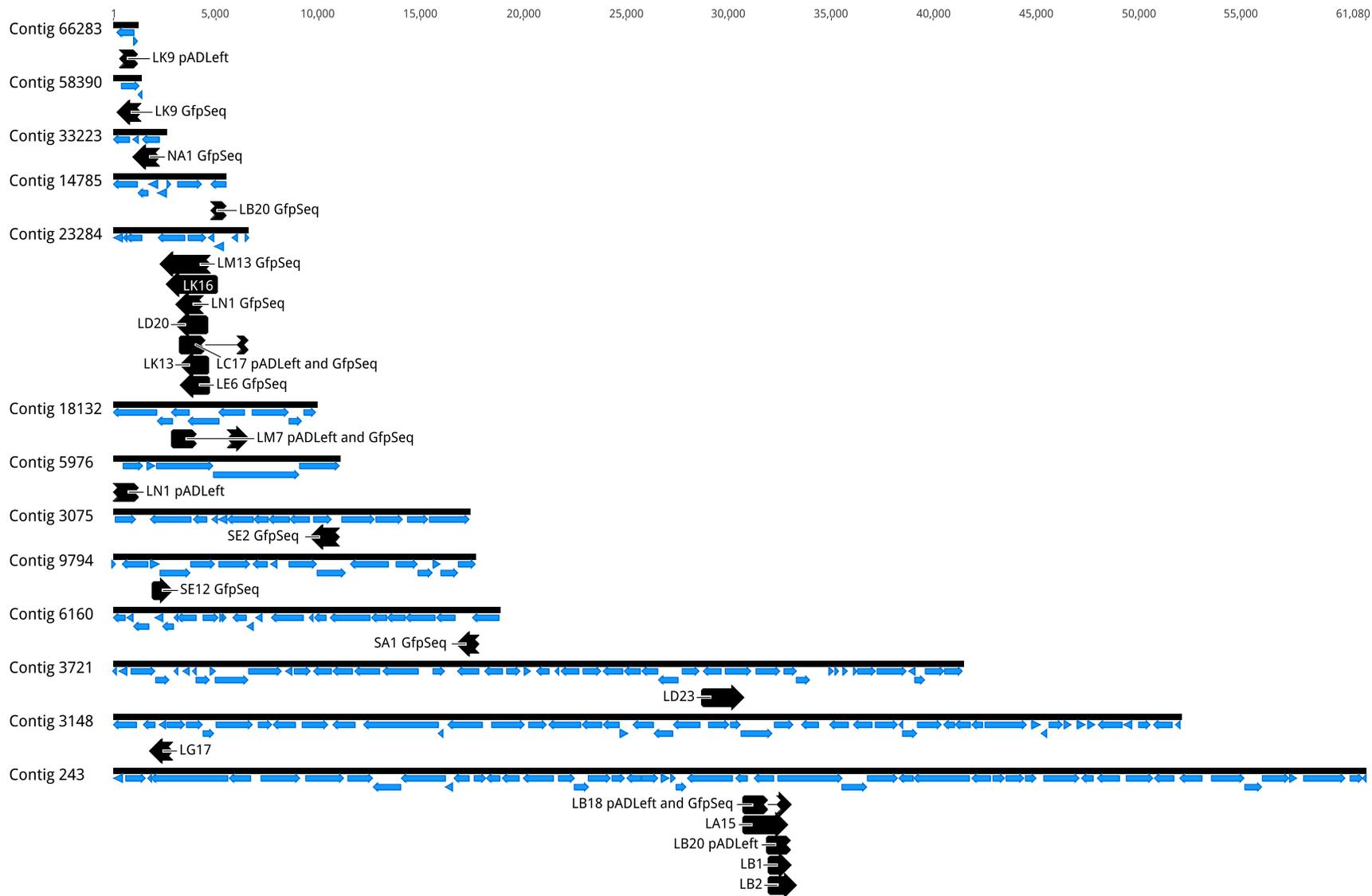
Table 4.3. Matches between SIGEX-recovered clone sequences and contigs assembled *de novo* from NGS data using various approaches.

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	ND	ND

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 (GfpSeq)	contig_3075 (17.4)	97.9	contig_6 (19.7)	97.9	ND	ND
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 or contigs using Geneious v. 6.1

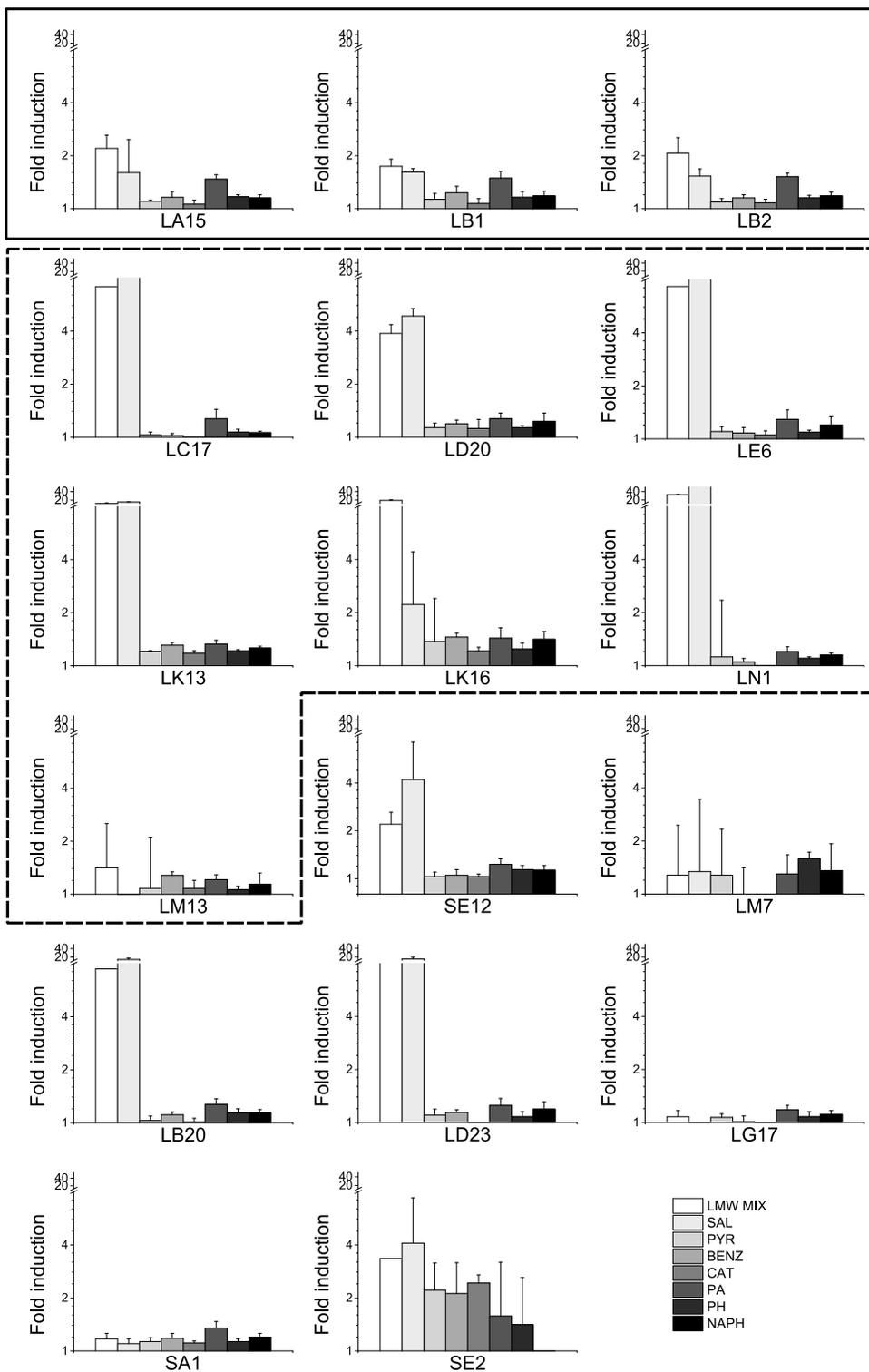
Figure 4.4. Overview of aromatic-inducible SIGEX-recovered clones (thick arrows), showing ORFs predicted by MetaGeneMark (thin arrows). This demonstrates that the relatively small plasmid-based clones can be mapped to a larger genomic context to obtain sequence information downstream and upstream that would otherwise be impossible to obtain using SIGEX, a functional screen, alone.



4.3.5 Determining Biological Roles for Sequences Surrounding SIGEX Clones

The aromatic-inducible SIGEX clones are inducible to different extents by a given chemical, and individual clones also show specific patterns of expression in the presence of different chemicals (Figure 4.5). For example, SE2 exhibited a significant level of expression in the response to multiple chemicals; in contrast, expression in other clones (LN1: salicylate) was more specific. Most clones align very well to contigs. For instance, LA15, LB1, and LB2 aligned with average 96.0 % pairwise identity to contig 243, which increased to 99.7% when only the *acrB* gene and promoter was considered. Slightly less identity was observed in LM13, LK16, LN1, LD20, LC17, LK13, and LE6, which align with an average of 92.3% identity to contig 23284. Interestingly, LA15, LB1, and LB2 share similar GFP expression patterns in the presence of various chemicals; on the other hand, LM13, LK16, LN1, LD20, LC17, LK13, LE6 have several SNPs, and the expression patterns for these clones are more variable, with some (LM13, LK16) even showing reduced sensitivity to salicylate and the LMW-aromatic mixture. In fact, LM13 contains a frameshift mutation in a HTH domain of *nahR* that likely alters the function of the transcriptional activator, while LK16 has a large region deleted from the gene because it is a chimera (see below).

Figure 4.5. GFP induction of aromatic hydrocarbon-inducible SIGEX-recovered clones in microtitre plates using a variety of LMW aromatic compounds at a concentration of 100 μ m. Fold induction relative to an empty vector control is shown as the average of 4 samples, with error bars representing standard deviation. Solid line box: clones map to contig 243. Dashed line box: clones map to scaffold 23284. The y-axes are represented on a log₂ scale that is broken from 8 through 12 to allow comparison between highly variable clones. SAL, salicylate; PYR, pyrene; BENZ, benzoate; CAT, catechol; PA, phenylacetic acid; PH, phenol; NAPH, naphthalene; LMW MIX, an equimolar mixture of the preceding chemicals.

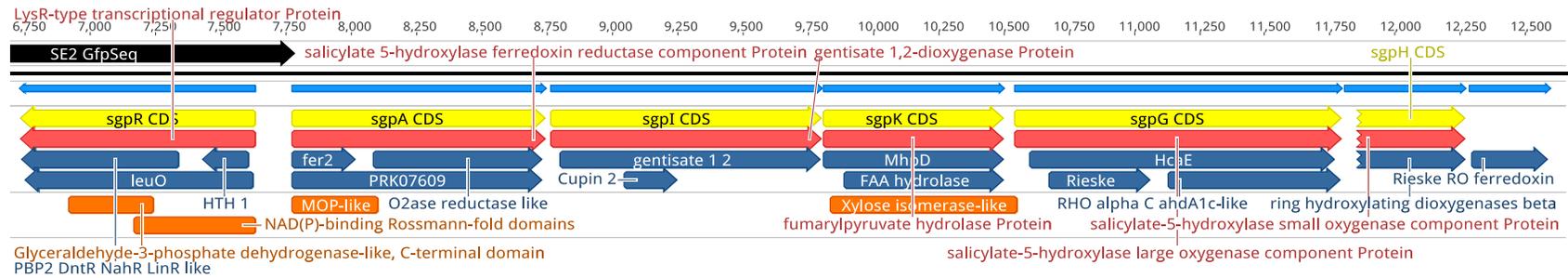


The clone SE2, which mapped to contig 3075, aligned with a LysR type regulator (DntR/NahR/LinR-like) divergently transcribed from GFP, a promoter/intergenic region, and a salicylate-5-hydroxylase ferredoxin reductase gene toward GFP. However, only the first 15 bp of the reductase gene were included in the SIGEX-recovered clone; thus, it represents an example where none of the catabolic genes were captured by SIGEX. However, by mapping it to contig 3075 (Figure 4.6A), 6 genes were identified in the operon downstream of SE2, all of which were highly similar (98-100% a.a. identity) to genes on the plasmid pAK5 from *Pseudomonas putida* (GenBank accession FJ859895; Izmalkova et al., 2013). The possibility that contig 3075 is of plasmid origin was also supported by the cBar plasmid prediction software. Similarly, in clone SA1, the SIGEX-recovered fragment contained a partial operon (Figure 4.6B), but by mapping it to contig 6160 we were able to discover the downstream genes, including catechol dioxygenases and biphenyl degrading genes. This operon was highly similar (100% a.a. identity for the downstream metabolic genes) to a previously characterized operon on pDTG1, another *P. putida* plasmid (Dennis & Zylstra, 2004; Park et al., 2002) from strain NCIB 9816-4. The first 17,447 bases of contig 6160 aligned with 99.9% nucleotide identity to pDTG1, at which point *tnpA* is present on pDTG1, but is absent on the contig. The remainder of contig 6160 matches with 100% nucleotide identity to pDTG1. In this case, cBar failed to report that the sequence was plasmid-derived. Several unique clones (LM13, LK16, LN1, LD20, LK13, LE6 and LC17) aligned to the same region of contig 23284, but with different start and stop points. This contig contained a salicylate degradation cluster (Figure 4.6C and

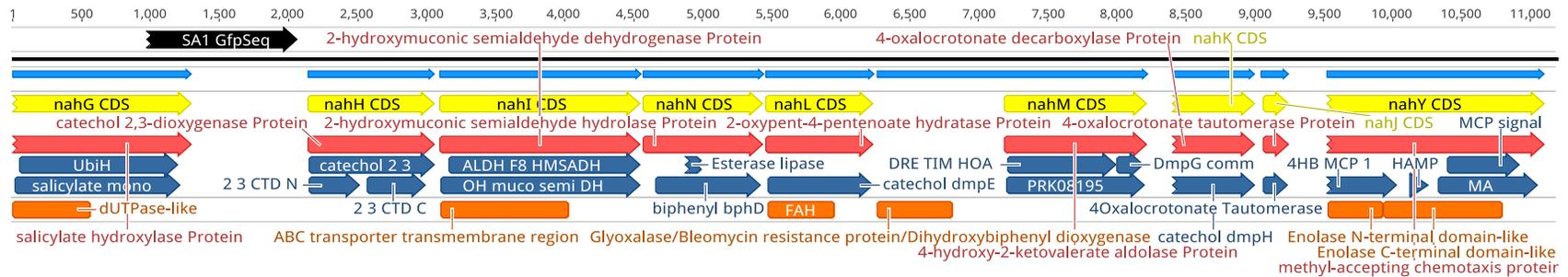
Figure 4.7A). The SIGEX clones terminate at various sites within the first gene in the operon, a salicylate 1-monooxygenase. Although the SIGEX clones align with only this gene, the genes in the operon predicted on contig 23284 that are downstream from the SIGEX clones are most closely related to a chromosomally encoded salicylate degrading operon (including a catechol 2,3-dioxygenase, a 2-hydroxymuconic semialdehyde dehydrogenase, and a chloroplast-type ferredoxin NahT, each at 100% nucleotide identity) found in *Pseudomonas stutzeri* (CCUG 29243).

Figure 4.6. Annotation of aromatic degrading operons on assembled contigs, with catabolic genes found downstream of the mapped aromatic-inducible SIGEX-recovered clones (black arrows). Green arrows indicate ORFs predicted by MetaGeneMark, BLAST and CDD annotations are shown as dark blue arrows; other colors represent various protein domain annotations. **(A)** A portion of contig 3075 aligning to SIGEX clone SE2. The genes in this predicted operon are 98-100% identical to plasmid pAK5, while the sequence of SE2 (aligned to the regulator *sgpR* and intergenic region plus 15 bp of *sgpA*) shares 97.9% nucleotide identity to the contig. The genes predicted on the contig constitute a salicylate-gentisate degradation pathway. **(B)** A portion of contig 6160 aligning to clone SA1. This operon contains several biphenyl-degrading genes from plasmid pDTG1. **(C)** Contig 23284 aligning to multiple aromatic-inducible clones. This salicylate degradation cluster is encoded chromosomally in *Pseudomonas stutzeri* (CCUG 29243).

A. Contig 3075



B. Contig 6160



C. Contig 23284

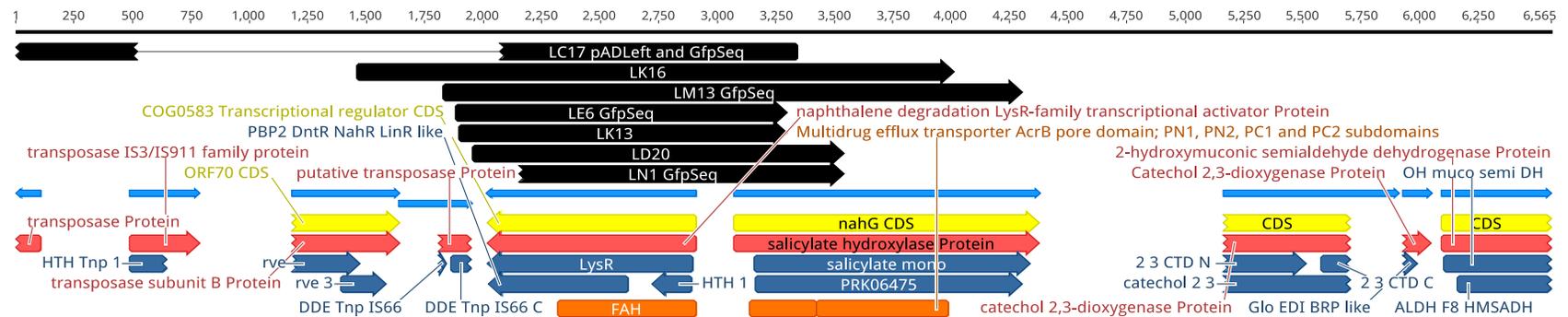
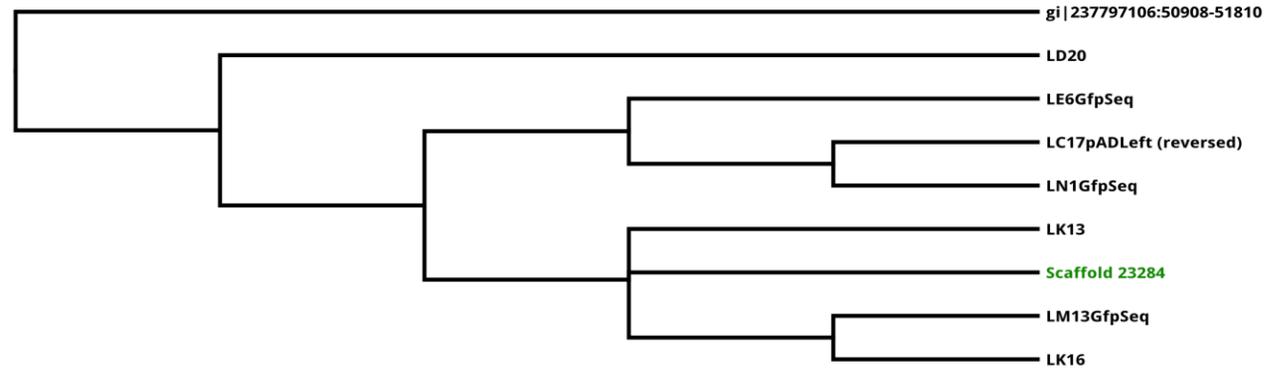
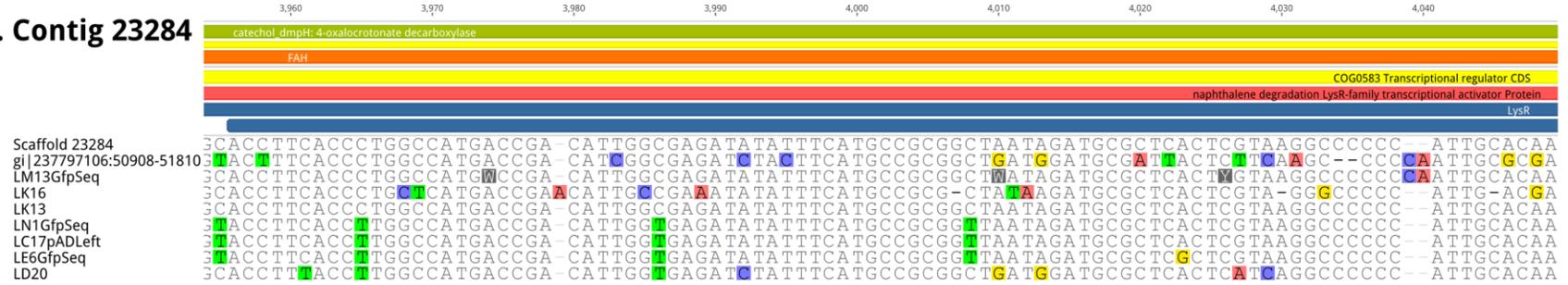
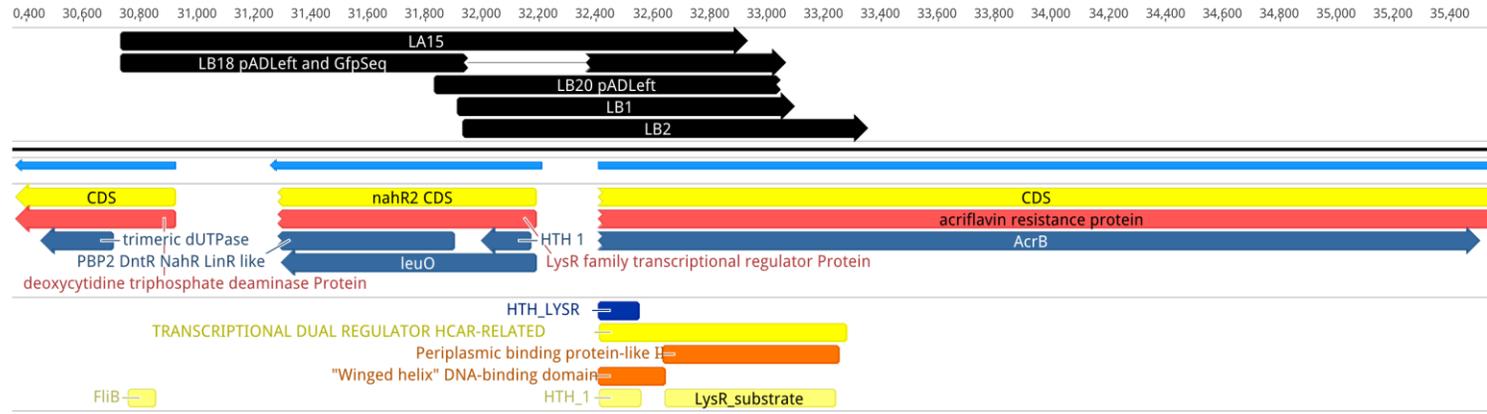


Figure 4.7. SIGEX-derived clones mapping to the same contig likely originated from different individuals within the microbial community. **(A)** SNPs and a dendrogram showing related but distinct genotypes isolated from the metagenome. This likely derives from SNPs between unique strains carrying the same genes within the community. **(B)** Clones LA15 and LB18, black arrows, have the same starting point but end at different points in the gene (right, indicated by arrow head pointing in direction of GFP expression).

A. Contig 23284



B. Contig 243



Clone LD23 aligns with only 81.2% nucleotide identity to contig 3721, possibly indicating that the SIGEX clone represents a rare instance in which the clone does not correspond to assembled NGS sequence. Nonetheless, contig 3721 contains a salicylate degradation cluster with an archetypal gene organization of *nahR* divergently transcribed from the *nahG* salicylate hydroxylase. Downstream of *nahG*, we found several genes exhibiting variable levels of identity (45 % to 95%) to AH degradation/transport genes found in different gammaproteobacterial species, suggesting that this contig may be derived from a hitherto uncharacterized bacterium.

Contigs 3148, 9794, and 243 each apparently contain operons for efflux/drug transport proteins. These are systems that generally represent a stress response and are responsible for removing toxic compounds (typically small molecules like antibiotics) from the cell. Other aromatic-inducible SIGEX clones mapped to contigs that were relatively short (*e.g.*, K9, NA1) or had unknown biological relevance (*e.g.*, LB20 and LM7).

4.3.6 Variation between Clones Mapped to Identical Metagenomic Contigs

In some of the clones, regions of overlap existed such that it was possible to identify SNPs or variants (Figure 4.7A). In the case of contig 23284, we found that 3 clones (LD20, LE 6GfpSeq, and LC17-pADLeft) contained several identical SNPs, and others, such as LM13, contained unique variations. Furthermore, when the overlapping sequences were analyzed by making a dendrogram (using a distantly related *nahR* gene, GI:237797106, to root the tree), the apparent SNP-containing sequences clustered together, indicating a common evolutionary origin. Note that

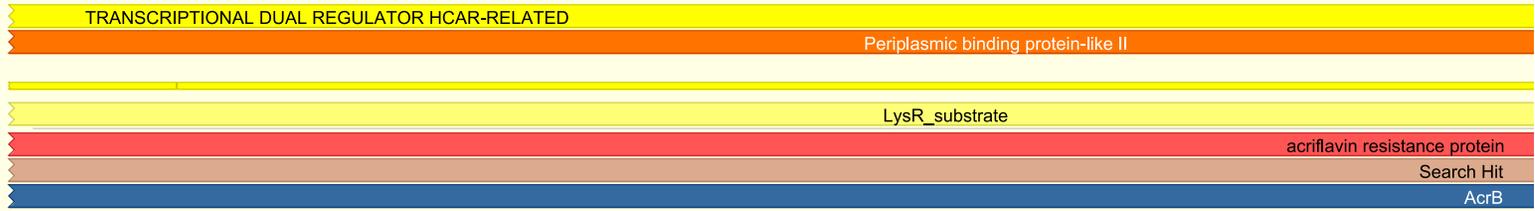
these are *not* mutations that arose during library propagation within SIGEX experiments, as they are present on different restriction fragments as shown by their relative end-points. Another instance, found on contig 243, demonstrates that the upstream termini of LA15 and LB20 both corresponded to the same nucleotide (Figure 4.7B); however, at the GFP-expressing termini, although each clone maps to the same gene, they terminate at different restriction sites within it. This is remarkable given the high level of nucleotide identity between the clones, and suggests that SIGEX was able to retrieve the same gene – from the same organism, or at least closely related organisms – more than once from the same metagenome, but on a different restriction fragment in the original library. Contig 243 also represents a completely novel gene arrangement; there are no long matches to any database entries, and it is only through *de novo* assembly and gene prediction that it is possible to assign functions to larger segments of this previously undescribed sequence.

As another metric for the validity of mapping the SIGEX clones, we looked for cases where both the forward and reverse Sanger reads mapped to the same contig but were not closed by end-sequencing, such as LB18 (Figure 4.7B). In this case, the Sanger reads aligned to a 2.4 kb span on contig 243, corresponding to the ~2.2 kb PCR amplicon from LB18 produced using the sequencing primers. This indicates that the clone was mapped to a genomic region of the appropriate length, even without knowledge of the middle sequence. We also asked whether any of the SIGEX clones may have been chimeric; *i.e.*, if the library creation process resulted in the incorporation of multiple metagenome fragments. We found that a

low proportion of clones (3 of the 21 in this study; LB20, LN1, and LK16) did contain chimeras; however, they were easily identified based on alignments with NGS-derived contigs where an abrupt change in % nucleotide identity, adjacent to a Sau3AI restriction site, is found (Figure 4.8).

Figure 4.8. Chimeric sequences found in a subset of SIGEX clones. A chimeric sequence was easily determined by the sudden divergence from its reference sequence (Contig 243) in the middle of a SIGEX clone (LB20). All examples of chimeric sequences in the SIGEX clones are preceded by a nearby GATC sequence, highlighting the possibility that the sequences derive from two independent restriction fragments.

32,892 32,902 32,912 32,922 32,932 32,942 32,952 32,962 32,972



1. Scaffold 243

- 2. LA15
- 3. LB20pADLeft
- 4. LB18GfpSeq
- 5. LB1
- 6. LB2

CCATTTGGCGGTAAAGTCCGACAAGTTCAGATTGATCTTGATCCTGCCGCTTTGCAGGCGCGCGGCGTATCCGGTCAGGATGTCGCTGCC
 CCATTTGGCGGTAAAGTCCGACAAGTTCAGATTGATCTTGATC
 CCATTTGGCGGTAAAGTCCGACAAGTTCAGATTGATCTTGATCCTAAGTAACTAACTAGATTAAAGAGGAGATAACATATGAGTAA
 CCATTTGGCGGTAAAGTCCGACAAGTTCAGATTGATCTTGATCCTGCCGCTTTGCAGGCGCGCGGCGTATCCGGTCAGGATGTCGCTGCC
 CCATTTGGCGGTAAAGTCCGACAAGTTCAGATTGATCTTGATCCTGCCGCTTTGCAGGCGCGCGGCGTATCCGGTCAGGATGTCGCTGCC
 CCATTTGGCGGTAAAGTCCGACAAGTTCAGATTGATCTTGATCCTGCCGCTTTGCAGGCGCGCGGCGTATCCGGTCAGGATGTCGCTGCC

4.4 Discussion

This chapter used a metagenomics approach combining the *de novo* assembly of NGS reads with sequence analysis of clones recovered in Chapter 3 using a phenotypic assay (SIGEX). We were able to map aromatic-inducible genes, recovered from an AH contaminated soil using SIGEX, to sequences assembled and annotated *de novo* from Illumina sequence data (Figure 4.4). Each of these procedures, taken individually, would lack important biological information. In the case of SIGEX, the sequence of nearby genes is absent, making it difficult to pinpoint with much certainty the genetic structure of the operon from which the clone originated. Conversely, NGS assembly on its own lacks a) the genetic or biochemical evidence to support *in silico* predictions, and b) a method to narrow down the immense number of predicted genes to those of interest for the study at hand (Suenaga, 2012). Combining these two approaches facilitated the identification of biologically relevant genes and operons that were present downstream or upstream from known inducible promoters (Figure 4.6).

We used two alternative assembly methods in this study: *de novo* assembly using IDBA-UD, and targeted assembly using PRICE. Complete *de novo* assembly gave us access to the overall best sequence data (Table 4.3). However, using directed assembly through PRICE is faster compared to most *de novo* assembly programs, and can be carried out on consumer-grade computers since much less RAM is required relative to *de novo* assembly. In a few cases, some of the contigs from the PRICE assembly were better matched to SIGEX clones than were their counterparts from the *de novo* assembly data, indicating that, if the resources are

available, it is beneficial to run both types of analysis to establish the best overall match for each clone.

We found that *de novo* assembly of our dataset provided a sufficiently large N50 for the purposes of mapping several complete operons. The quality of a metagenome assembly, as measured by the N50, depends mainly on the species diversity present in the sample. Compared to other deep sequencing studies on soils (Delmont et al., 2012a), we obtained relatively large contigs (with a scaffolded N50 of 9.1 kb, and ~2000 sequences >26.8 kb). This is attributable to the fact that we may have reduced the species richness of the sample (Suenaga, 2012) by aerating and agitating the contaminated soil in a bioslurry, thereby enriching for species capable of degrading the complex mixture of xenobiotics found in the Rock Bay soil.

A very high level of identity was observed between the SIGEX clones and NGS-derived contigs, with an average of >95% identity at the nucleotide level observed between IDBA-UD *de novo* assembled contigs and the sequences of SIGEX-clones obtained from Sanger sequencing (Table 4.3). Some of these differences can be attributed to population variation (Figure 4.7A), since identical SNPs were discovered in several SIGEX clones (each of which originate from a unique fragment of metagenomic DNA). Our confidence in mapping fragments to contigs was bolstered by targeted assembly of NGS reads using PRICE: this procedure limited contig extension to regions hypothesized to be relevant for aromatic metabolism.

We show that contigs of similar size and composition map to the clones, regardless of the assembly process (Table 4.3).

Many contigs pointed us directly to a database entry containing matches for each putative gene identified in a predicted operon (e.g., contigs 3075, 6160, and 23284). In these cases, interpretation of the biological significance of the clones mapping to these contigs was straight forward. Furthermore, these previously described gene clusters are known to be inducible by some of the compounds we tested using SIGEX. For example, contig 3075 contains sequences highly similar to the *Pseudomonas* plasmid pAK5, which has been characterized by Izmalkova et al. (2013). In *P. putida* harbouring pAK5, gentisate 1,2-dioxygenase activity was induced by naphthalene or salicylate (Izmalkova et al., 2013). This corroborates our finding that clone SE2 contains a salicylate inducible promoter driving expression of the *sgp* operon; we also note that it was inducible by salicylate, and was also upregulated approximately 1.5 to 2.5 fold by pyrene, benzoate, catechol, phenylacetic acid and phenol. Similarly, an operon containing the *nah* genes was found on contig 6160, which exhibits high similarity to pDTG1; this operon is regulated by *nahR* and is induced by salicylate (Dennis & Zylstra, 2004; Park et al., 2002; Schell & Poser, 1989). Clone SA1 maps to contig 6160 and was weakly inducible by phenylacetic acid, phenol, and naphthalene. Several salicylate inducible clones map to contig 23284, which in turn is highly similar to an operon found in *P. stutzeri* CCUG 29243 (a.k.a. AN10). The *nah* gene cluster in AN10 is

reconstituted on contig 23284 and represents another example of a known salicylate inducible operon recovered with SIGEX (Bosch et al., 1999).

In several cases, no database entries existed that shared a similar genetic structure to that predicted for the contig. For example, contig 3721 shares some similarity to a region of *P. putida* plasmid NAH7 and a separate region to the *Azotobacter vinelandii* DJ chromosome (as well as several others), but we were unable to identify any existing characterized sequences with the exact gene organization found on contig 3721. Taken together, the predicted genes compose an operon putatively responsible for degrading phenol (based on the metabolic genes designated RockBay_127618 – 127625) as well as the uptake of aromatic compounds (RockBay_127613 – 127615, which are similar to *benK* and *benF*). The genes in this operon each align to sequences from a different species, making taxonomic classification difficult. However, we note a similarity in the operon structure of the metabolic genes to that described by Suenaga et al. (2009) for phenol catabolizing genes. This is surprising since the clone LD23 (aligning to contig 3721) was most inducible by salicylate, not phenol. Thus, contig 3721 could represent a novel operon that has yet to be isolated from any cultivated species.

Aromatic degrading plasmids (or other mobile elements), especially those classified as the NAH-type of naphthalene degrading elements, are widespread in environments contaminated with coal tar and AHs (Brunet-Galmés et al., 2012; Fernández et al., 2012; Izmalkova et al., 2013; Li et al., 2004; Park et al., 2002; Takizawa et al., 1994). The examples shown in this study demonstrate that several different plasmids carrying genes for aromatic degradation were present in the

community we characterized, most of which have some unique aspects of gene organization and SNPs, as has been previously found (Suenaga et al., 2009). However, it remains unclear what role (if any) these mobile elements play in HMW-aromatic degradation. Although it is beyond the scope of this study, time course analysis of the bioslurry indicates that HMW PAHs were biodegraded significantly over the course of the experiment (Whynot, 2009), and determining what genes are involved in this process will be the subject of future work on this site. The use of novel tools such as the biodegradation database, used in Chapter 5, will assist in distinguishing genes that may be involved in biodegradation (Fang et al., 2013).

This study used NGS to complement existing metagenomic data in a combined approach linking phenotypically characterized metagenomic clones to larger contigs assembled *in silico*. It was shown that the genomic sequence surrounding the aromatic-inducible clones could be predicted, as demonstrated by the facts that a) each clone mapped to a NGS contig with high confidence (often >95% nucleotide identity), b) end-sequenced clones were mapped to the same NGS contig with a span corresponding to the size of the clone, and c) some of the contigs themselves contain predicted genes that correspond with >99% nucleotide identity to operons from previously sequenced organisms. Furthermore, the genes identified upstream and downstream from the inducible clones were shown to have biological significance to the metabolism of AHs that were used as inducers. This indicates that, for a community sequenced with sufficient depth, *in silico* assembly and prediction tools can be used to target the computational analysis of biologically relevant subsets of the metagenome.

Chapter 5.

Characterization of Microbial Communities in a Contaminated Site using Next-Generation Sequencing

5.1 Introduction

In a complex environment such as soil, microbial communities are composed of many organisms that interact with one another in numerous ways that are not yet understood. The interplay between factors such as horizontal gene transfer (HGT), cross feeding, and genetic variation within clonal populations are just a few examples of what makes those interactions difficult to establish. Contaminated sites comprise environments that are of special interest to microbiologists, since they may contain communities of microbes enriched with physiological characteristics of biotechnological interest; however, much of their genetic diversity remains untapped. Until recently, it was impossible to obtain the data necessary to evaluate the entire metagenomic community composition and structure of such an environment (Delmont et al., 2012b). However, with the advent of NGS, whole-metagenome sequencing has recently emerged as a viable method for determining taxonomic and functional relationships in soil metagenomes. Although some information obtained through shotgun metagenomics has been ground-breaking (Tyson et al., 2004; Venter et al., 2004), the results from sequence assembly and annotation performed entirely *in silico* must be interpreted with caution, due to

several shortcomings of current sequencing technologies and the computational processes used for analysis. These limitations include the inadequate depth of coverage (as well as read length) obtained during sequencing and the incomplete availability of reference sequences for the annotation of taxa without cultured representatives (Henry et al., 2011; Vogel et al., 2009; Wommack et al., 2008).

In this chapter, we aim to identify genetic elements that may be involved in xenobiotic transformation by using shotgun metagenome sequencing of a PAH-contaminated site metagenome, along with a variety of computational tools for the identification of gene function and taxonomic classification. The PAH-degrading elements described in Chapter 2 will partially guide the functional analysis. The data presented in this chapter are aimed to augment and inform the findings from Chapters 3 and 4, in which the same metagenome was screened for aromatic-inducible genes using SIGEX.

5.2 Methods

5.2.1 Metagenomic DNA Isolation and Sequencing

The Rock Bay soil was heavily contaminated with a variety of toxic aromatic compounds (Chapter 3; Whynot, 2009). Metagenomic DNA was isolated from the Rock Bay soil bioslurry samples as described in Chapter 3. Isolated DNA was sequenced with the Illumina HiSeq 2000 platform (Genome Québec, Montréal, Québec) for a total of three lanes of sequence data, collected in 100 bp paired-end (PE) reads (see Chapter 4, Methods, for details). One library, used for two lanes of sequencing, contained 255 bp inserts (designated RB-255); the other, used for one lane of sequencing, contained 447 bp inserts (designated RB-447). Raw data was filtered (Genome Québec, Montréal, Québec) prior to use in downstream bioinformatics applications; the analysis statistics of these sequence reads are shown in Table 5.1.

Table 5.1. Metagenome sequences uploaded to MG-RAST for analysis. Some statistical measures of the sequence data are shown.

MG-RAST ID (Name)	Total starting base pairs	Base pairs, post MG-RAST QC	Post QC mean GC content	MG-RAST Identified Protein Features	MG-RAST Identified rRNA Features	α -Diversity from MG-RAST Annotations
4494328.3 (RB-255-3-R1)	20,339,741,000	13,445,814,355	59 \pm 8 %	75,081,064	930,100	276.632
4494329.3 (RB-255-3-R2)	20,339,741,000	13,052,740,865	59 \pm 8	70,976,492	899,494	279.302
4494326.3 (RB-255-4-R1)	20,424,765,600	13,456,923,931	59 \pm 8	72,243,106	933,669	279.328
4494327.3 (RB-255-4-R2)	20,424,765,600	13,134,969,802	59 \pm 8	71,630,228	901,115	278.495
4494330.3 (RB-447-3-R1)	21,786,407,600	11,717,480,627	58 \pm 9	48,776,035	1,014,919	278.263
4494331.3 (RB-447-3-R2)	21,786,407,600	11,004,495,154	58 \pm 9	57,219,998	916,427	279.830
Total (raw data)	125,101,828,400	75,812,424,734	NA	NA	NA	NA
4514941.3 (Contigs - IDBA-UD assembly)	N/A	407,472,907	58 \pm 8	227,123	294	148.132

5.2.2 Assembly of Metagenomic Sequences and MG-RAST Analysis

The *de novo* assembly of raw reads into contigs is described in detail in Chapter 4. Briefly, IDBA-UD (Peng et al., 2012) was used to obtain assembled sequences using kmer sizes of 20 to 100 and a step size of 1, using all 3 paired end lanes of Illumina sequence as input. Assembled contigs, as well as raw sequence data, were uploaded to MG-RAST for annotation; this data is available under the accession numbers listed in Table 5.1. MG-RAST was used to determine taxonomic and functional annotations by looking for similarities in the sequences within the M5NR database. The M5NR protein database used by MG-RAST is comprised of non-redundant protein and rRNA sequences originating from the following databases: GenBank, IMG, KEGG, PATRIC, RefSeq, SEED, SwissProt, TrEMBL, and eggNOG (functional and organism classifications); COG, KO, NOG, and Subsystems (functional hierarchy annotations); Greengenes, SILVA LSU, RDP, and SILVA SSU (rRNA). The functional and taxonomic annotations from raw reads (using the RB-255-3R1 dataset, since it had the best QC score; MG-RAST ID 4494328.3) were compared to those from the assembled contigs, to establish if any difference exists between the datasets. KEGG pathway analysis was done using KeggMapper with an E-value cut-off of $1e-5$ and 60% identity, with a minimum alignment length of 15 bp or 15 a.a.

5.2.3 Identification of Biodegradation Genes

The biodegradation gene database (BDG) described by Fang et al. (2013) was used to probe for genes that might have biodegradation activity that were not detected using SIGEX (Chapter 3). The BDG contains a list of approximately 50,000

non-redundant protein sequences, and their NCBI accessions, of a variety of manually curated genes known to be important for biodegradation. The BDG is comprised of the following gene families: *alkB*, *benA*, *bph*, *bphA1*, *bphA2*, *carA*, *dbfA1*, *dxnA*, *dxnA-dbfA1*, *glx*, *mmoX*, *npah*, *p450*, *ppah*, and *ppo*. Note that many of these were described in Chapter 2 in the context of PAH biodegradation, but comparatively few of these gene families were identified using SIGEX in Chapter 3.

To determine if entries in the BDG database had matches within the Rock Bay metagenome, a local BLASTp search was performed using an E-value cut-off of 1e-5. Each of the ~477,000 proteins predicted by MetaGeneMark from the IDBA-UD *de novo* assembly (Chapter 4) was queried against the BDG database. Hits were filtered at a cut-off of 90% pairwise protein identity, in accordance with previous work using such database annotation methods (Fang et al., 2013; Kristiansson et al., 2011; Shi et al., 2013). Gene classes were identified using the HMMER profiles (available at <http://fungene.cme.msu.edu/>; Fish et al., 2013) for each gene in the BDG database: first, the HMMER profiles were combined into an HMM database (using *hmmbuild* and *hmmcompress* with default parameters). Next, filtered BDG database BLAST hits were used as queries in an *hmmsearch* search of the new BDG HMM database (using default parameters). The filtered hits from the BDG database were then analyzed for their taxonomic relationships using MEGAN (Metagenome Annotator) version 5.0.83 beta, with a minimum support of 1, bit score of 40, maximum expected value of 0.01, minimum complexity of 0.44, and a top percent of 100 (Huson et al., 2007).

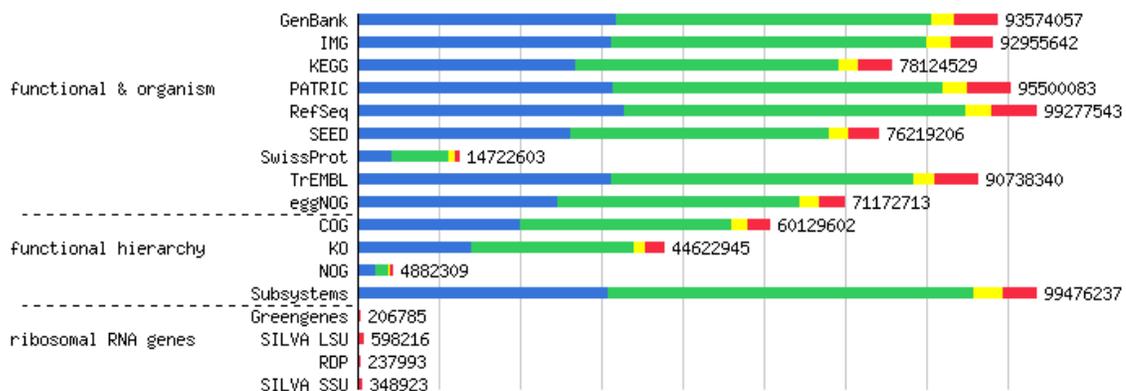
5.3 Results

5.3.1 Overview of Annotations

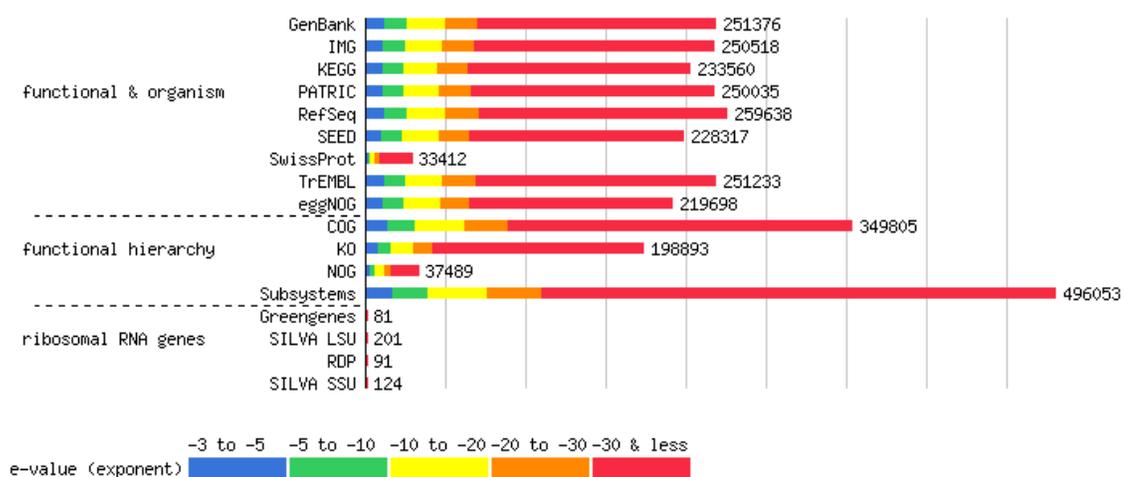
MG-RAST was used to assess either raw Illumina data (annotations for Illumina reads presented throughout this chapter are exclusive to RB-255-3R1, MG-RAST ID 4494328.3 unless otherwise stated) or IDBA-UD assembled contigs for the presence of protein and rRNA features. Of the 136,625,753 raw Illumina reads, 92.8% of those contained predicted ORFs; of the 117,116 annotated contigs, 100% contained predicted ORFs. In the ORFs found on raw Illumina reads, 59.2% contained identified protein features; of those, 92.7% were classified into functional categories. From ORFs found on assembled contigs, 75.6% contained protein features, and 84% of those were assigned a functional category. The hit distributions within each database for each metagenomic dataset are shown in Figure 5.1. For rRNA prediction, the raw Illumina reads gave rise to 930,100 identified rRNA features, while the assembled contigs gave rise to 294 identified rRNA features.

Figure 5.1. Source hits distribution of MG-RAST annotated features for (A) Illumina reads (RB-255-3-R1), MG-RAST ID 4494328.3 and (B) IDBA-UD assembled contigs, MG-RAST ID 4514941.3. Length of bars corresponds to the relative number of hits in each database; colors represent different E-values. Although the proportion of high-confidence hits is increased in the assembled sequence dataset, the number of overall hits is significantly higher for the Illumina reads.

A. MG-RAST source hits distribution for Illumina reads



B. MG-RAST source hits distribution for IDBA-UD assembled contigs



5.3.2 Taxonomic Analysis

Metagenomic samples were analyzed for their taxonomic classifications at the levels of domain, phylum, class, order, family, and genus using the M5NR database. The Illumina reads were classified as 97.4% belonging to the domain bacteria, while the assembled contigs were classified as 98.6% bacteria. At the phylum level, Illumina reads were annotated as 87.8% Proteobacteria, 3.2% Actinobacteria, 2.8% Firmicutes, and 1.0% Bacteroidetes; assembled contigs were found to contain 93.1% Proteobacteria, 1.2% Actinobacteria, 1.9% Firmicutes, and 0.7% Bacteroidetes. Other phyla were present in smaller magnitudes for both samples, the relative proportions of which are shown in Figures 5.2 and 5.3. The most frequently annotated genus in both the Illumina reads (29.7%) and in assembled contigs (32.5%) was *Pseudomonas*. Many different genera were annotated (9364 for Illumina reads, and 1327 for assembled contigs), but most were in significantly smaller relative proportions, as shown in Figures 5.4 and 5.5. For comparison, annotations of rRNA within the raw Illumina reads and assembled contig datasets are shown in (Figure 5.6).

Figure 5.2. Phylum-level classification of Illumina sequence reads as determined using the M5NR database in MG-RAST (MG-RAST ID 4494328.3). Labels are shown only for those phyla present in sufficient numbers to be visible on the graph.

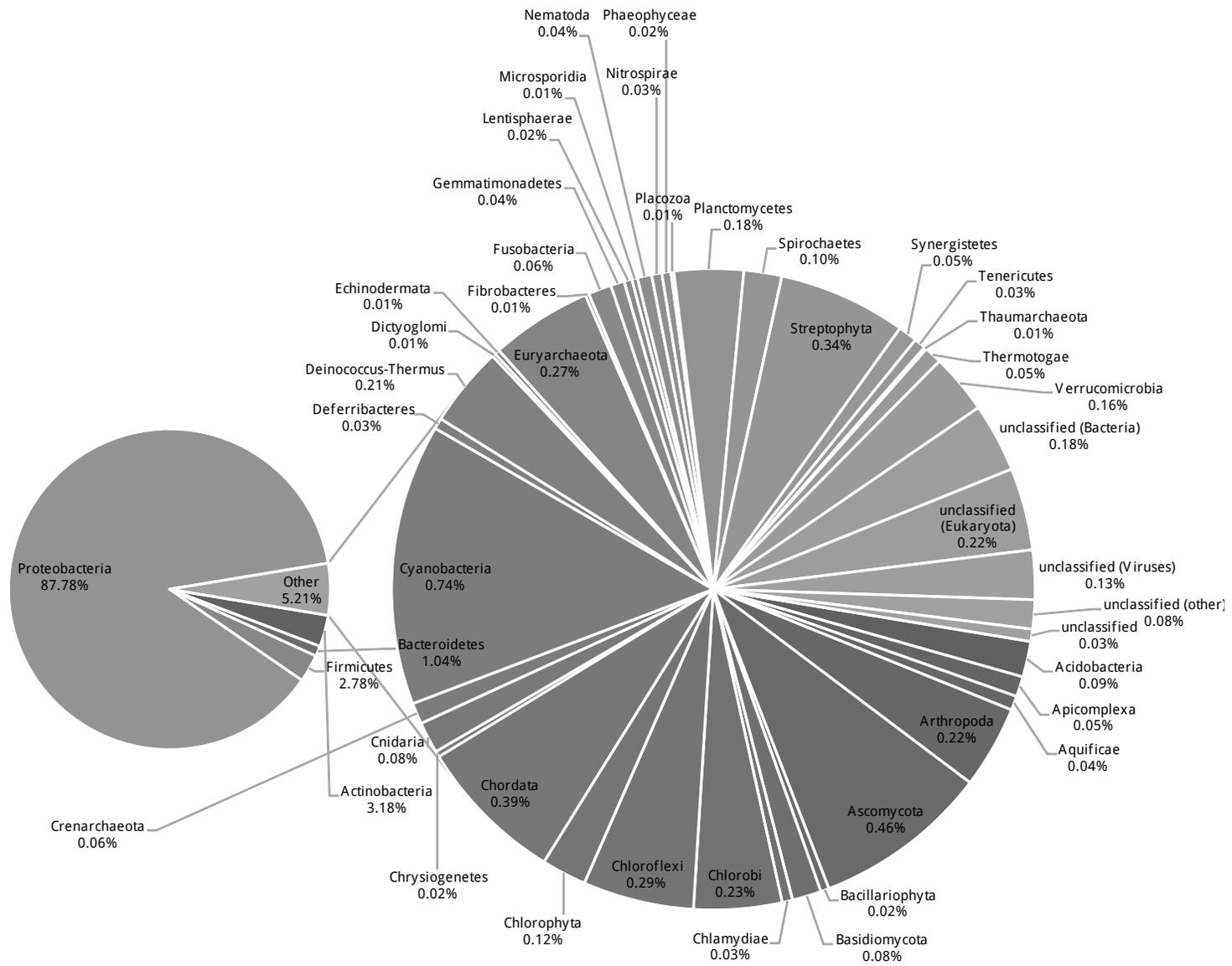


Figure 5.3. Phylum-level classification of IDBA-UD assembled contigs as determined using the M5NR database in MG-RAST (MG-RAST ID 4514941.3). Labels are shown only for those phyla present in sufficient numbers to be visible on the graph.

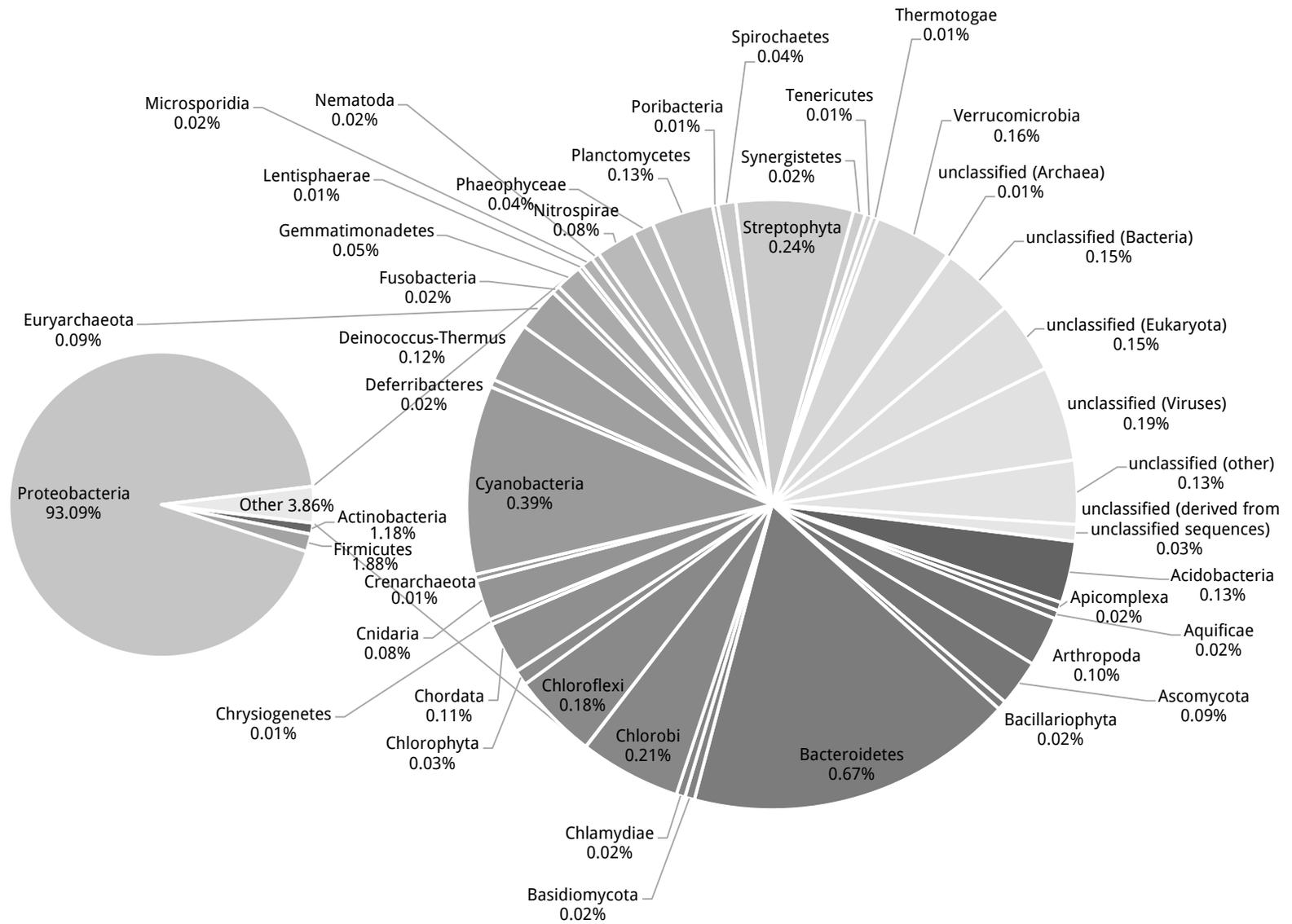
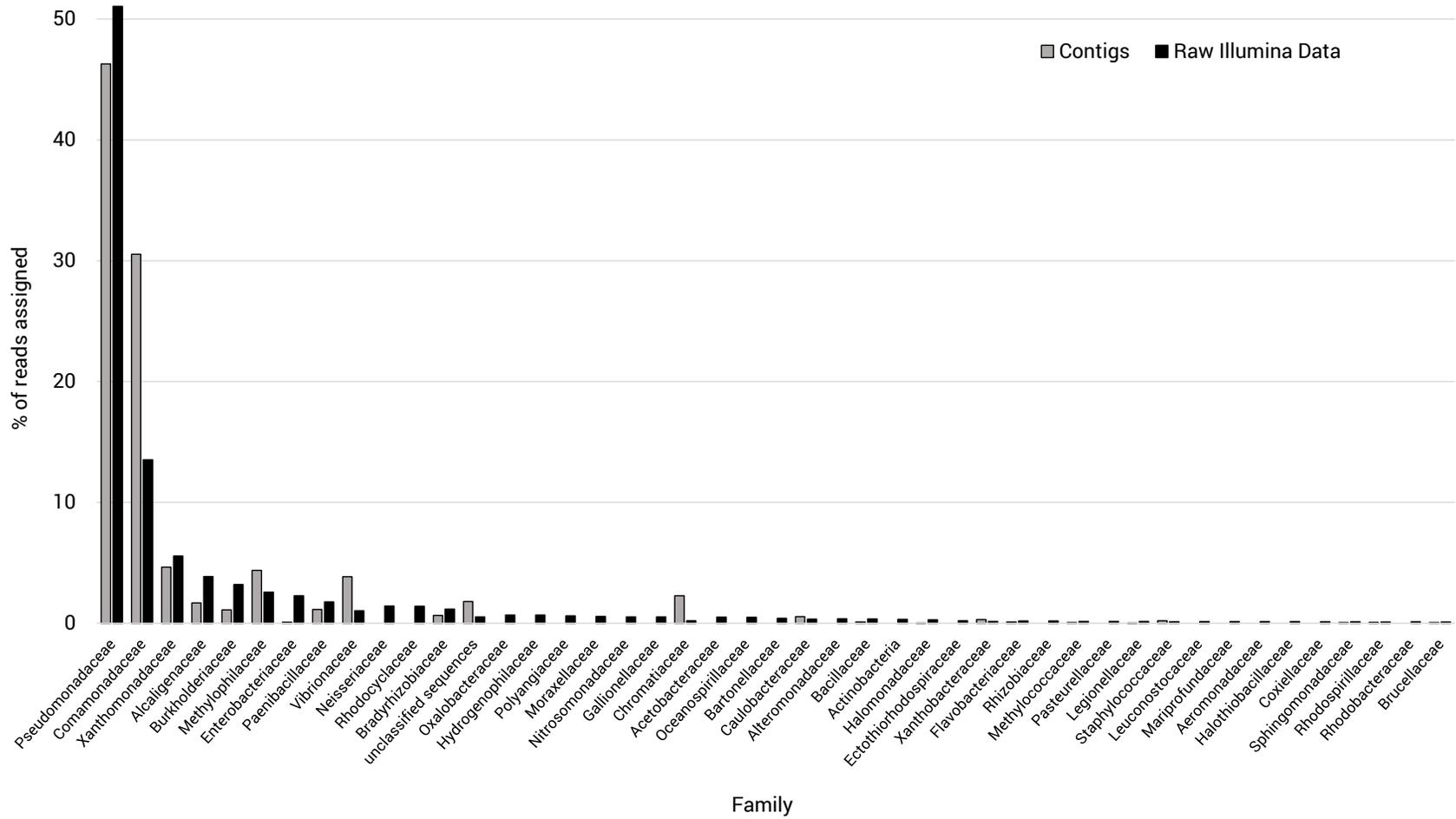


Figure 5.4. Genus-level classification of Illumina sequence reads as determined using the M5NR database in MG-RAST (MG-RAST ID 4494328.3). Only the 100 most common genera are shown; genera that are not present in sufficient numbers to be visible are not labeled.

Figure 5.5. Genus-level classification of IDBA-UD assembled contigs determined using the M5NR database in MG-RAST (MG-RAST ID 4514941.3). Only the 100 most common genera are shown; genera that are not present in sufficient numbers to be visible are not labeled.

Figure 5.6. Comparison of bacterial rRNA annotations between raw Illumina reads and assembled contigs (black and grey, respectively). The height of each bar represents the number of reads assigned to that family. Annotations are from the LSU, SSU, M5RNA, RDP, Greengenes, and ITS databases, and were filtered using the MG-RAST workbench using a cutoff of 97 % identity and an E-value of $1e-5$ with an alignment length cutoff of 15. Visualization was restricted to bacterial sequences, and those with a relative abundance greater than 0.1% in the Illumina raw data.



5.3.3 MG-RAST Analysis of Aromatic Metabolism

Based on the functional annotations made by MG-RAST, we examined the datasets containing either Illumina reads or assembled contigs for genes that may be involved in the degradation of aromatic compounds. The SEED “Subsystems” annotation database contains a category called “Metabolism of aromatic compounds”, which encompasses a wide variety of metabolic enzymes and subclasses of enzymes (*e.g.*, “Metabolism of central aromatic intermediates”, which itself contains subclasses). Similar, though slightly varied, proportions of these enzymes were annotated in the Illumina (Figure 5.7) and assembled contigs (Figure 5.8) datasets. A larger proportion of the total dataset for the assembled contigs, compared to raw Illumina reads, contained sequences annotated as “Metabolism of aromatics” (4.1% compared to 2.5%). These annotations indicate the widespread presence of genes involved aromatic metabolism within the microbial communities found in the Rock Bay soil bioslurry. This was demonstrated further by KEGG pathway analysis: complete (or nearly complete) pathways for several known aromatic xenobiotics were present for benzo[a]pyrene (Figure 5.9), benzoate/catechol (Figure 5.10), and toluene/xylene (Figure 5.11). Not all pathways were complete, however, as exemplified by the naphthalene/anthracene degradation pathway (Figure 5.12).

Figure 5.7. Functional category breakdown of reads classified as “Metabolism of aromatics” in the SEED Subsystems database using MG-RAST annotations of Illumina reads (MG-RAST ID 4494328.3). Features within the subsystem “Metabolism of aromatics” are found on 2.5% of all the annotated contigs.



Figure 5.8. Functional category breakdown of reads classified as “Metabolism of aromatics” in the SEED Subsystems database using MG-RAST annotations of assembled contigs (MG-RAST ID 4514941.3). Features within the subsystem “Metabolism of aromatics” are found on 4.1% of all the annotated contigs.

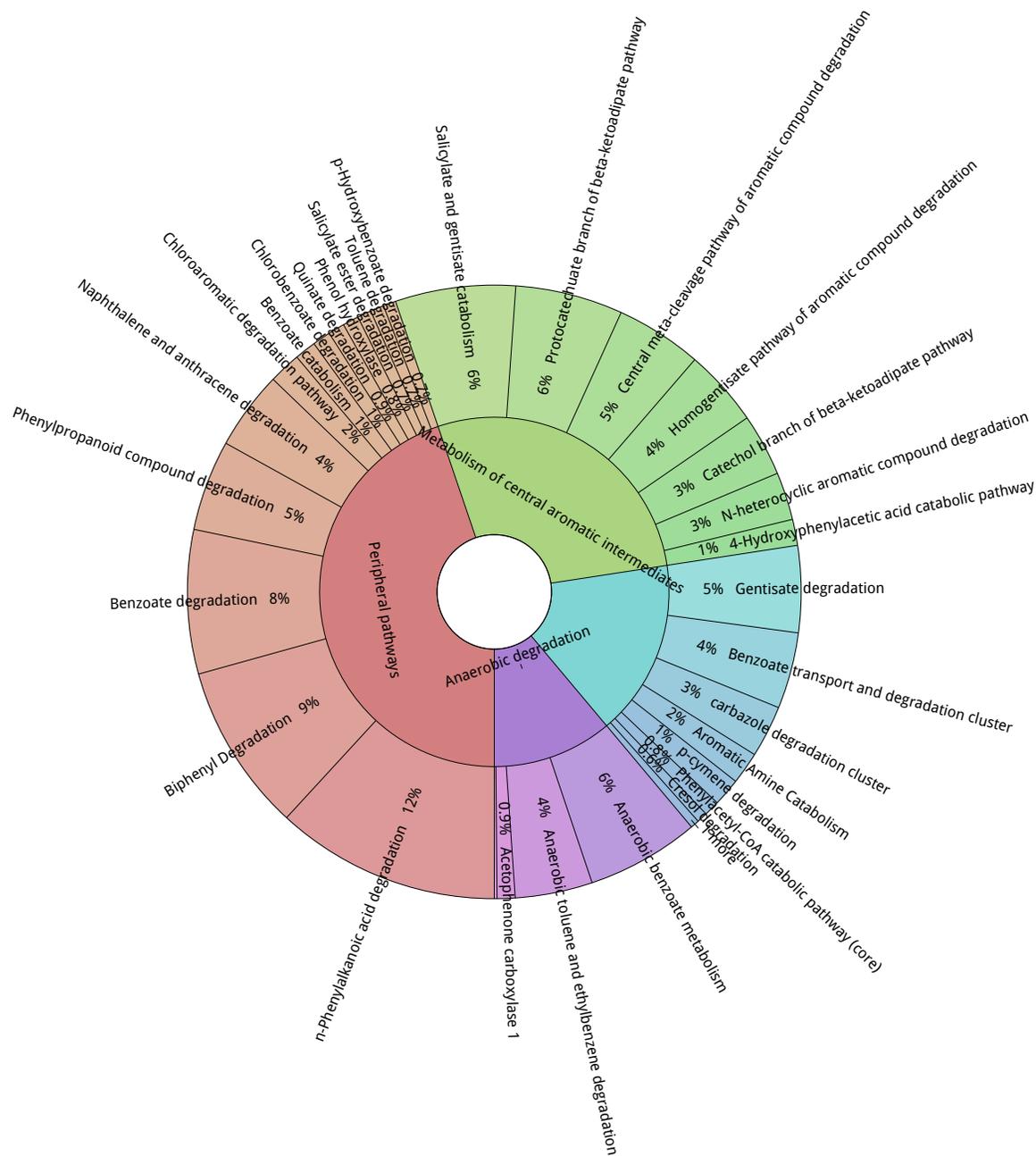


Figure 5.9. KEGG pathway analysis showing the enzymes required for metabolism of benzo[a]pyrene. Shaded boxes represent enzymes for which genes were found in assembled contigs (MG-RAST ID 4514941.3) – in this example the complete pathway was present.

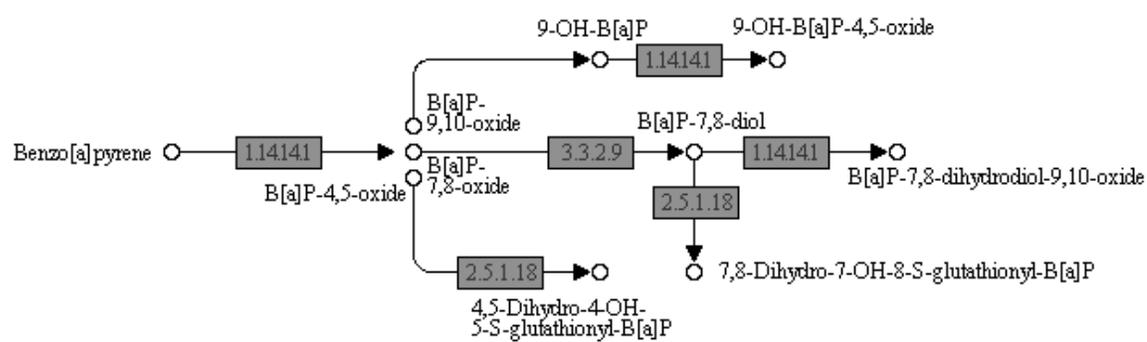


Figure 5.10. KEGG pathway analysis showing the enzymes required for the metabolism of benzoate and catechol through hydroxylation. Dark shaded boxes indicate enzymes for which genes were found in assembled contigs (MG-RAST ID 4514941.3) and light shaded boxes indicate those found in Illumina reads (MG-RAST ID 4494328.3). All enzymes required for the conversion of catechol to pyruvate or acetaldehyde and acetyl-CoA are present. No shading indicates that no match was present.

Figure 5.11. KEGG pathway analysis showing the enzymes required for the metabolism of toluene and xylene. Dark shaded boxes indicate enzymes for which genes were found in assembled contigs (MG-RAST ID 4514941.3) and light shaded boxes indicate those found in Illumina reads (MG-RAST ID 4494328.3). No shading indicates that no match was present.

TOLUENE AND XYLENE DEGRADATION

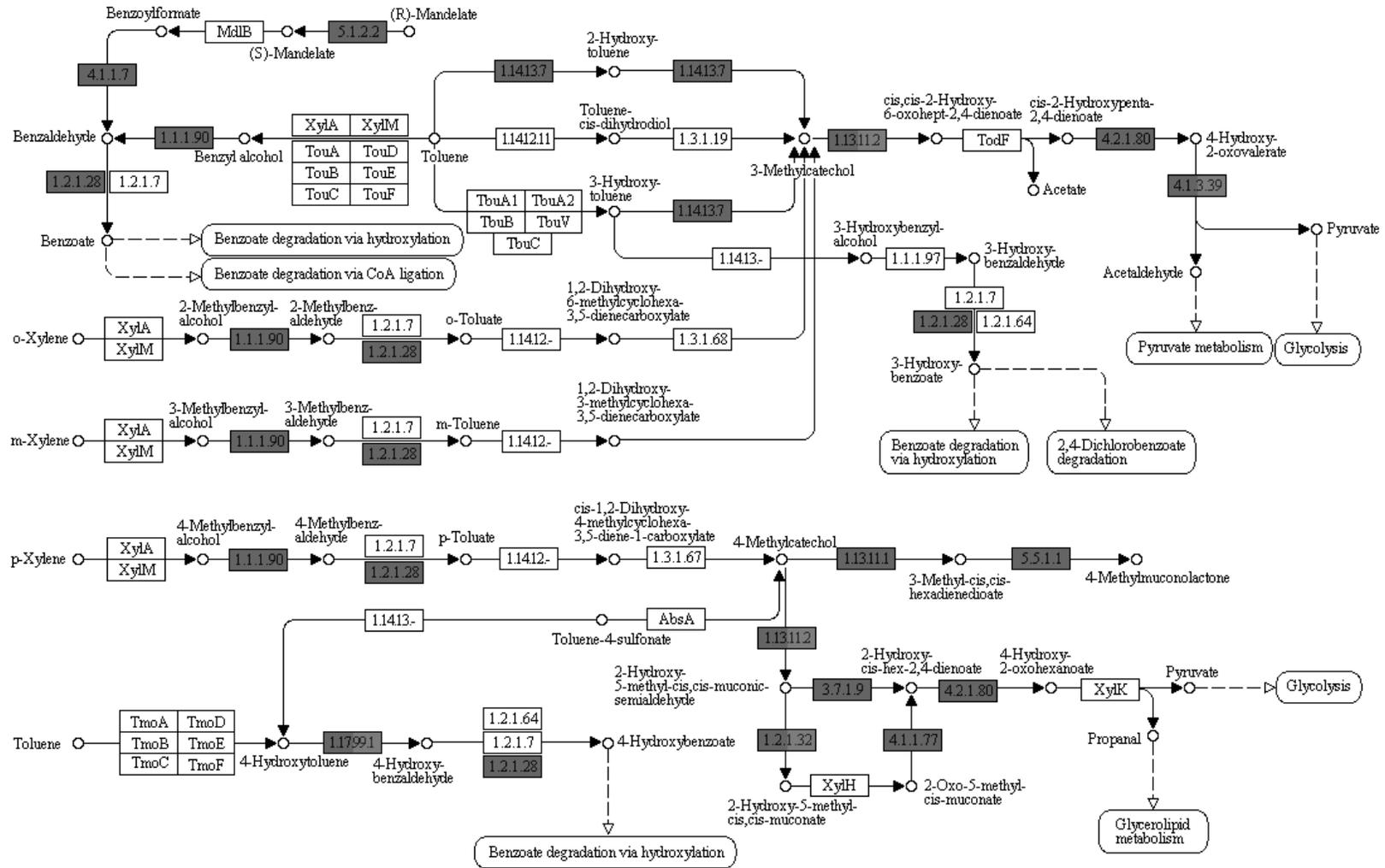
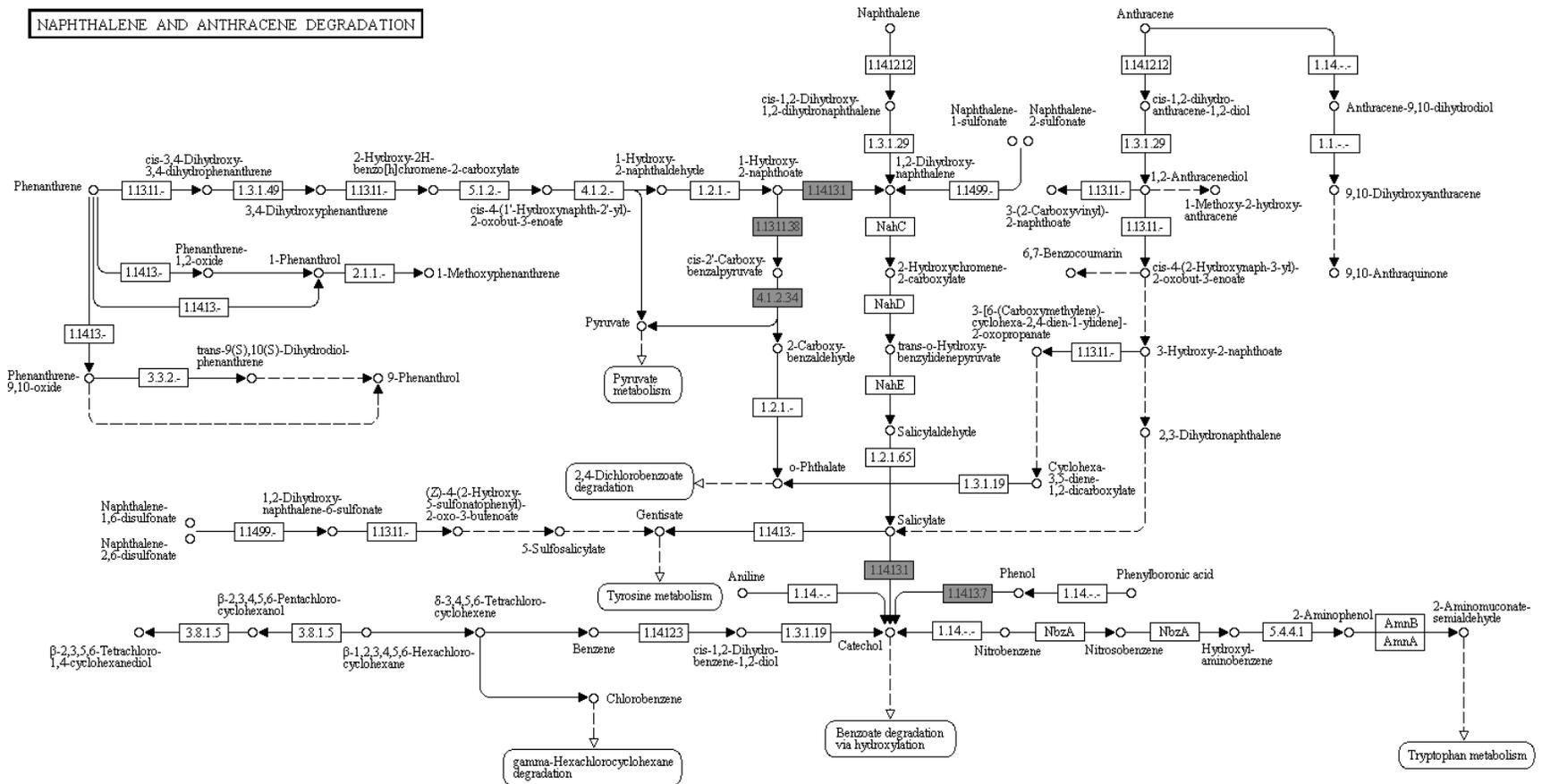


Figure 5.12. KEGG pathway analysis showing the enzymes required for the metabolism of naphthalene and anthracene. Shaded boxes represent enzymes for which genes were found in assembled contigs (MG-RAST ID 4514941.3). No shading indicates that no match was present.

NAPHTHALENE AND ANTHRACENE DEGRADATION



00626 11/16/09
 (c) Kanehisa Laboratories

5.3.4 Biodegradation Gene Identification

The BLASTp query of 477,784 MetaGeneMark-identified proteins (derived from *de novo* assembled NGS scaffolds of the Rock Bay metagenome) vs. the 49,090 BDG database proteins gave rise to 25,451 hits (around 5.3% of the Rock Bay proteins); of those, 1607 corresponded to a unique (non-redundant) entry in the BDG database. Following filtering at 90% amino acid identity, 417 hits remained, of which 246 constituted unique entries in the BDG database. The taxonomic relationships of these final hits and their relative proportions (calculated without the removal of multiple hits to the same sequence) is shown in Figure 5.13. All of these sequences align to matches within 64 different species, most of which are Proteobacteria (of those, 21 fall under the genus *Pseudomonas*). A more detailed breakdown of the gene classes found within each species, as well as the overall relative proportions of each gene class, is shown in Figure 5.14. A quantitative species-centric view of the prevalence for each gene is shown in Figure 5.15.

Figure 5.13. A phylogram showing the taxonomy of biodegradation genes found in the Rock Bay PAH-contaminated site metagenome. The size of each green circle corresponds to the abundance of sequences assigned to that taxonomic group (assigned sequences are also enumerated beside taxa names). Proteins were identified using the biodegradation gene database (Fang et al., 2013) as a BLAST database for MetaGeneMark annotated protein queries from assembled contigs. Taxonomic relationships were determined using the NCBI taxonomy, and visualized through MEGAN 5.

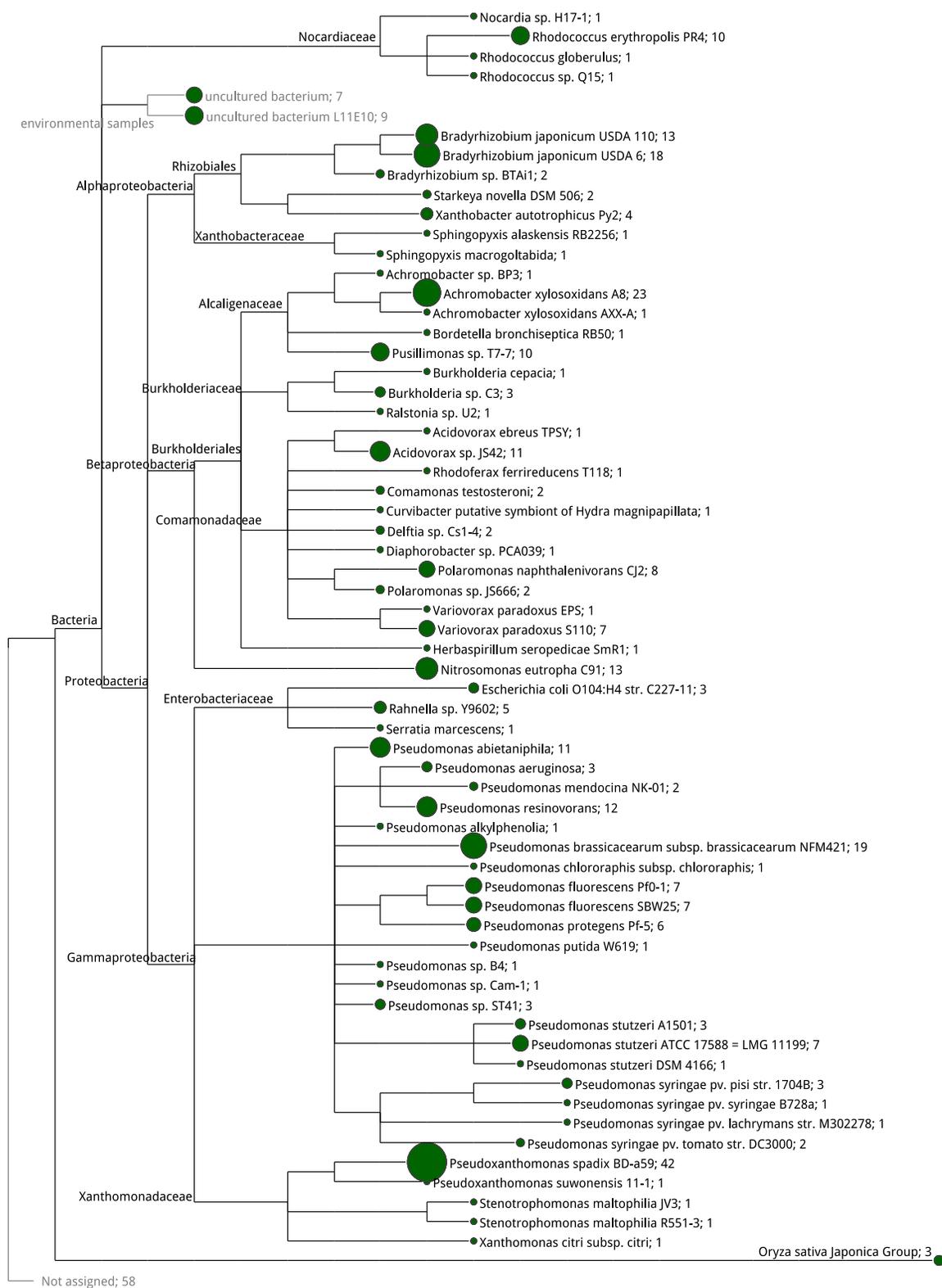


Figure 5.14. Relative proportions of biodegradation genes in each species annotated within NGS contigs from the Rock Bay PAH-contaminated site metagenome. The height of each bar shows the prevalence of each gene class identified from the BDG database within each species, with each gene class represented by a unique color. The inset graph shows the prevalence of each gene class independently of the species from which it originates.

Figure 5.15. The species distribution (left) within each gene class in the BDG database. Each brick (colored by species) corresponds to four sequences that were annotated within each species (with white boxes corresponding to between 1 and 3 sequences), and are distributed above the gene class into which they are categorized.

5.4 Discussion

MG-RAST was used to determine taxonomic and functional relationships for sequences in datasets derived from Illumina sequencing and assembled contigs (raw Illumina reads assembled using IDBA-UD). The sequences that were assembled into contigs prior to annotation generally gave rise to higher confidence hits than the raw Illumina reads, although there were fewer sequences in total (Figure 5.1). This is explained by the fact that assembly gives rise to sequences longer than the 100 bp that comprise raw reads; consequently, there is a higher chance of obtaining a longer (and therefore more statistically significant) match to a database entry (e.g., of the Subsystems hits, those with an E-value of less than $1e^{-30}$ comprised 74.6% of the assembled reads, but only 4.9% of raw Illumina reads).

The taxonomic groups identified using MG-RAST were generally consistent between Illumina reads and assembled contigs. By far, the most prevalent group found in the shotgun sequence data is the phylum Proteobacteria, annotated on approximately 90% of the sequences from raw reads as well as assembled contigs (Figures 5.2 and 5.3). The genus *Pseudomonas*, represented at about 30% in both datasets, was most common (Figures 5.4 and 5.5); this is consistent with the observation that Pseudomonads are common in aerobic environments rich in organic material. Furthermore, Pseudomonads are often implicated in metabolically diverse processes, especially xenobiotic metabolism (Labana et al., 2007;

Silby et al., 2011). *Pseudomonas* was also the most common genus within taxonomic assignments of SIGEX clones in Chapter 3.

While a sizeable percentage of the total sequences were annotated by MG-RAST as involved in aromatic metabolism (2.5% - 4.1%), there was a significant assortment of different enzyme classes present in the metagenome. The features annotated from the M5NR database show that the genetic requirements of nearly-complete pathways for a wide variety of aromatic compounds are present in the sequenced metagenomic DNA (Figures 5.9, 5.10, 5.11, and 5.12). Evidence presented in Chapters 3 and 4 showed that the assembled contigs contained several operons responsible for aromatic degradation (e.g., naphthalene, salicylate), and those same elements were upregulated by the compounds on which they act, as indicated by GFP upregulation in transcriptional fusions with those genes (e.g., contig 23284 and associated SIGEX clones); similar pathways were identified using KEGG mapper. The fact that more complete pathways were annotated in the assembled contigs dataset vs. raw Illumina reads is indicative of the statistical powers afforded from having longer matches to database entries.

It is also evident that several pathways are present in this metagenome that were not retrieved using SIGEX library screening, even though attempts were made. For instance, a pathway for benzo[a]pyrene metabolism exists (Figure 5.9), but no benzo[a]pyrene inducible clones were recovered in Chapter 3. However, this method is limited in its ability to predict certain functions: for example, the naphthalene / anthracene pathway is incomplete according to the KEGG annotations (Figure 5.12), but chemical evidence from soil bioslurry experiments

indicates that anthracene was degraded quickly by this microbial community (Whynot, 2009). Annotations derived from organisms that are distantly related to representatives in metagenomic sequences, or, annotations that are made using low similarities to reference sequences, may skew the results of pathway analysis. This suggests that more biochemical data is needed to assess the *bona fide* functions of uncultured genes.

The BDG database provided a method of annotating genes within the Rock Bay metagenome that constitute previously characterized elements known to be involved in biodegradation. *Pseudomonas* was found to be the most prevalent genus, in accordance with both MG-RAST annotations of all Illumina data as well as the annotations of SIGEX clones. There is a distinct absence of biodegradation genes derived from *Mycobacterium* and *Sphingomonas* species (each genus represents 0.5% of the raw Illumina reads; assembled contigs are comprised of 0.2% *Mycobacterium* and 0.3% *Sphingomonas*), an observation that might be interpreted as surprising given the wealth of literature that has portrayed *Mycobacterium* and *Sphingomonas* as common PAH-degrading members of soil communities (Chapter 2). Several known PAH-degrading genera were, however, observed at higher relative abundances ($\geq 1.0\%$), all of which constitute Gram-negative Proteobacteria. This includes *Achromobacter* (Vinas et al., 2005; 1.5% of Illumina reads; 3% of contigs), *Acidovorax* (Eriksson et al., 2003; Jones et al., 2011; Singleton et al., 2009; 3.1% of Illumina reads; 6.5% of contigs), *Bordetella* (Eriksson et al., 2003; 2.2% of Illumina reads; 2.7% of contigs), *Burkholderia* (Tittabutr et al., 2011; 5.4% of Illumina reads; 3.1% of contigs), *Stenotrophomonas*

(Boonchan et al., 1998; 1.0% of Illumina reads; 1.4% of contigs), *Variovorax* (Eriksson et al., 2003; 1.1% of Illumina reads; 2.4% of contigs), and *Xanthomonas* (Hamann et al., 1999; 2.6% of Illumina reads; 2.2% of contigs). No Gram-positive genera known to metabolize HMW PAHs were present at more than 0.5% (*Mycobacterium*).

Although the most frequently characterized HMW PAH-degrading taxa were not found in high numbers, several of the PAH-degrading genes described in Chapter 2 were identified within the Rock Bay metagenome using the BDG database. For instance, the *bph* and *bphA1* genes, encoding carbazole dioxygenases, were identified in metagenomic sequences; however, none of the SIGEX clones had similarity to those genes. In fact, when predicted genes from contigs containing mapped SIGEX sequences were used as queries, only Contig 3075 (containing the *Pseudomonas putida* plasmid pAK5 salicylate-gentisate pathway) had genes that were found in the BDG database at a cutoff of $\geq 90\%$ a.a. identity; the genes on this contig were assigned to the gene classes *bphA2*, *carA*, and *dxnA1-dbfA1*. The remaining biodegradation gene classes – which were each found in predicted Rock Bay protein sequences (*alkB*, *benA*, *bph*, *bphA1*, *dbfA2*, *glx*, *mmoX*, *npah*, *p450*, *ppah*, and *ppo*) – were not found on contigs to which SIGEX clones were mapped. In some cases, including *alkB* (alkane-1-monooxygenase), *dbfA2* (dibenzofuran dioxygenase), *glx* (glyoxal oxidase), and *mmoX* (methane monooxygenase), this is because inducers for those genes were not used in the SIGEX experiments. However, representatives from the remaining gene classes should have, theoretically, been detected with the inducers used in SIGEX

experiments; this is supported by a wide variety of literature in Pseudomonads and other organisms. Namely, *benA* (benzoate dioxygenase) from *P. putida* is known to be inducible by benzoate (Cowles et al., 2000); *bph* and *bphA1* (biphenyl dioxygenases) from *Pseudomonas pseudoalcaligenes* KF707 can be regulated by salicylate (Fujihara et al., 2006); *npah* (naphthalene dioxygenase) from *P. fluorescens* is induced by naphthalene and salicylate (Kamath et al., 2004); members of the *ppah* (phthalate dioxygenase) family, such as the *ant* operon from *P. fluorescens* are inducible by benzoate (Retallack et al., 2006). It is unclear whether it would be possible to detect *p450* (aromatic cytochrome p450 enzymes) and *ppo* (benzenediol oxidase) with the inducers used in Chapter 3. The failure to detect these genes within the SIGEX clones may indicate either a shortcoming of the SIGEX library itself, the induction procedure, or a lack of expression of those genes in *E. coli*. Alternatively, this might suggest that the BDG database comparisons were too stringent; using a similarity cutoff of less than the generally accepted 90% may increase the number of genes found, but lower stringency searches would also increase the number of false positive hits.

Many of the initial BLASTp hits that were obtained from the BDG database were removed by the stringent scoring methods that were employed; however, many of the hits that were discarded may still encode genes that are relevant to biodegradation, but have yet to be characterized. Thus, this type of analysis provides a high specificity, in that it is unlikely to identify false positives for biodegradation genes, given the high stringency of the search algorithms; but, it

may have low sensitivity in that many false negatives could arise due to the presence of uncharacterized biodegradation genes.

Overall, the data obtained by NGS is much broader in nature than that obtained by a phenotypic screen. This is advantageous when looking for large-scale trends, such as pathway analysis or taxonomy reports. However, it is also more challenging to narrow the scope of target genes and to determine their actual physiological roles in cells and in the community. This chapter demonstrates that shotgun sequencing of metagenomic DNA can provide a wide range of information to researchers for data mining. In an era where it is becoming more and more feasible to sequence the soil metagenome (Vogel et al., 2009), it is crucial that the massive amount of data is curated and analyzed, and where possible, compared to studies looking at a matched metagenomic sample using several different methodological approaches.

Chapter 6.

Applications of SIGEX for the Design of Whole-Cell Bioreporters

6.1 Introduction

Substrate-induced gene expression (SIGEX; Uchiyama et al., 2005) was initially designed with the intention of recovering novel catabolic operons, as utilized and discussed in Chapters 3 and 4. However, because SIGEX enables the discovery of potentially novel metagenomic clones that are inducible by substances being screened for, we hypothesize that it might also be well-suited as a tool for the recovery of novel bioreporters. Whole-cell bioreporters are a rapid method to measure the bioavailable amount of a test compound. Using a reporter protein such as GFP or LacZ, whose gene is transcriptionally fused to a promoter that is regulated in response to a test compound, it is possible to measure the dose-response relationship between the compound and the induction of the reporter gene. Although several bioreporters have been developed (Hynninen & Virta, 2010), there are still many highly toxic environmental pollutants (found in contaminated sites) for which no bioreporters exist; additionally, increasing pollutant uptake by bioreporter strains may improve detection strategies (van der Meer et al., 2004). In this chapter, we present the argument that the SIGEX methodology can be used as a tool to discover useful novel bioreporters. Furthermore, we show that these novel whole-cell bioreporters (created using various host organisms) can be used

in combination with the FCM model used in SIGEX experiments to explore different physiological parameters affecting bioreporter function.

The efficacy of a bioreporter hinges upon its ability to take up (either actively or passively) the chemical being detected. The structure of the bacterial outer membrane (OM) plays an important role in determining mechanisms of uptake for environmental substances (Hancock, 1984). Moreover, the dynamic interplay with environmental parameters including pH, O₂ content, mineral and ion concentrations, DOC, and temperature – among many others – can alter the properties of the OM, either via direct interactions or through alterations in gene expression (Nikaido, 2003). Since the environment contains complex mixtures of chemicals, many of which may influence membrane permeability, it is important to understand how these factors might influence xenobiotic uptake. LPS is a component of the OM in Gram-negative bacteria which is involved in the protection of cells from environmental insults – especially hydrophobic compounds – and is also responsible for the immune response to bacteria in humans (Wang & Quinn, 2010). LPS is assembled sequentially outward from the inner-most lipid A that is attached to an inner and outer core of sugars (which are conserved in *E. coli* and *Salmonella*, and are usually conserved among families; Heinrichs et al., 1998). These components are assembled on the inner membrane at the cytoplasmic face, and then flipped into the periplasm where core-lipid A is ligated to the O-antigen (a highly variable component of LPS with more than 170 unique O-antigens known in *E. coli*; Heinrichs et al., 1998; Wang & Quinn, 2010). This structure is subsequently flipped to the OM where it comprises the outer-most interface with

the environment. The entire structure is stabilized by various phosphate groups, located on several of the sugars, which form bridges with one another through divalent cations (mainly Mg^{2+} and Ca^{2+}). As the LPS can sequester and prevent the uptake of many small molecules, it is necessary to study its structure in the context of environmental pollutants.

An overarching problem addressed in this thesis is understanding the bacterial responses to pollutants found in contaminated sites. Mercury poses a threat to the health of aquatic ecosystems due to its propensity to bioaccumulate within food webs (Wolfe et al., 1998). Contamination with Hg inevitably leads to the formation of MeHg, which is highly neurotoxic even in small doses; the formation of MeHg is of significant concern because it is produced mainly through biochemical reactions carried out by bacteria (Ullrich et al., 2001). Since bacteria represent the first step in the bioaccumulation of Hg compounds, and some strains also possess the ability to detoxify it through the well-characterized and widespread *mer* operon (Mathema et al., 2011), it is important to understand factors that affect Hg uptake in bacteria (Yamaguchi et al., 2007). It has been shown that Hg uptake is diminished by the presence of Ca and Mg ions in the *E. coli* whole-cell Hg bioreporter pRB28 (Daguené et al., 2012). This leads to the important corollary that within aquatic ecosystems where cation decline occurs, entrance of Hg into the food chain may be enhanced. That Mg and Ca cations stabilize the LPS through cross-bridging between phosphates (Hancock, 1984) indicates that the LPS itself

may be an important factor in Hg uptake; however, the mechanism through which Hg crosses the OM has yet to be elucidated.

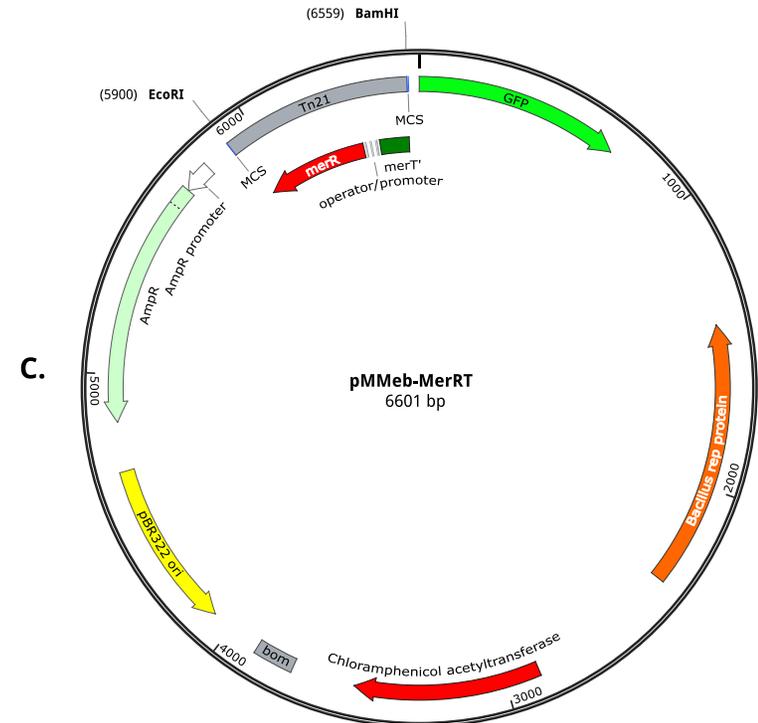
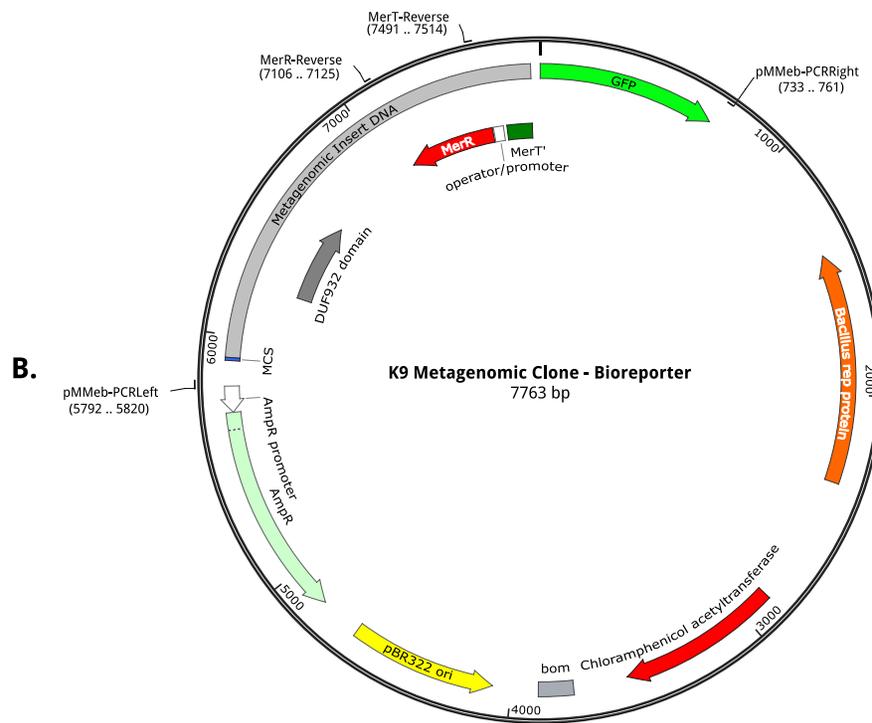
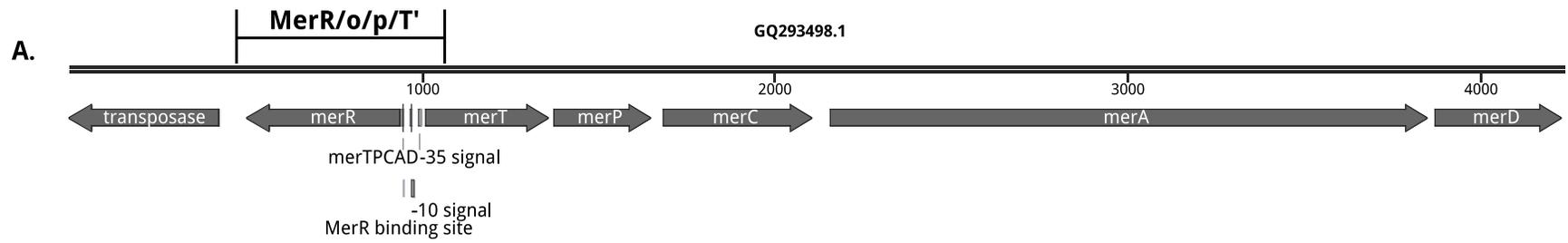
The first objective of this chapter is to show that SIGEX-recovered clones can be used to make novel bioreporter constructs. Second, to examine the possible role of LPS in bacterial permeability to Hg, the novel GFP-based bioreporter was used to assess the uptake of Hg in a variety of LPS-truncated mutants. We hypothesize that in a more truncated LPS, less cross-bridging would occur, and therefore Hg uptake would increase. The effect of Mg is examined in the context of this hypothesis. We report here that truncations in the LPS can enhance Hg uptake, and that the role of Mg in hampering Hg uptake, as reported by Daguéné et al. (2012), appears to be an effect that is independent of the LPS core.

6.2 Methods

6.2.1 Bioreporter Recovery from Metagenomic Samples using SIGEX

A metagenomic library derived from the CFB Petawawa military testing range was screened using SIGEX as described in Chapter 3. A clone containing a mercury-inducible genetic element was recovered. The first 490 bp of the cloned sequence aligned with 100% nucleotide identity to the *merR* and *merT* genes of *Nitrosomonas europaea* ATCC 19718; this clone shares only 85.0% nucleotide identity to the sequence of *E. coli* transposon Tn21 (upon which existing mercury bioreporters are based; Figure 6.1A). Therefore, it was tested directly as a novel whole-cell bioreporter in the mercury bioassay as described below. Furthermore, the MerR/operator/promoter/T' (o/p) region and only the o/p/T' region were cloned separately into the vector pCR2.1-TOPO by PCR amplification (Figure 6.1B). For both constructs, the forward primer pMMeb-PCR-Right (5'-CTCGGCGGATTTGTCCTACTCAAGCTTGC-3') was used, which included amplification of the GFP reporter gene; for the construct including MerR, the reverse primer MerR-Reverse (5'-TGAGCGTGTCGTCATCCATG-3') was used, and for the construct of MerT' and its promoter, the reverse primer MerT-Reverse (5'-GCAAAAAGCGC-CAATGGTCAGATTC-3') was used.

Figure 6.1. The *mer* operon and the K9 bioreporter clone used in this chapter. **(A)** The *mer* operon from *E. coli* Tn21. In the absence of Hg²⁺, MerR represses the operon through binding at the operator/promoter region, as well as its own transcription in a Hg²⁺ independent manner (Ross et al., 1989); in the presence of Hg²⁺, MerR undergoes a conformational change that activates transcription of the operon (Brown et al., 2003). **(B)** The metagenomic clone K9, recovered using SIGEX. Because it contains a partial *mer* operon (similar to *Nitrosomonas europaea*), it exhibits increased GFP expression in the presence of Hg²⁺. Primer binding sites are indicated for experiments where removal of extraneous metagenomic DNA (MerR-Reverse) and deletion of the *merR* gene (MerT-Reverse) was performed. **(C)** The novel GFP-based Hg bioreporter, pMMeb-MerRT, constructed by cloning the EcoRI/BamHI fragment from pRB28 (Selifonova et al., 1993) into the compatible sites in pMMeb. Illustrations created using SnapGene Viewer.



6.2.2 Bioreporter Constructs

The bioreporter constructs (except for the metagenome-isolated bioreporter) used in these experiments are derived from pRB28 (Selifonova et al., 1993). A BamHI/EcoRI fragment carrying the MerR/o/p/T' region from pRB28 (indicated in Figure 6.1A) was cloned into pMMeb into the corresponding restriction sites upstream of GFP. The resulting construct, pMMeb-MerRT (Figure 6.1C), was transformed into *E. coli* BW25113 (wild-type strain), and a variety of mutants containing knockouts for various LPS assembly genes, including *rfaF*, *rfaG*, *rfaI*, *rfaJ*, *rfaL*, *rfaQ*, *rfaS*, and *rfaY*, all of which were obtained from the Keio collection (Baba et al., 2006) and are described in Table 6.1. Transformations of the bioreporter plasmids into the various hosts were done using standard methods for calcium chloride competent cells (Sambrook & Russell, 2001).

Table 6.1. Mutated genes in *E. coli* used for bioreporter membrane permeability experiments. Functions from Heinrichs et al. (1998).

Mutated gene	Alternate gene names	Function in LPS assembly
<i>rfaF</i>	<i>waaF</i>	HepII transferase
<i>rfaG</i>	<i>waaG</i>	HexI transfer: UDP-glucose:(heptosyl) LPS α 1,3-glucosyltransferase
<i>rfaI</i>	<i>waaI, waaO</i>	HexII transfer: UDP-galactose:(glucosyl) LPS α 1,3-galactosyltransferase; HexII transfer: UDP-glucose:(glucosyl) LPS α 1,3-glucosyltransferase
<i>rfaJ</i>	<i>waaR, waaT</i>	HexIII transfer: UDP-glucose:(glucosyl) LPS α 1,2-glucosyltransferase; HexIII transfer: UDP-galactose:(glucosyl) LPS α 1,2-galactosyltransferase
<i>rfaL</i>	<i>waaL</i>	Lipid A core:surface polymer ligase
<i>rfaQ</i>	<i>waaQ</i>	HepIII transferase
<i>rfaS</i>	<i>waaS, wabA</i>	Unknown; possibly involved in the formation of α -Gal-1 \rightarrow 7-Kdo substitution or in the formation of α -Rha-1 \rightarrow 5-Kdo substitution
<i>rfaY</i>	<i>waaY</i>	Involved in phosphorylation of HepII

6.2.3 Mercury Bioreporter Assay

Assays were carried out as described in Daguéné et al. (2012). A single colony inoculated in 3 mL of LB + antibiotics. This culture was incubated for 6 h at 37 °C with shaking at 200 rpm, then diluted 1:100 into 5 mL of GMM + antibiotics (Difco M9 salts supplemented with 0.3% glucose, 1mM MgSO₄, 1 µg/mL Thiamine, and trace elements (Barkay et al., 1998), and incubated for 16 h. This culture was diluted in 20 mL GMM + antibiotics, and incubated under the same conditions until an OD 600 of 0.6-0.7 was reached. The cells were harvested by centrifugation and washed with 67 mM Pi (NaH₂PO₄/K₂HPO₄, pH 7.1), then re-suspended in an equal volume of 67 mM Pi. The OD 600 was set to 0.4, and a 1:100 dilution of this was used in the assay. Assay medium was 67 mM Pi, 0.9 mM (NH₄)₂SO₄, and 5 mM glucose. If magnesium was added to the assay medium, it was diluted from a 1 M MgSO₄ stock into a separate stock of assay medium to a final concentration of 1 mM. Cells were pre-incubated in the assay medium for 15 min prior to the addition of mercury. The assay was started by the addition of 100 ng/L HgCl₂, and each experiment included a non-mercury containing blank control.

6.2.4 Measurement of Bioreporter Gene Expression

In contrast to Daguéné et al. (2012), GFP was used as the reporter gene instead of luciferase. GFP fluorescence was measured using the FACSaria flow cytometer 6 h after the assay was initiated. Fifty-thousand events were collected for each sample using a 488 nm laser and fluorescence was measured using a 530/20 nm band-pass filter. Experiments were carried out using at least 3 biological replicates. Refer to Chapter 3 for detailed flow cytometry methods.

6.3 Results

6.3.1 Characterization of a Novel Bioreporter Recovered using SIGEX

A mercury inducible clone, designated K9 following its isolation from the CFB Petawawa contaminated site metagenome using SIGEX, was exposed to increasing concentrations of mercury and observed using flow cytometry to measure GFP expression (Figure 6.2). By using flow cytometry, it was possible to account for variations in cell numbers, and obtain an average “per-cell” quantification of gene expression. As shown in Figure 6.2, it is possible to visualize gene expression for the entire cell population, allowing the determination of measures of spread and the population’s shape, in addition to a single mean value for each sample. A dose-response relationship was observed during mercury induction (Figure 6.3), showing a linear increase in GFP fluorescence with increasing concentrations of mercury. Linear regression showed an r^2 value of 0.9991 when the bioreporter was examined in *E. coli* strain GS071 and an r^2 value of 0.998 in the strain DH10b. The promoter remained inducible following cloning of the MerR/o/p/T’ region (using PCR to remove the upstream surrounding metagenomic DNA), but inducibility was abolished by the removal of MerR.

Figure 6.2. Mercury induced expression of clone K9 (in *E. coli* DH10B) isolated from the Petawawa military contaminated site metagenome using SIGEX. Boxes showing histograms of GFP expression represent increasing concentrations of mercury, from left to right. Red lines indicate expression of the blank no-mercury control, and green lines indicate expression of the mercury-induced culture.

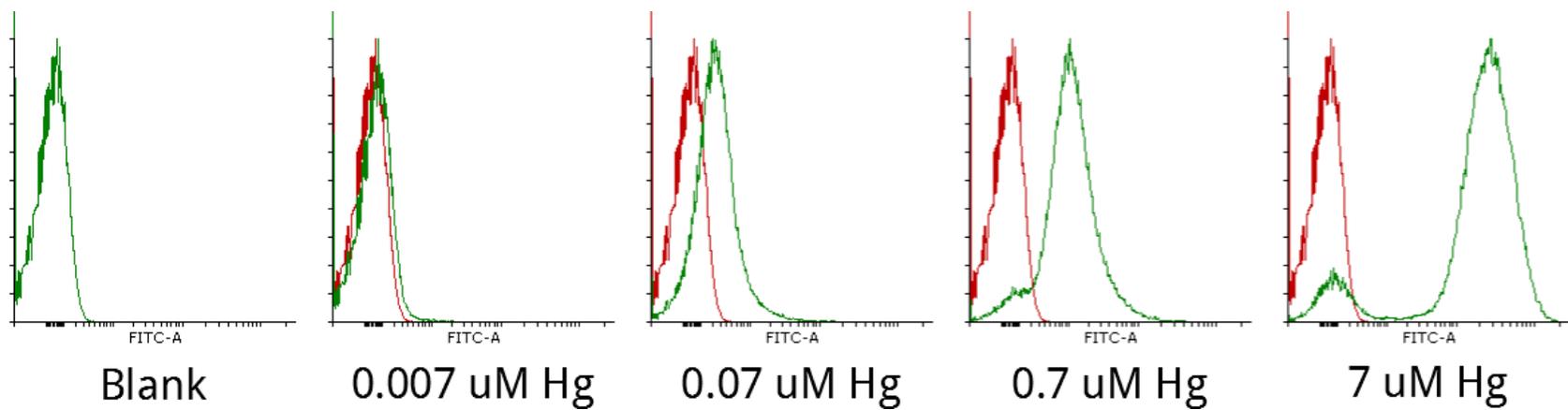
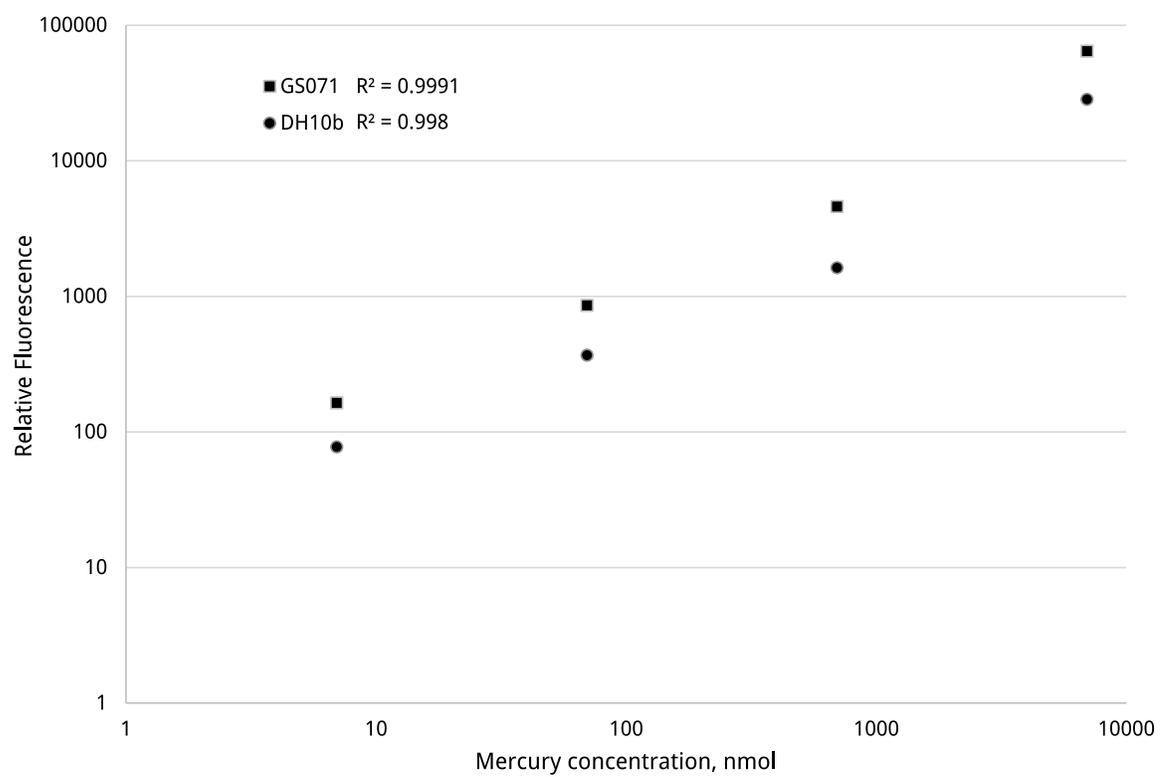


Figure 6.3. Mercury induction of metagenomic bioreporter clone K9 in *E. coli* DH10b and GS071. Flow cytometry was used to measure GFP expression. Regression analysis of individual replicates indicates that the dose-response correlates linearly (r^2 of 0.9991 and 0.998 in GS071 and DH10b, respectively) with increasing amounts of mercury. Strain GS071 shows a slightly elevated response relative to DH10b.



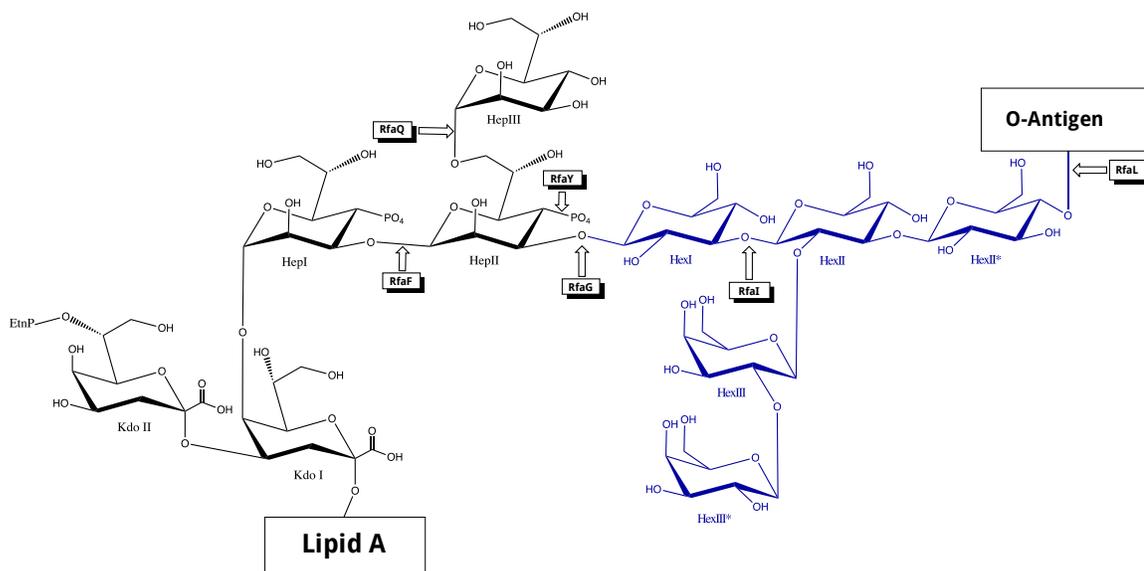
6.3.2 Bioreporter Assays in LPS Mutants using Flow Cytometry Analysis

We measured bioavailable Hg (using the pMMeb-MerRT construct) in a variety of *E. coli* mutants that possessed truncations in the LPS to different degrees. The degree of LPS truncation generally corresponded to a higher amount of Hg uptake (Figure 6.4), with *rfaF* showing the most uptake, and *rfaL* showing the least, relative to wild-type BW25113. Hg uptake is enhanced in the mutants *rfaS*, *rfaQ*, *rfaG*, *rfaI*, *rfaJ*, and *rfaF* between 1.17-fold (*rfaS*) and 1.75-fold (*rfaF*). With *rfaY* and *rfaL*, Hg uptake is *decreased* 1.38 and 1.10 fold, respectively. Data for Hg uptake are reported as the fold-change in relative fluorescence units compared to a non-Hg containing blank.

We measured Hg uptake in the LPS mutants with pre-exposure of the bioreporter cultures to 1 mM Mg^{2+} in the assay media. The addition of Mg^{2+} resulted in decreased Hg uptake in nearly all cases. The amount by which Hg uptake was altered is shown in Figure 6.5. The reduced uptake in the presence of Mg^{2+} does not appear to correlate with the amount of LPS present, as the presence of 1 mM $MgSO_4$ decreases Hg uptake regardless of the extent to which the LPS is truncated, with no apparent relation to where the mutation is found. In only a single instance, the *rfaG* mutant, Mg^{2+} is shown to increase Hg uptake.

Figure 6.4. Effect of LPS truncations on Hg uptake **(A)** The LPS structure and synthesizing enzymes of the *E. coli* R1 core oligosaccharide, with arrows indicating the point of truncation corresponding to each mutant. Black residues are part of the inner core; blue residues constitute the outer core. Inner and outer cores are highly conserved between species while the O-antigen accounts for a large degree of variation. Figure created by Mike Jones, licensed under Creative Commons, 2010, and modified by Matt Meier. **(B)** Response of the pMMeb-MerRT bioreporter to mercury, using various LPS mutants (each shown in the schematic above) as the host organism. The order on the bar graph from left to right represents increasing Hg uptake; more uptake is observed in mutants with highly truncated LPS (*rfaG*, *rfaI* and *rfaF*; $p < 0.05$ compared to WT is denoted by asterisks), while those with little effect on the inner LPS (e.g., *rfaL*, which prevents ligation of O-antigen) show a response similar to the BW25113 wild-type. Inductions were performed using 100 ng/L HgCl₂; values are calculated from the average of three biological replicates, and error bars represent standard error of the mean.

A. Structure of *E. coli* LPS and enzymes involved in its synthesis.



B. pMMeb-MerRT bioreporter response in various LPS mutants.

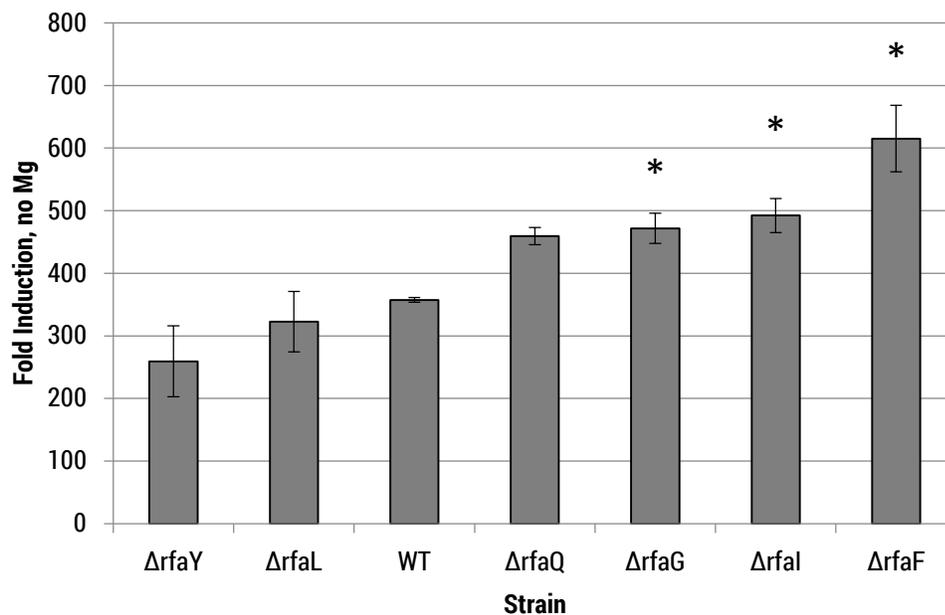
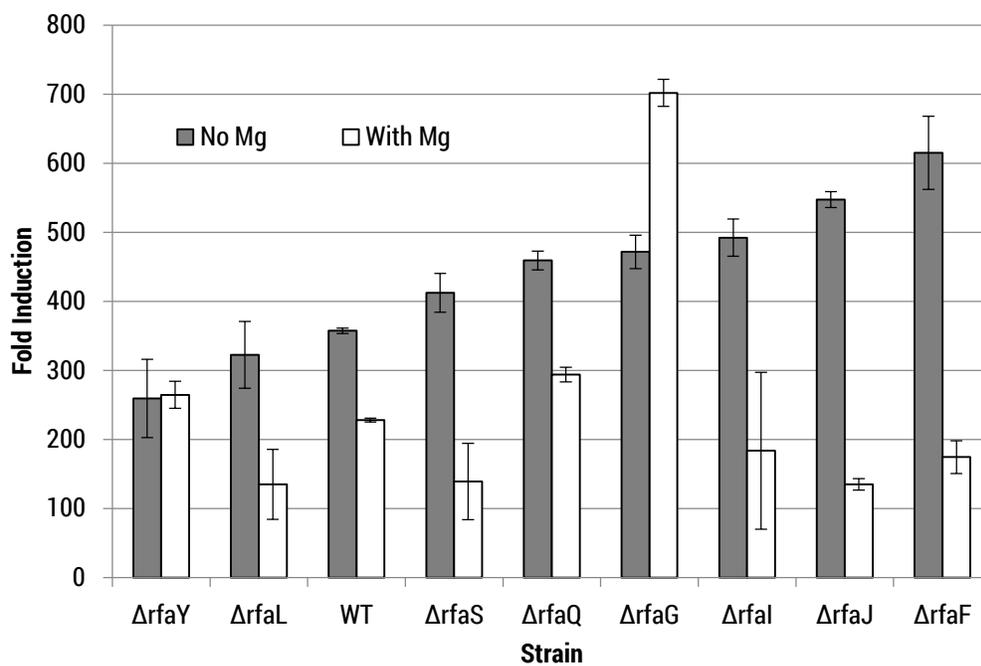
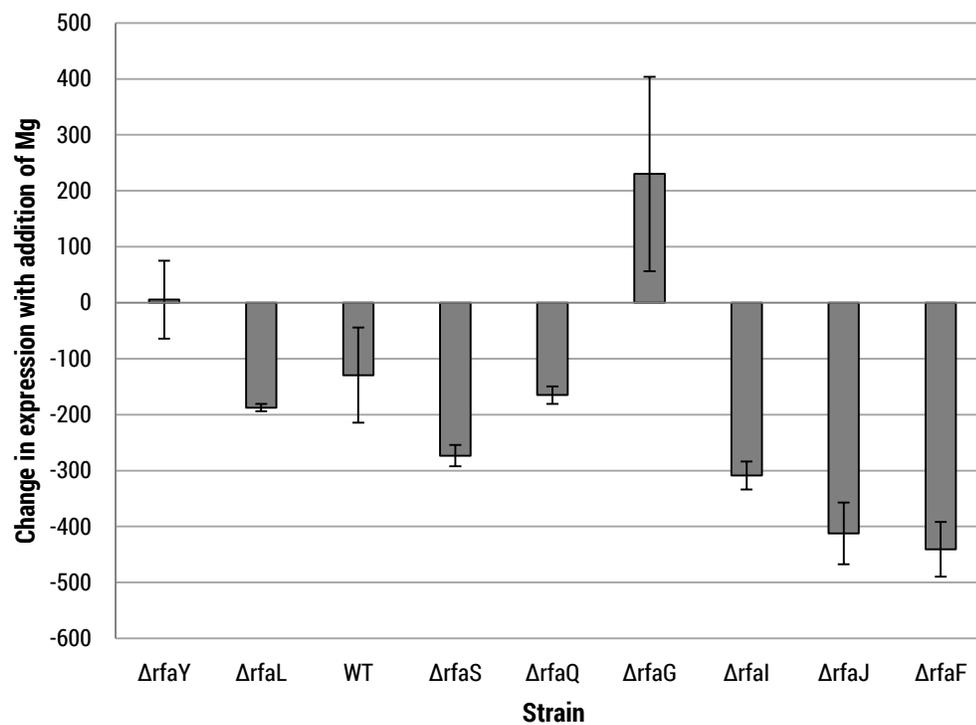


Figure 6.5. Effect of Mg^{2+} on Hg uptake measured by flow cytometric analysis of GFP expression in various bioreporter hosts carrying the pMMeb-MerRT construct. Inductions were performed using 100 ng/L $HgCl_2$. **(A)** Fold induction of GFP expression, relative to the blank, measured with and without pre-incubation of the bioreporter cells with 1 mM Mg^{2+} . Error bars represent standard error of the mean, calculated from three biological replicates. **(B)** The amount by which expression (measured in arbitrary fluorescence units), relative to the blank, is changed by pre-exposure of the bioreporter cells to 1 mM Mg^{2+} . The order from left to right represents increasing Hg uptake in the *absence* of Mg. Error bars represent the standard deviation of three biological replicates.

A. Effect of pre-incubation with Mg on fold induction.



B. Absolute differences of expression following pre-incubation with Mg.



6.4 Discussion

6.4.1 Using Metagenomic Clones as Novel Bioreporters

The main function of a bioreporter is to reveal an increase in the expression of a reporter gene in the presence of a chemical of interest. Thus, the discovery of novel biosensing genetic elements is well suited to metagenomic methods such as SIGEX because they are already designed to use a reporter gene for the recovery of genes that are transcriptionally induced by a chemical of interest. By exploiting the diversity present in metagenomic samples, it may be possible to use SIGEX to recover many new transcriptional units that are upregulated by a wide variety of chemicals, and thereby design a multitude of bioreporters for various chemicals. This chapter demonstrated the use of SIGEX in the discovery of a metagenomic fragment that functions as a novel whole-cell bioreporter based on a clone recovered from a metagenomic plasmid library. The clone, K9, expressed GFP with a linear dose-response to increasing concentrations of mercury, demonstrating its utility as a bioreporter (Figure 6.3). The sequence of the K9 clone corresponded to the *merR* transcription factor gene, an intergenic region containing the *mer* operator and promoter region, and the partial *merT* gene, aligning with high confidence to those genes from *Nitrosomonas europaea*. By comparing the cloned MerR/o/p/T' region to only the o/p/T' region from this metagenomic fragment (knocking out the transcription factor), it was shown that the *merR* gene and the *merT* promoter were responsible for the observed induction.

Using a metagenomic fragment containing the partial *mer* operon enabled a comparison to the previously described pRB28 bioreporter, which uses the *lux*

reporter system (Selifonova et al., 1993). Although we found that GFP was generally less sensitive compared to the existing bioreporters, detecting nanomolar as opposed to picomolar quantities (van der Meer et al., 2004), it has the advantage of relying on a more persistent reporter protein (GFP has a half-life of more than one day, while luciferase's is approximately 3 h). Furthermore, since GFP is endogenously fluorescent, it is not as resource-intensive for the cells as is light production through luciferase. The high ATP requirement of luciferase may cause metabolic stress that could alter the overall cellular physiology. Finally, GFP is easily measured using flow cytometry, which enables single-cell measurements; this can provide data regarding the population distribution of gene expression, which is more informative than an average reading (Figure 6.2).

6.4.2 Mutations in the LPS Core Result in Increased Hg Uptake in Bioreporters

We show that truncations in the LPS coat (particularly those affecting sugars upstream the O-antigen, *i.e.*, within the inner and outer core) result in enhanced uptake of Hg as measured by the Hg bioreporters described here. The only LPS mutants that we tested that do not show increased Hg uptake are *rfaL* and *rfaY* ($p=0.35$ and $p=0.18$, respectively). The enzyme encoded by *rfaL* is thought to be the O-antigen ligase (Wang & Quinn, 2010), and therefore would not influence core interactions (since *E. coli* K-12 strains do not contain the O-antigen). Thus, it is not surprising that its Hg uptake is similar to the wild type. In contrast, *rfaY* is involved in the phosphorylation of the HepII sugar in the inner core (Heinrichs et al., 1998) and its mutation causes sensitivity to crystal violet, indicating that it

increases permeability of the OM. However, it does not appear to be more permeable to Hg (relative to the wild type), even though all the other *rfa* mutants tested here show increased Hg uptake. Overall, our results indicate that the LPS is likely involved in inhibition of Hg uptake. This imposes important implications for organisms in the environment, insofar as any conditions which destabilize the LPS could potentially allow for more Hg to enter into bacterial cells, enabling inorganic Hg ions to ultimately accumulate as more toxic and biologically active Hg in food chains. Because of this, it may be important to monitor environmental factors that affect outer membrane permeability, particularly the presence of divalent cations, in environments where elemental Hg is at risk of being absorbed.

6.4.3 Magnesium Hampers Hg Uptake in Bioreporters Independent of LPS Truncations

This study also confirms the results of Daguéné et al. (2012), where it was shown that a variety of divalent base cations (Mg and Ca) hamper Hg uptake through a hitherto unknown mechanism. In this study, a different reporter gene (GFP as opposed to luciferase) and different strains of *E. coli* (BW25113 and derivatives, as opposed to HMS174) were used to reach the same conclusion. This indicates that the effect of divalent base cations hampering Hg uptake is a biologically significant observation, independent of construct and bacterial strain.

The addition of excess Mg can cause the LPS to become more crystalline (Nikaido, 2003) and may decrease the permeability of the OM. When present in millimolar quantities, Mg forms bridges between phosphate groups in the LPS resulting in an impermeable mesh-like structure which is more stable and melts at

a higher temperature than Mg-depleted LPS (Nikaido, 2003). As such, our hypothesis was that in a more truncated LPS, less cross-bridging would take place (especially when phosphates were removed), and as a consequence, Hg uptake would be less affected by Mg treatment. However, this was not observed: therefore we are forced to conclude that there is some other reason that Mg hampers Hg uptake (such as Mg occupying binding sites for Hg that would otherwise be available to sequester excess Hg). The notable exception to this is *rfaG*, where the addition of Mg actually *enhances* Hg uptake. This particular mutant is truncated at a point that readily exposes a phosphate group (Figure 6.4A) to the outer portion of the LPS. We speculate that the LPS structure in the *rfaG* mutant might interfere with a binding mechanism for divalent cations, resulting in less membrane sequestration of both Mg and Hg, and therefore enhancing uptake.

6.4.4 Conclusions

This chapter demonstrated the use of SIGEX for the discovery of a novel biosensing genetic element from a metagenomic library. Furthermore, it was shown that whole-cell bioreporters can be used with SIGEX-based flow cytometric analysis to perform bioreporter assays to identify dose-response relationships under a variety of experimental conditions. Furthermore, we show that the bioavailability of environmentally relevant substances such as Hg depend to some extent on properties of the outer membrane in bioreporters. Not only might this speak to the physiology of environmental organisms involved with xenobiotic transformation, but it may also be used to design bioreporters with increased sensitivity, through

the incorporation of LPS mutants as part of a battery of hosts used for bioreporter constructs.

Although it is possible to detect and identify xenobiotics using a wide variety of physical chemistry methods (*e.g.*, mass spectrometry), the methods used to do so have certain disadvantages: for example, they lack the ability to discern the bioavailability of the substances being analyzed, and they tend to be expensive, requiring specialized personnel to operate. As a provisional alternative, whole-cell bioreporters can quickly and inexpensively determine the bioavailable concentrations of chemicals via light or fluorescence production (*lux* or GFP), as well as through other reporter genes. For that reason, this work emphasizes the importance of uncovering new classes of bioreporters and understanding how they interact with host physiology in order to optimize their use.

Chapter 7.

Summary and Conclusions

7.1 Summary of Findings

The objective of this thesis was to study, in a culture-independent fashion, the role of diverse organisms in the transformation of xenobiotics found at contaminated sites. Two separate shortcomings of current metagenomic methods were addressed throughout this thesis: 1) the limited depth of sequence data used in most metagenome sequencing studies was overcome using high-depth reads of NGS in a contaminated soil community, and 2) the conventional reliance of functional gene characterization using similarity-based approaches was circumvented by the use of a gene-expression assay to enable the identification of genetic elements based on their transcriptional regulation. The conclusions from each chapter, as they relate to the broad goal of understanding xenobiotic transforming microbes found in contaminated sites, are summarized in this section.

Chapter 2

Objectives Evaluate and summarize our current understanding of aromatic hydrocarbon degrading genetic elements, using published data, to determine what aspects of our understanding of aromatic degradation remain uncharacterized.

Conclusions Chapter 2 aimed to survey existing knowledge on aerobic aromatic hydrocarbon degradation in bacteria, and how the genes involved are regulated at the transcriptional level. It was revealed that the regulation of PAH catabolising genes is poorly understood. Furthermore, most literature on the physiology of PAH metabolism derives from a relatively limited taxonomy (primarily *Mycobacterium* and *Sphingomonas*), which led us to the conclusion that additional information may be gained by using a variety of culture-independent methods to gain an accurate perspective on their catabolic capacity. The results from Chapters 3 to 5 address this issue.

Chapter 3

Hypothesis Aromatic-degrading elements are carried by individuals within the microbial community of a PAH contaminated site, and we can access a subset of those elements using SIGEX.

Conclusion The hypothesis was supported. By using SIGEX to screen for inducible genetic elements from uncultured bacteria in contaminated sites, we uncovered a multitude of metagenomic fragments that were

upregulated by various low-molecular weight aromatics (benzoate, salicylate, phenylacetic acid, catechol, naphthalene, and phenol). The DNA sequences contained in these clones, although limited in length, contained genes or gene fragments that are physiologically relevant to aromatic catabolism.

Chapter 4

Hypothesis Massively parallel NGS data can be integrated with Sanger reads from SIGEX to obtain the sequences of the regions surrounding SIGEX-recovered genetic elements.

Conclusion A major limitation of library-based studies is that plasmid inserts only contain several kb of DNA, which limits the number of genes that can be characterized; therefore, NGS was used to obtain over 125 Gb of 100 bp PE reads from the matched metagenomic DNA sample. The assembled sequences (approx. 400 Mb with an N50 of 9.1 kb) enabled mapping of the SIGEX-derived sequences to operon-sized contigs. This supported the hypothesis, and complemented the SIGEX clone sequences by providing a method to predict the identity and genomic context of upstream and downstream genes. It was possible to annotate genes, and, in many cases, entire operons (Appendix B), that were found on those contigs matching with a high statistical probability (*i.e.*, E-value = 0.0) to the SIGEX-derived sequences. As expected based on the sequence analysis from Chapter 3, the

surrounding annotated genes were functionally related to aromatic metabolism.

Chapter 5

Hypothesis Genes previously known to be involved in biodegradation that were *not* retrieved using metagenomic library screens may still be detectable within metagenomic NGS data; thus, some PAH-degrading genes may be present in the Rock Bay soil sample despite their absence in SIGEX experiments.

Conclusions The MG-RAST annotation pipeline and the biodegradation gene database were used to identify previously characterized aromatic-degrading gene sequences within the metagenomic data, as well as the taxa associated with those gene classes. The findings reveal that the taxonomic groups found on SIGEX clones, and the surrounding sequences (Chapters 3 and 4), are consistent with the taxa associated with biodegradation as determined by those *in silico* analyses, in that it was predominantly *Pseudomonas* species that were found to carry biodegradation genes. However, a notable difference is that more gene classes were recovered using the *in silico* tests. Moreover, PAH degrading genes were found (*e.g.*, *carA*) using the biodegradation gene database, but not SIGEX, supporting the hypothesis. Although some gene classes were excluded during library screening, the use of

similarity-based search tools enabled the identification of other potentially relevant sequences for biodegradation.

Chapter 6

Hypothesis SIGEX can be used for the design of novel whole-cell bioreporter constructs. These, and existing constructs, can be improved by the use of bacterial hosts with truncated lipopolysaccharide components.

Conclusion The hypothesis that bioreporters can be discovered using SIGEX was supported by the recovery of a mercury-inducible element from a metagenome library. The GFP reporter gene induction of this clone was found to increase with mercury concentration and it was shown that this response was the result of a mercury-responsive transcriptional regulator and promoter region upstream of a partial gene. Furthermore, it was shown through bioassays using LPS mutants that uptake of environmental contaminants may be enhanced through LPS truncation.

7.2 Contributions to Scientific Knowledge

The experiments undertaken in this thesis have contributed a significant quantity of metagenomic sequence data to existing databases and demonstrated that this data can be mined for genetic elements both *in silico* and in conjunction with *in vitro* gene-expression analyses. This thesis described the enhanced insights that were obtained, through the amalgamation of several independent metagenomic methodologies, for genetic elements involved with xenobiotic transformation in contaminated sites. By using multiple, independent approaches to characterize the same metagenomic sample, we were able to gain more insight into the relevance of particular genes and their potential for biodegradation capacity or their use in other biotechnological applications.

Based on the results of Chapter 2, it was evident that gaps persist in our understanding of how many xenobiotic-degrading genes are regulated. Therefore, in Chapter 3, a previously described (Uchiyama et al., 2005), though seldom used (Ekkers et al., 2012), methodology (SIGEX) was revised through the use of a novel plasmid vector, superior metagenomic DNA library creation, and advanced high-throughput single-cell sorting techniques, for recovering genes involved in xenobiotic transformation based on their transcriptional activity in the presence of an inducing compound. SIGEX was used to recover genes that were inducible by a variety of xenobiotics found in contaminated environments, and subsequently to obtain the DNA sequences comprising those genes, enabling the identification of

many biologically relevant domains and partial genes that were putatively responsible for degradation or elimination of aromatics, or other xenobiotics tested.

In Chapter 4, the sequences from Chapter 3 were expanded outwards and mapped to contigs derived from NGS data; annotated sequences contained complete, and in some cases, novel operons for xenobiotic catabolism. We used this data to explain the genomic context for each of the SIGEX-derived clones recovered in Chapter 3 (Appendix A). This is, to the best of the author's knowledge, the first time that functionally-derived sequences have been mapped to *de novo*-assembled shotgun sequenced metagenomic NGS data for the purpose of annotating the surrounding sequence. The fact that SIGEX-recovered clones 1) were sometimes recovered more than once in a single induction, and 2) contained independent restriction fragments that sometimes aligned to the same original sequence, suggests that the library was being exploited to the fullest extent possible with this particular host organism. Further increases in library size would likely have diminishing returns on the diversity of genes recovered.

Together, Chapters 3 and 4 constitute a novel and effective approach for characterizing metagenomes by combining a phenotypic screen with NGS sequence data and several relevant database annotation and search tools. We found that most of the *de novo* assembled sequences that were aligned to SIGEX clones contained high-confidence matches to existing elements; however, in some cases, it was shown that the aromatic-inducible sequences may represent novel arrangements of genes, some of which even contain easily discernable variation within the microbial populations (see Contigs 243 and 3721 in Appendices A1 and

A4, respectively). This supports the hypothesis that SIGEX can be useful for the discovery of novel inducible genetic elements. This work can be applied more generally in the sense that any functional screen (*i.e.*, not limited to SIGEX) can be used as input for the mapping process. In many cases, metagenomic screens of clone libraries or amplicon-based methods entail the study of comparatively short sequences (*i.e.*, a few kb). These types of investigations could be drastically improved by the incorporation of NGS data to obtain a more definitive context for the applicable metagenomic fragments. Furthermore, this means that past studies might be retrospectively examined in light of the new NGS technologies that have been developed, provided that archived samples of metagenomic DNA are still available for sequencing.

The benefit of employing *in vitro* gene expression systems such as SIGEX for the screening of relevant genes is that one can rapidly reduce the metagenomic space that must be analyzed (Kakirde et al., 2010) in a manner that is independent of previously characterized sequences (*i.e.*, no prior knowledge of sequence similarity is required to discover a physiologically relevant hit). However, this type of screening procedure is limited by the biological host in which it is done: specifically, heterologous expression will not always result in proper gene expression (Handelsman, 2004), due to the absence of transcription factors, insufficient uptake of the test compound, or other enigmatic genetic or environmental factors. Therefore, to complement the SIGEX screening (Chapter 3) and characterization of those clones using NGS contigs (Chapter 4), the focus of Chapter 5 was to determine the presence of biodegradative genes based solely on

sequence similarity. By using targeted database searches (*i.e.*, the biodegradation gene database) for manually curated genes that are well-characterized, we determined with high confidence the catabolic capacity of the PAH-contaminated metagenome in a manner that was independent of our phenotypic screen. This analysis revealed that many functional genes were present in the Rock Bay PAH-contaminated soil metagenome that were not detected by the phenotype-based SIGEX screening.

7.3 Future Directions

Many facets of this thesis could be pursued as areas of further investigation. Of the SIGEX clones analyzed, each could be characterized biochemically, including: the enzyme activity and specificity of the proteins encoded by genes contained therein; the promoter sequences and transcription regulators that are responsible for the observed inductions could be determined; site-directed mutagenesis could provide important information about enzyme functions. Furthermore, those clones could be developed into novel whole-cell bioreporters for the chemical assessment of bioavailable contaminants present in environmental samples. In addition, the assimilation of heavy isotopes of xenobiotic compounds could be performed using DNA-SIP during bioslurry treatments such as those used in Chapter 3, allowing even more efficient screening of metagenomic fragments that are involved in biodegradative processes by further narrowing the metagenomic space being explored. Combining DNA-SIP with SIGEX could potentially increase the ratio of positive clones discovered. As a supplement to SIGEX, using NGS to shotgun-sequence a DNA-SIP metagenome would increase the

likelihood of obtaining much longer contigs. This could even, perhaps, result in complete draft genomes of uncultured bacteria if sufficient read depth is attained, and a sequencing technology is used that provides longer reads (greater than the 100 bp reads used in this thesis) such that long repeats can be resolved during assembly. The *in silico* annotations that were already determined by this thesis could also be used for targeted cloning for the retrieval and characterization of biodegradation genes. Additionally, the metagenomic DNA sequences from this project, which are available on the MG-RAST server, can be mined in future studies for genes of interest that were not discussed in this thesis (e.g., DNA repair genes responding to genotoxic xenobiotics).

7.4 Concluding Remarks

Microbes (and living organisms in general) contain many complex genetic components that we are still in the early stages of understanding; consequently, comparative genomics must draw on many types of information to understand how differences in DNA sequences between organisms are rendered into diverse phenotypes. This thesis, rather than regarding phenotype-based screens and sequence-based screens as methods at odds with one another, recognizes that they provide balanced information that can, and should, be used cooperatively. The research presented here constitutes a unique approach for understanding metagenomes, and it is anticipated that such targeted analyses will be helpful for subsequent studies of environmental microbes.

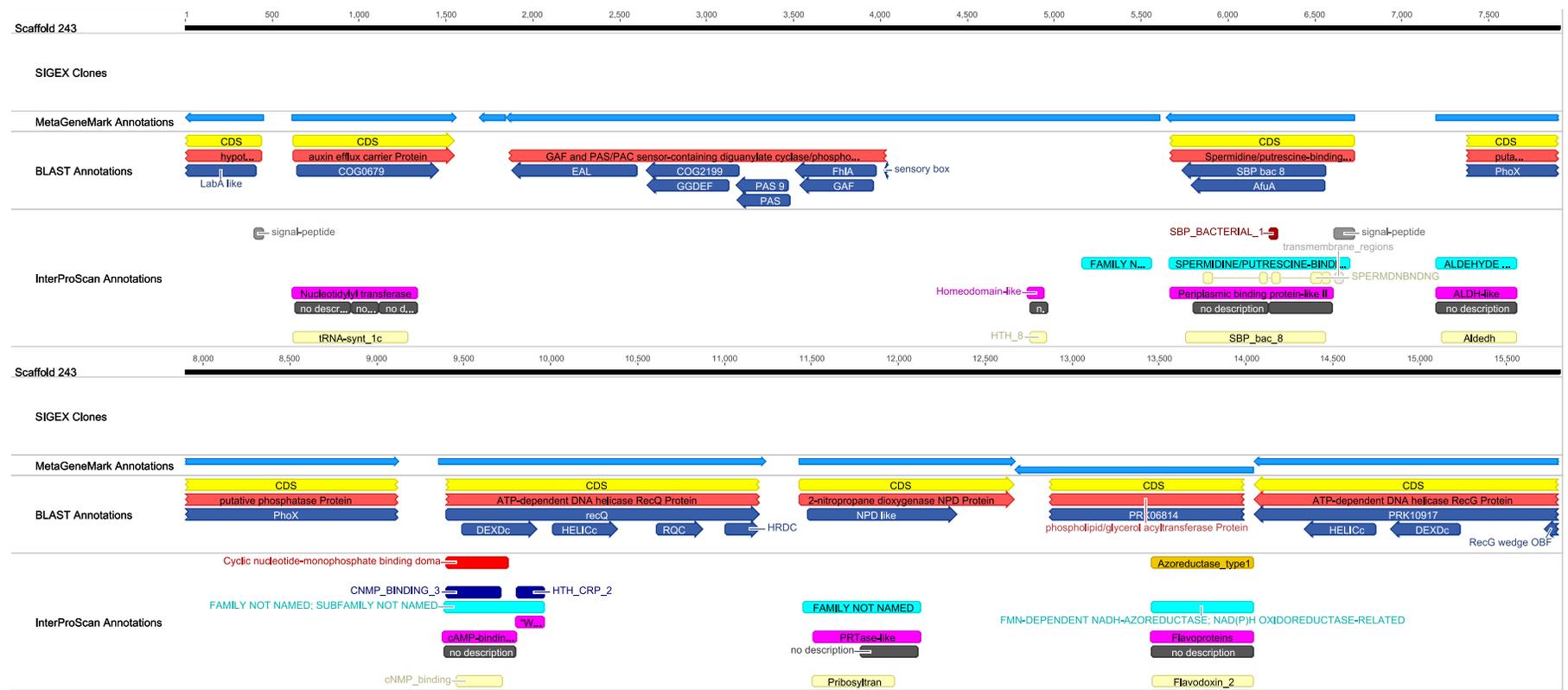
Appendices

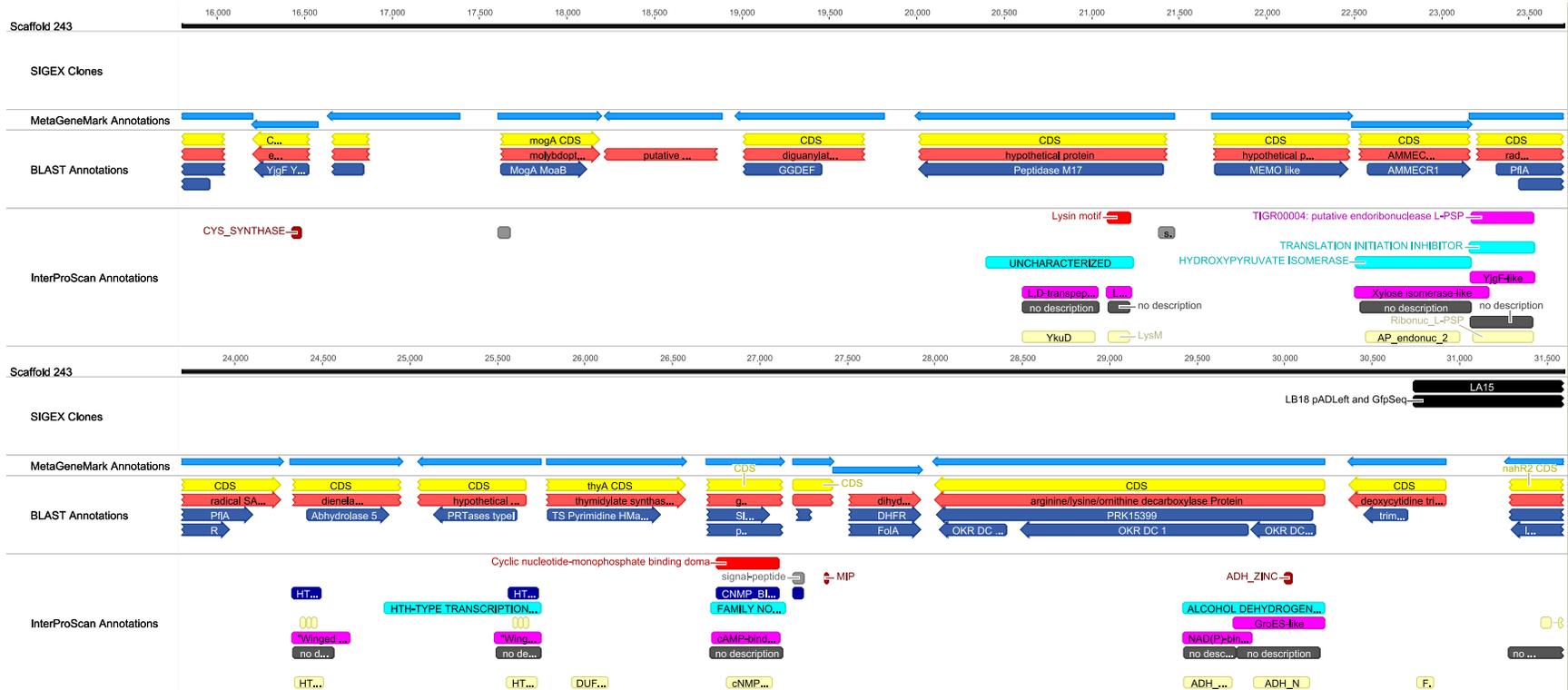
Appendix A. SIGEX-derived Clones Mapped to NGS Contigs

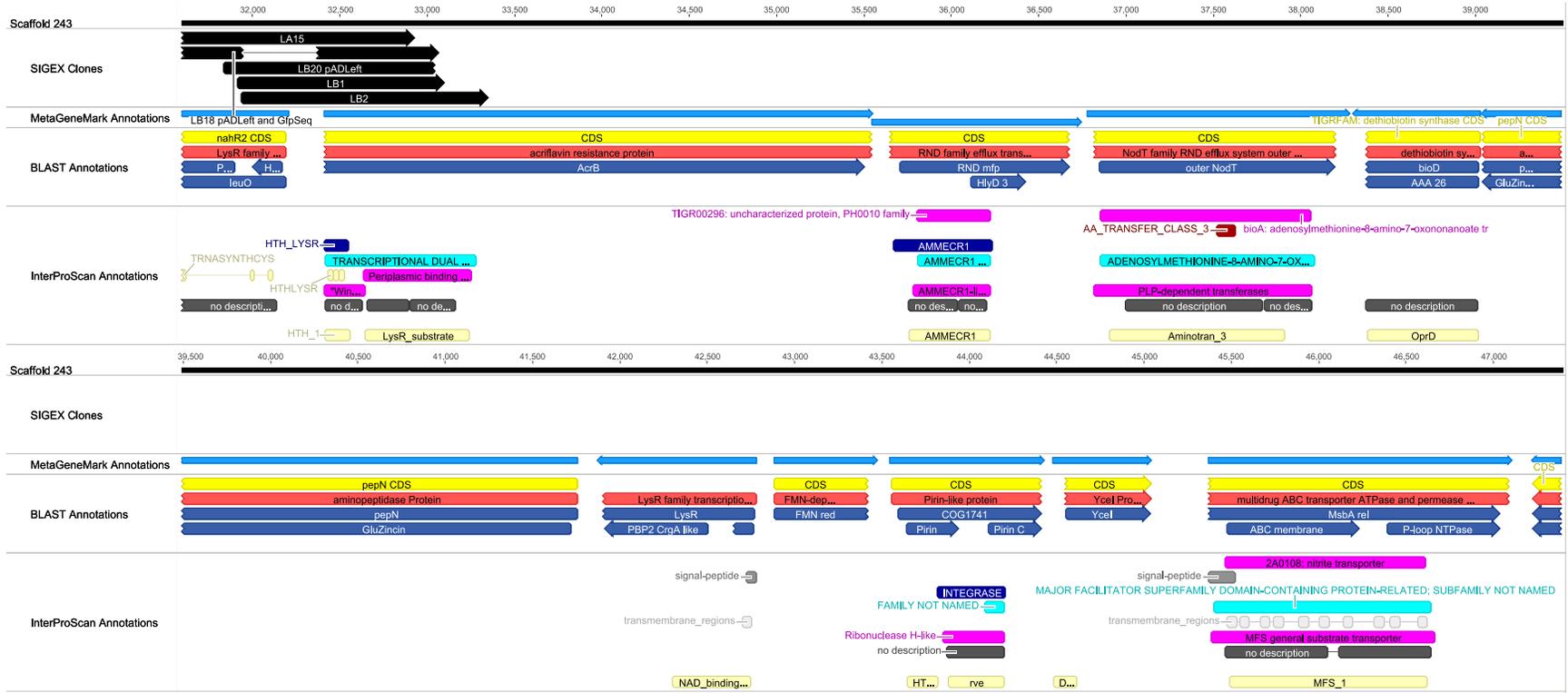
The figures in this appendix show alignments (performed by Geneious V6.0) between contigs (assembled by IDBA-UD) and consensus sequences of SIGEX-recovered aromatic inducible clones. Contigs are shown with MetaGeneMark annotations and their respective BLASTp and InterProScan hits. SIGEX clones are shown as black annotations, with arrow heads pointing in the direction of gene expression. Captions indicate the pairwise identity between each SIGEX clone and the contig.

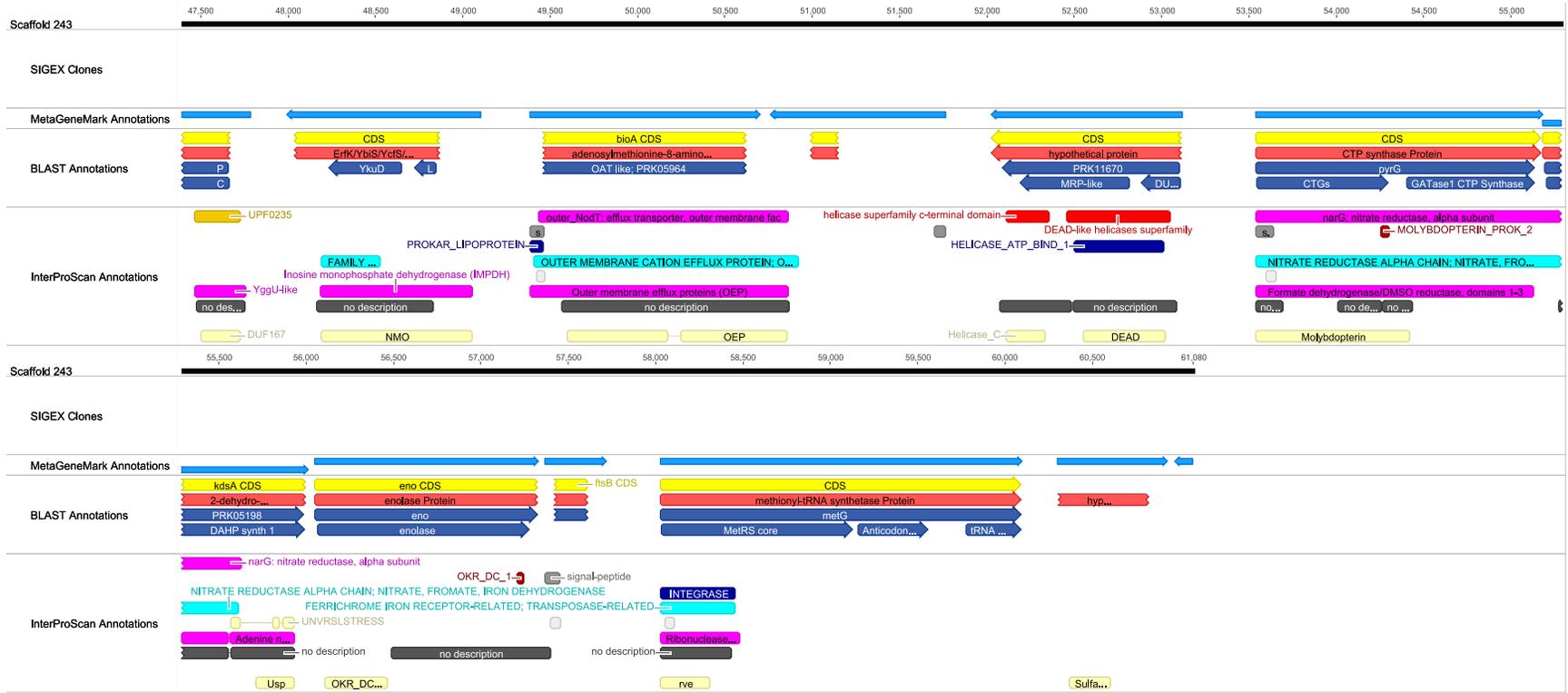
A.1. Contig 243.

Contig 243 aligned to LB18-pADLeft (91.2%), LB18-GfpSeq (100.0%), LA15 (99.0%), LB20-pADLeft (94.6%), LB1 (98.7%) and LB2 (97.9%).









A.2. Contig 3075.

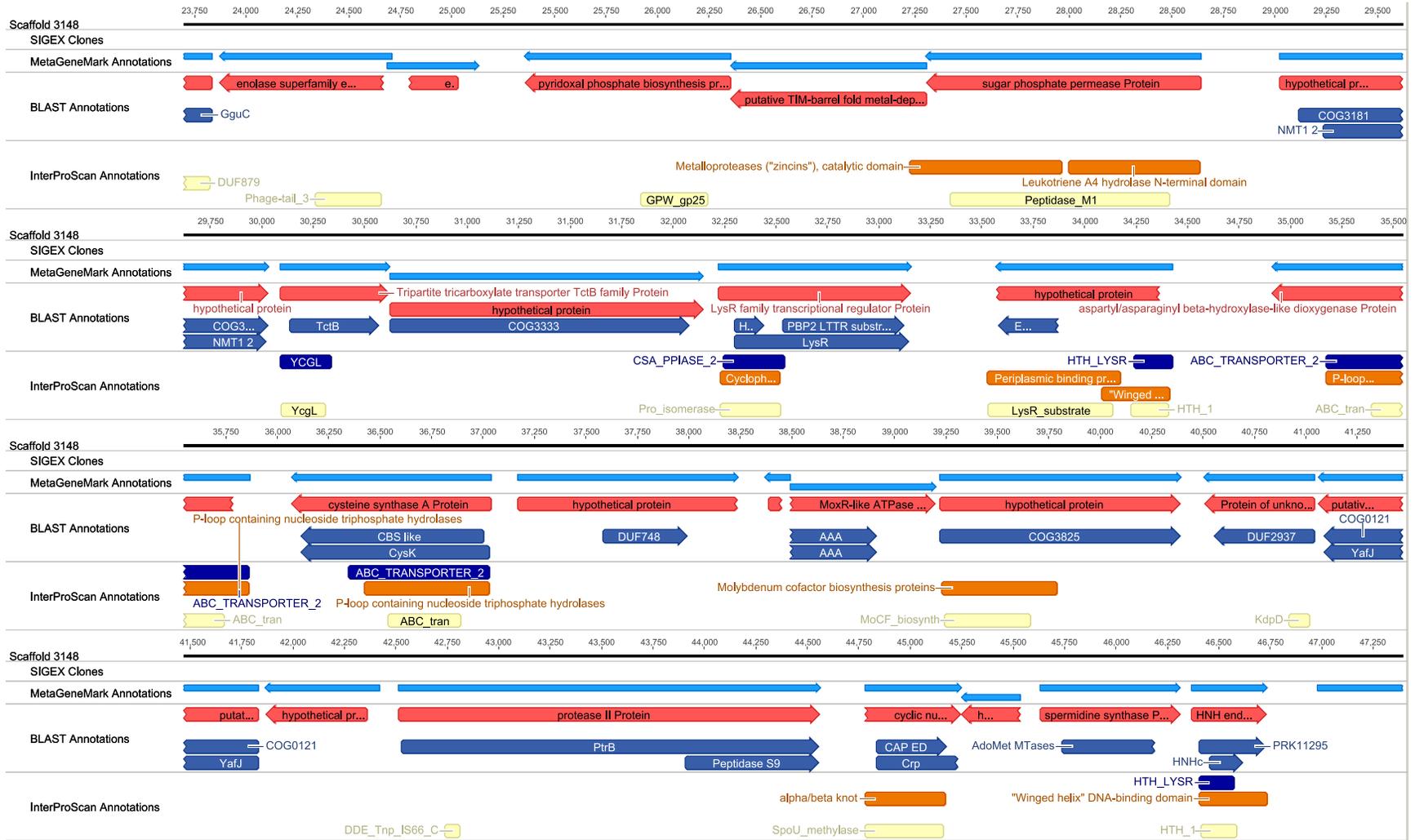
Contig 3075 aligned to SE2 (97.9%).

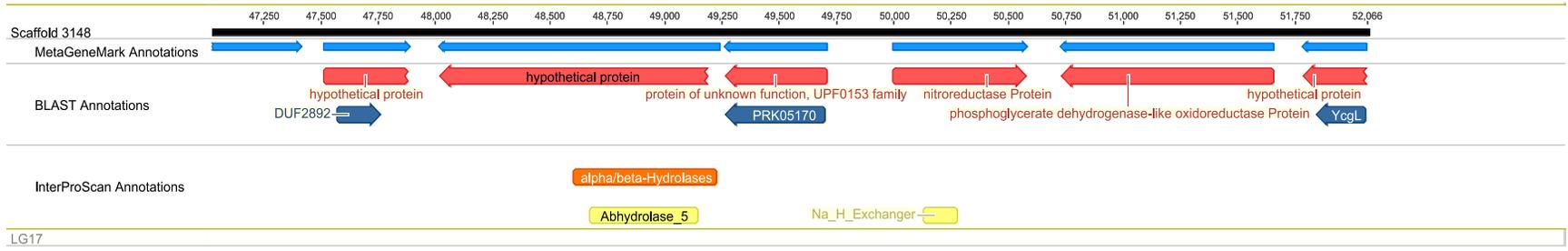


A.3. Contig 3148.

Contig 3148 aligned to LG17 (97.0%).

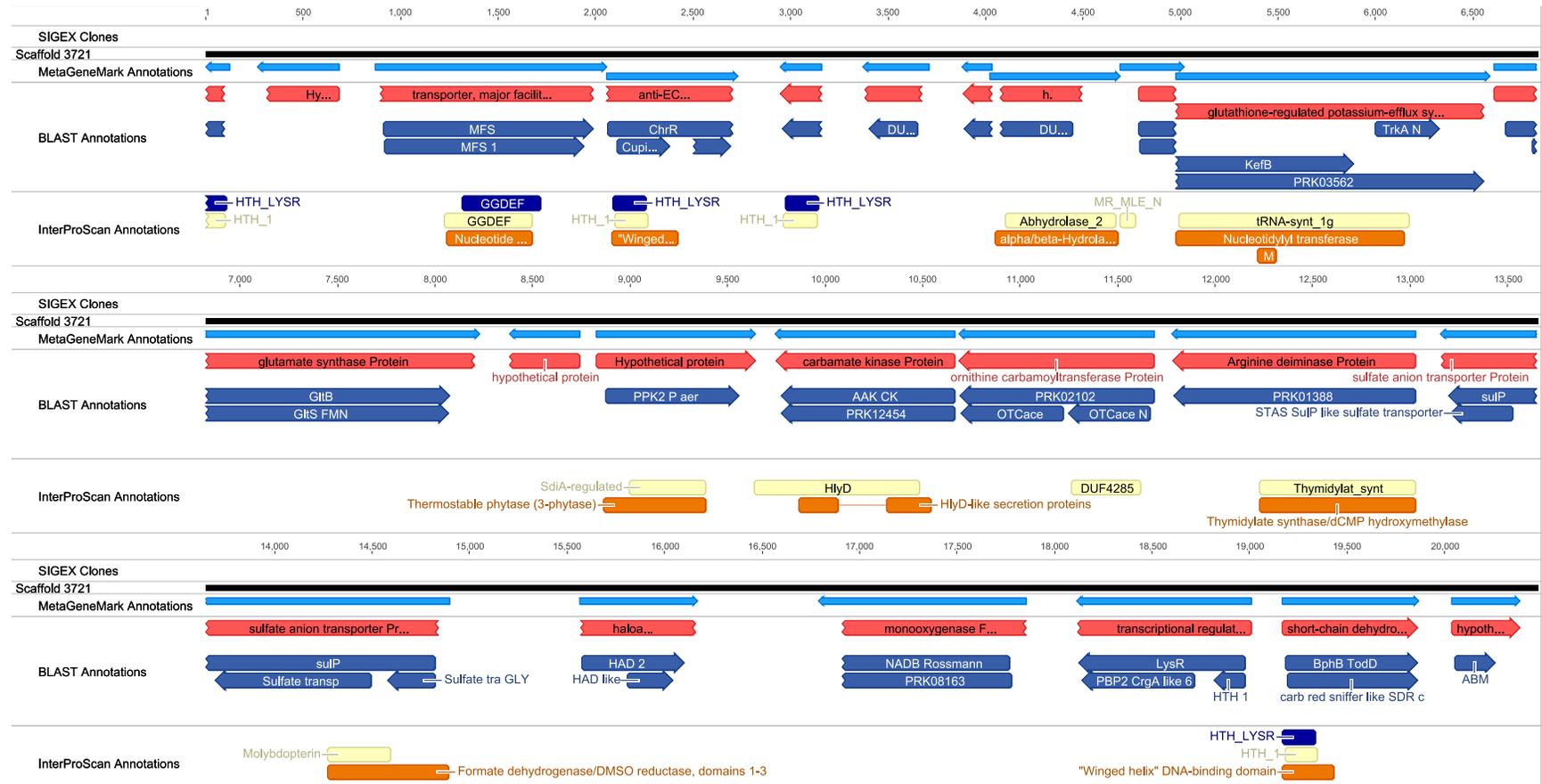


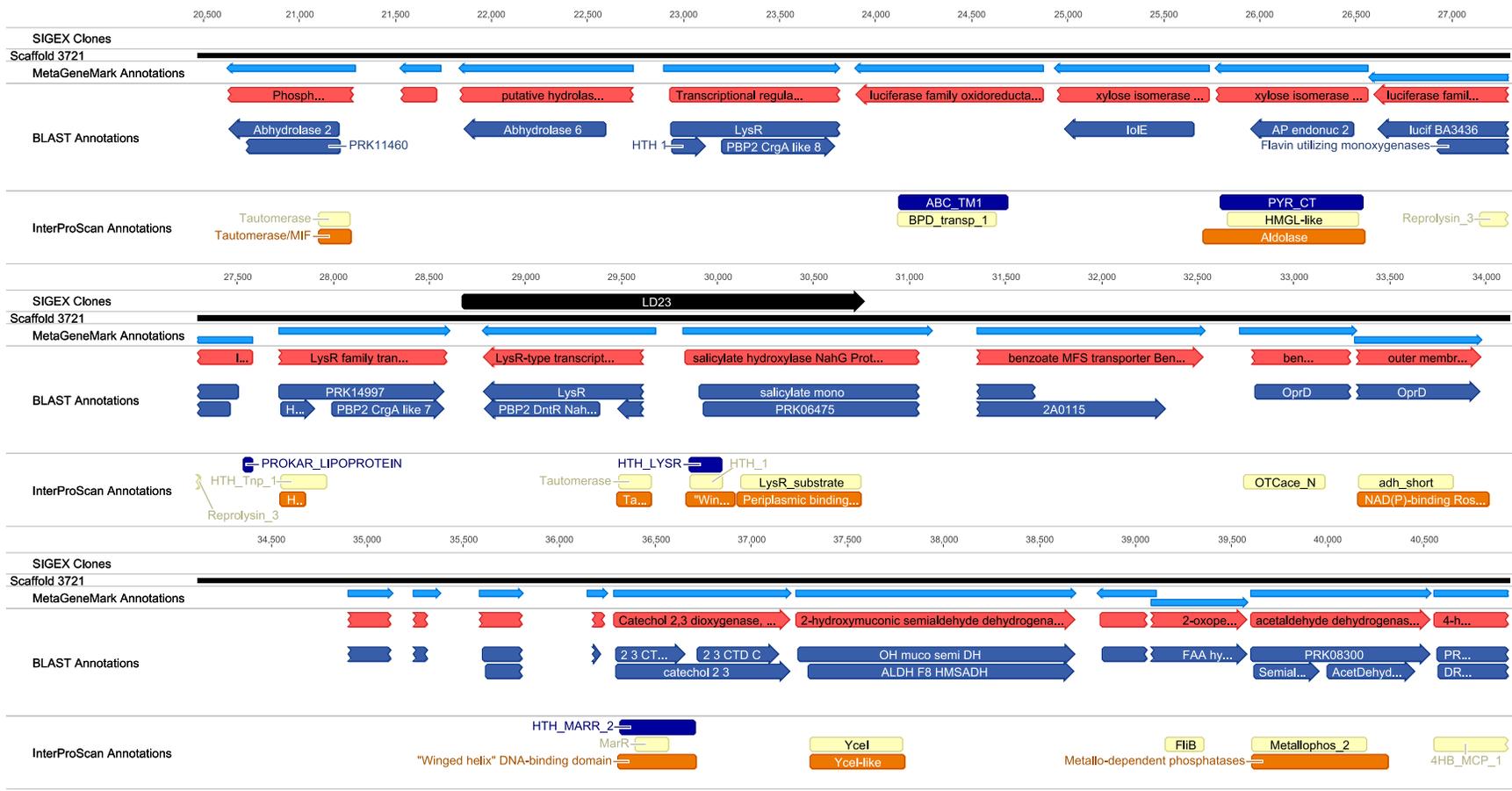




A.4. Contig 3721.

Contig 3721 aligned to LD23 (81.2%).

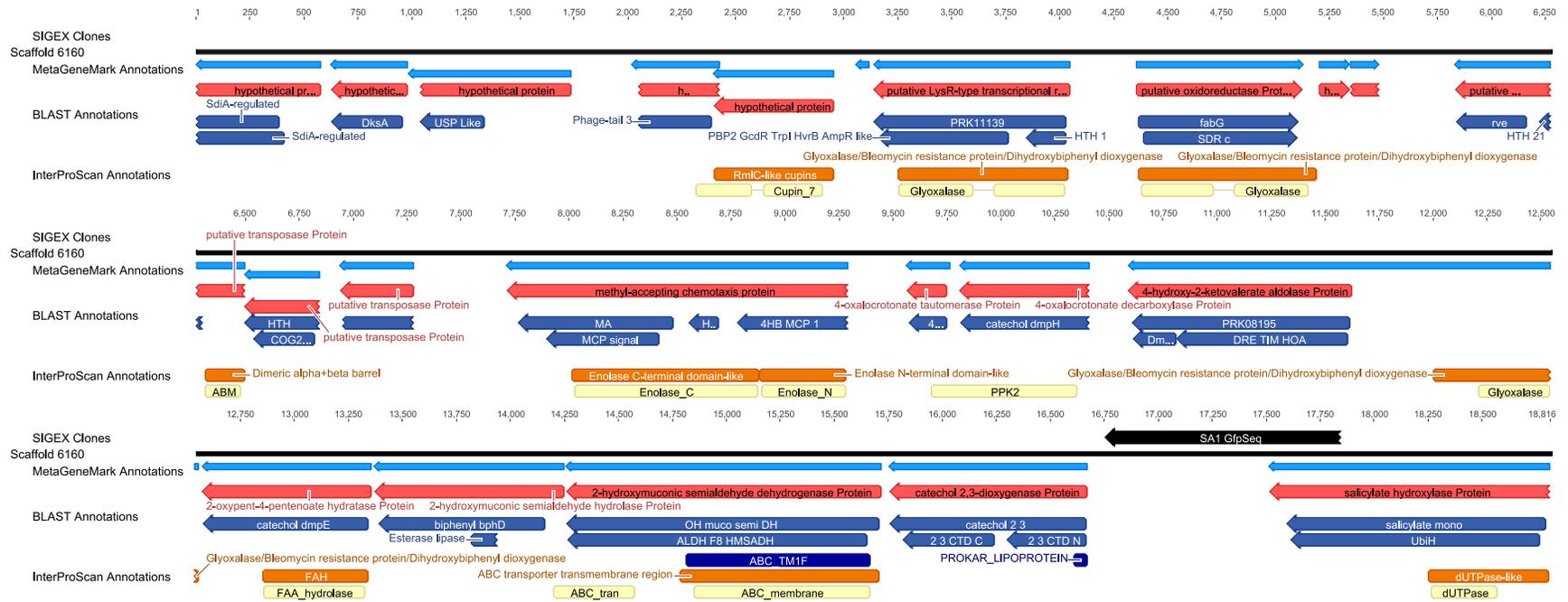




SIGEX Clones	
Scaffold 3721	
MetaGeneMark Annotations	
BLAST Annotations	
InterProScan Annotations	

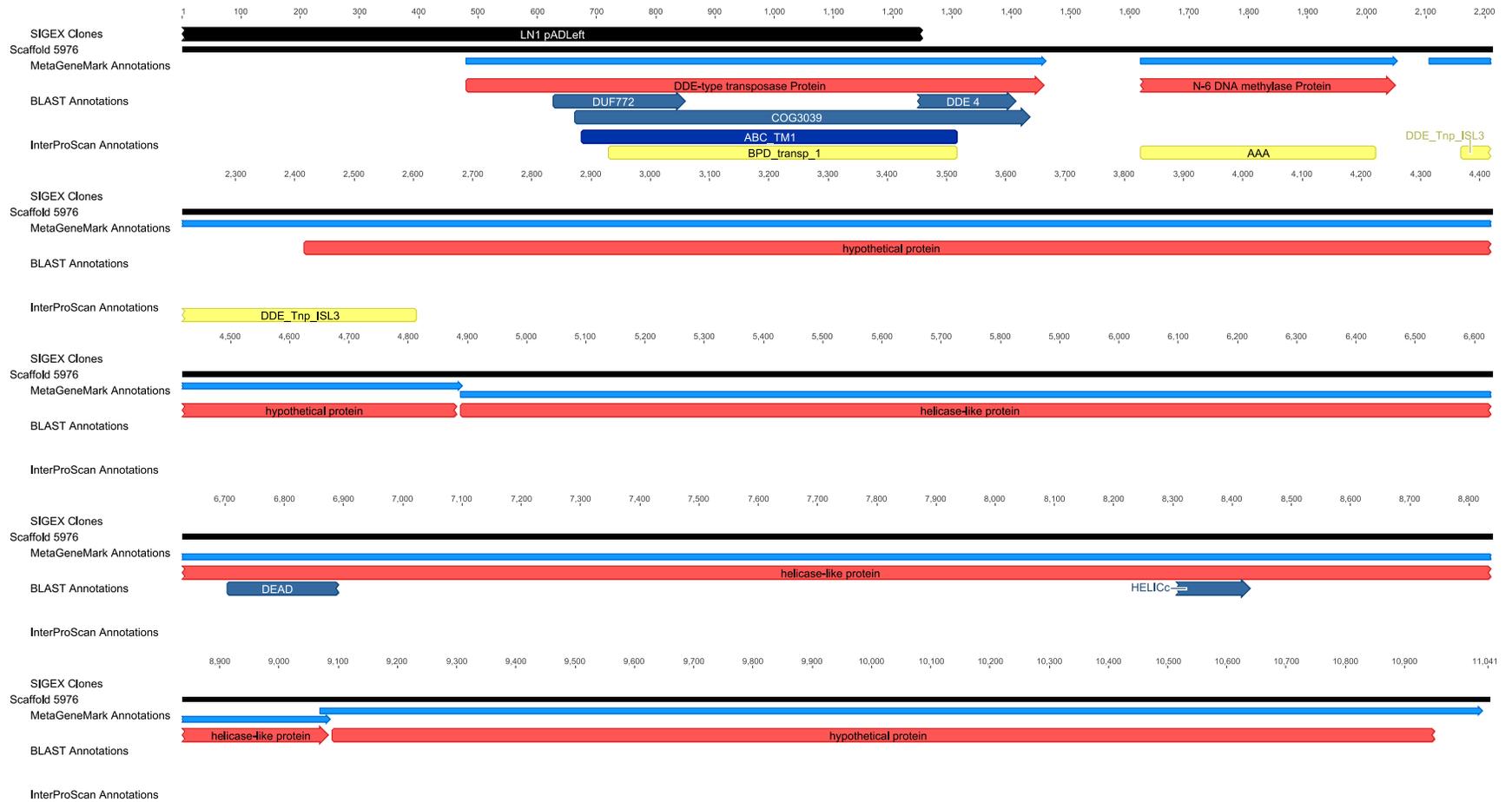
A.5. Contig 6160

Contig 6160 aligned to SA1 (95.0%).



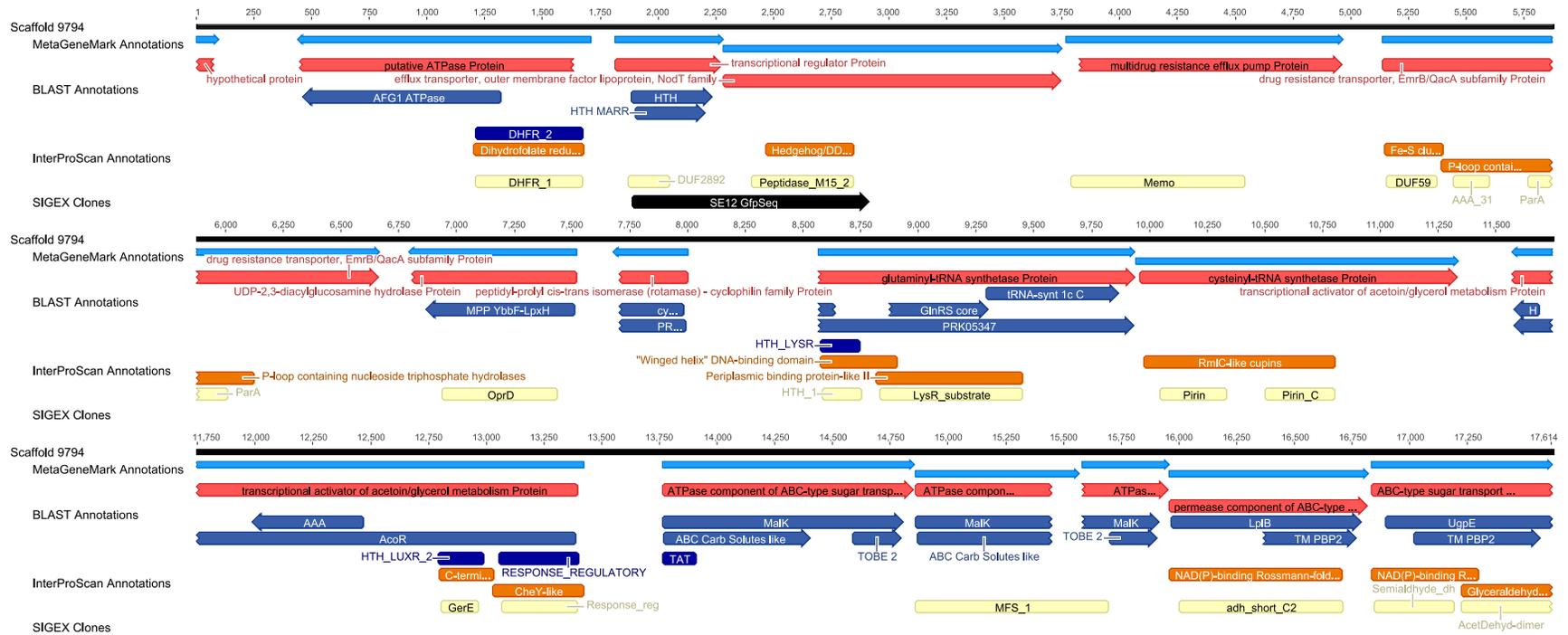
A.6. Contig 5976.

Contig 5976 aligned to LN1-pADLeft (88.5%).



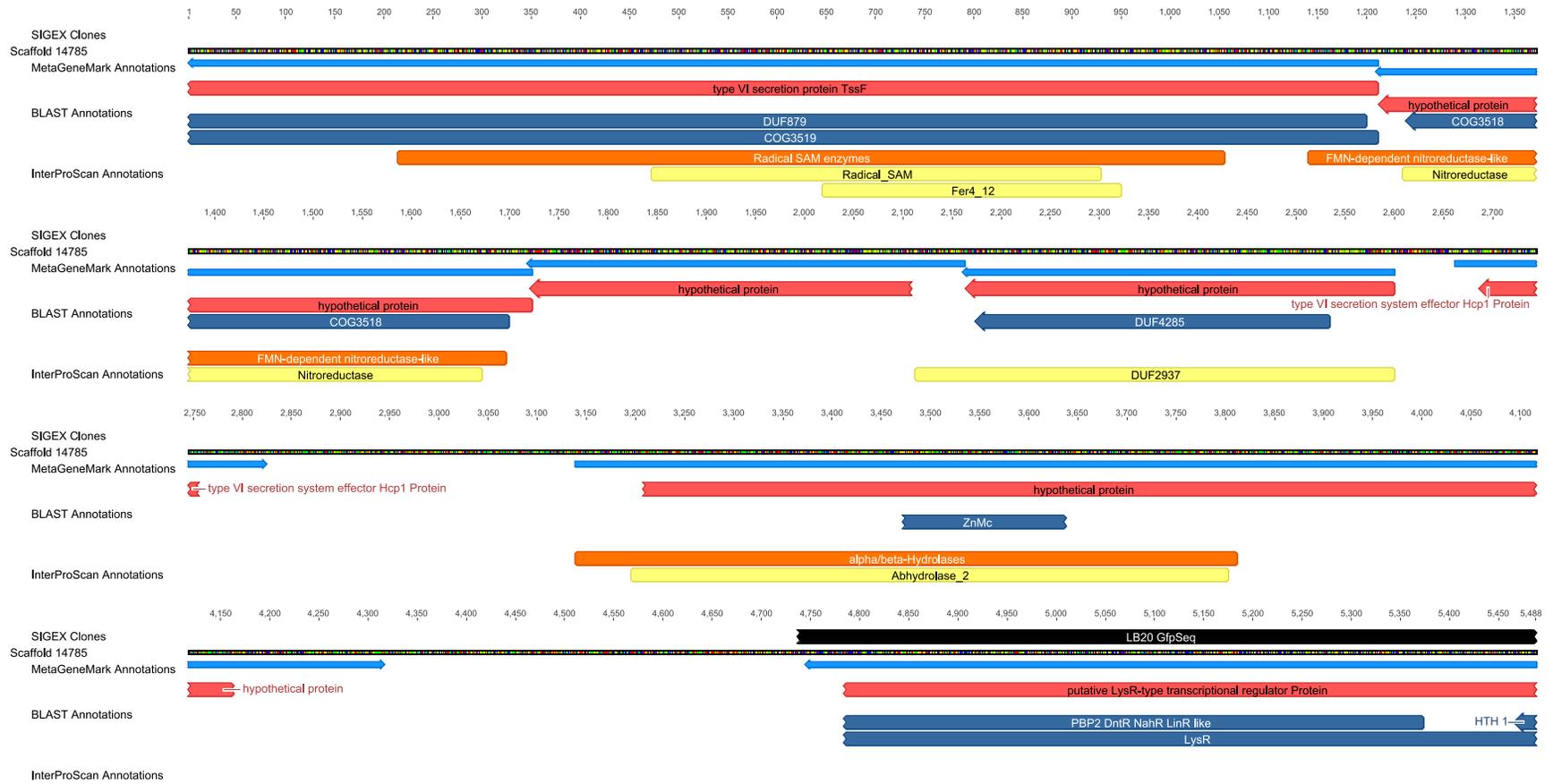
A.7. Contig 9794.

Contig 9794 aligned to SE12 (99.3%).



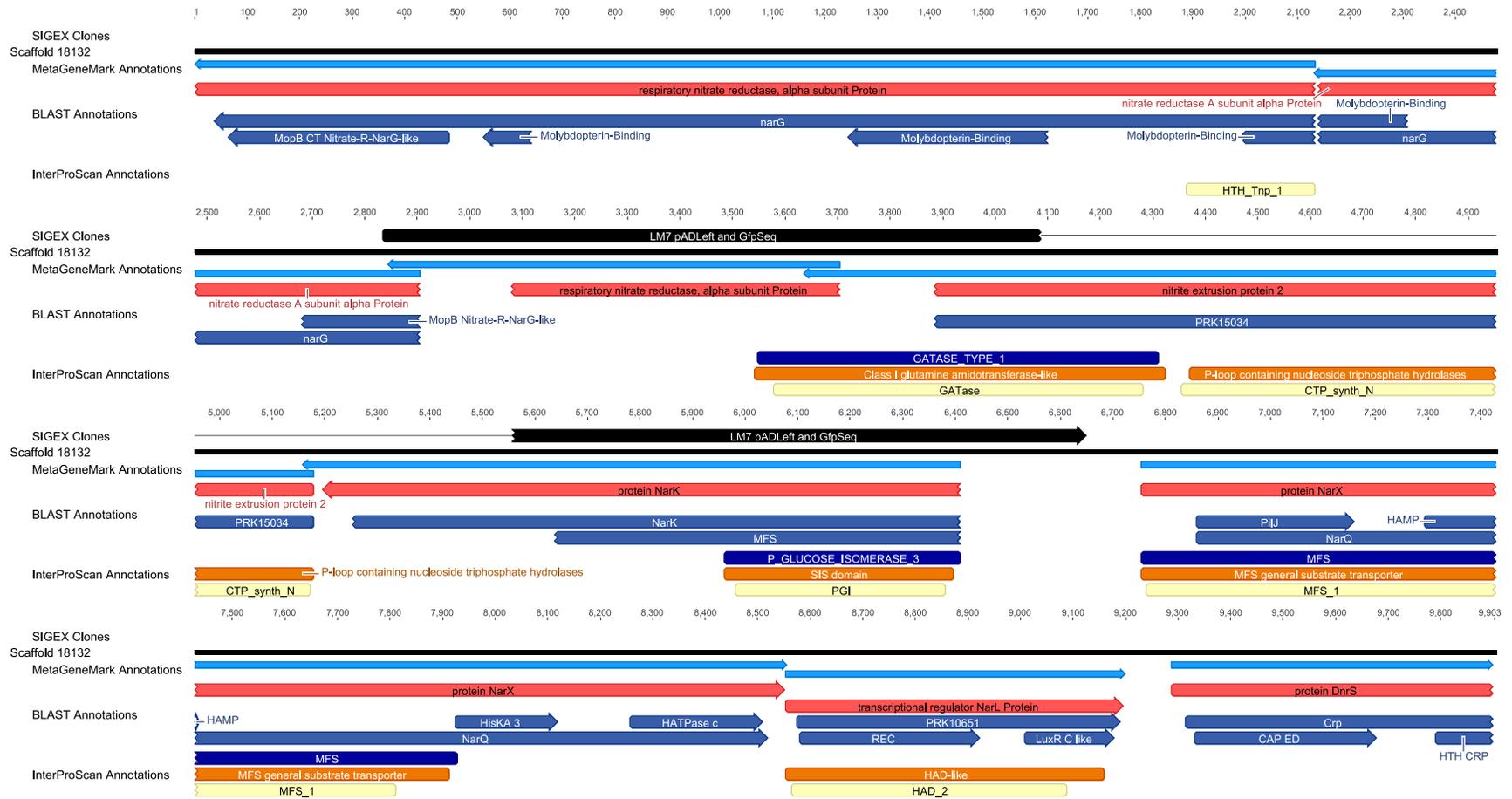
A.8. Contig 14785.

Contig 14785 aligned to LB20-GfpSeq (97.9%).



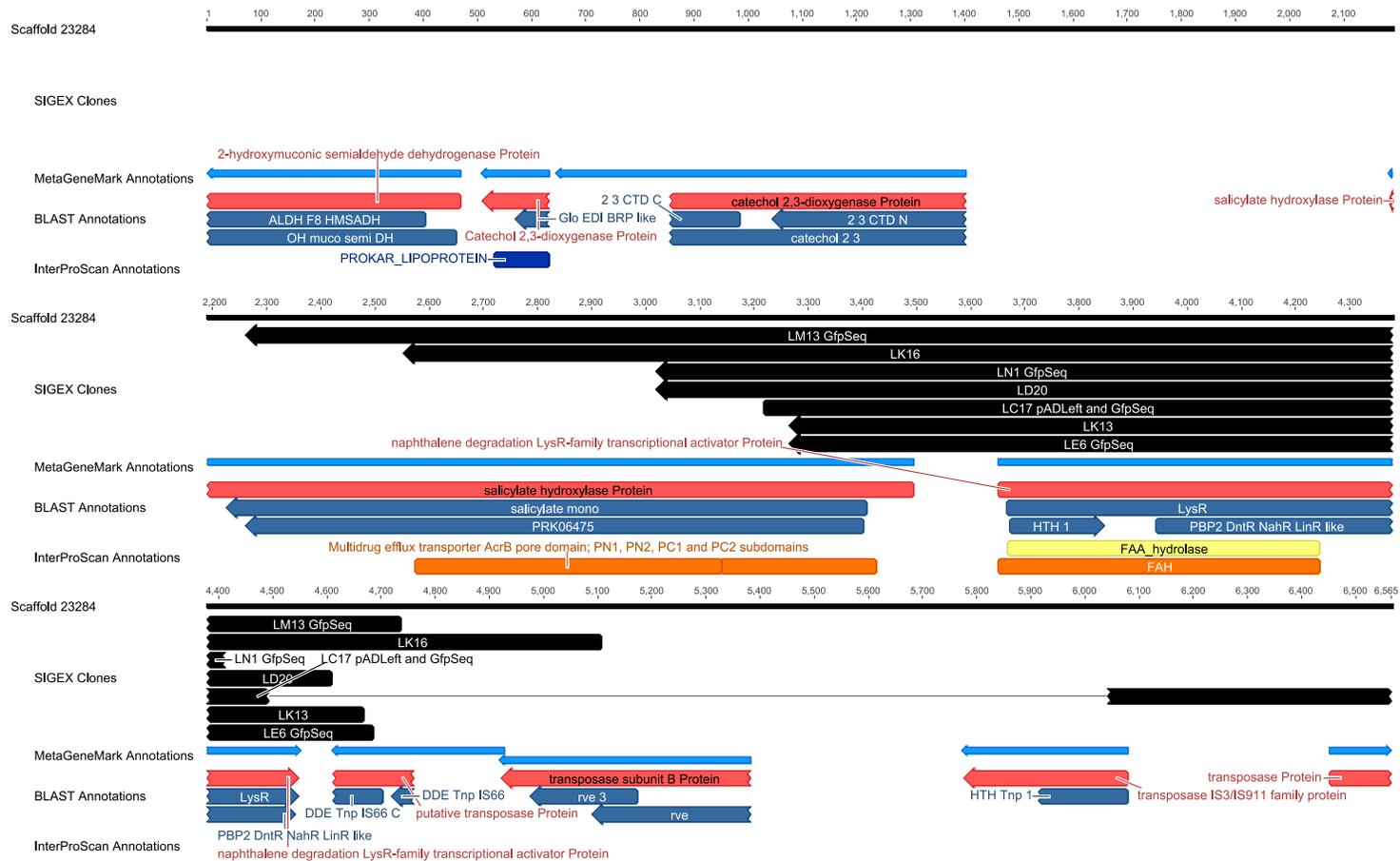
A.9. Contig 18132.

Contig 18132 aligned to LM7 (85.7% for pADLeft read, 96.7% for GfpSeq read).



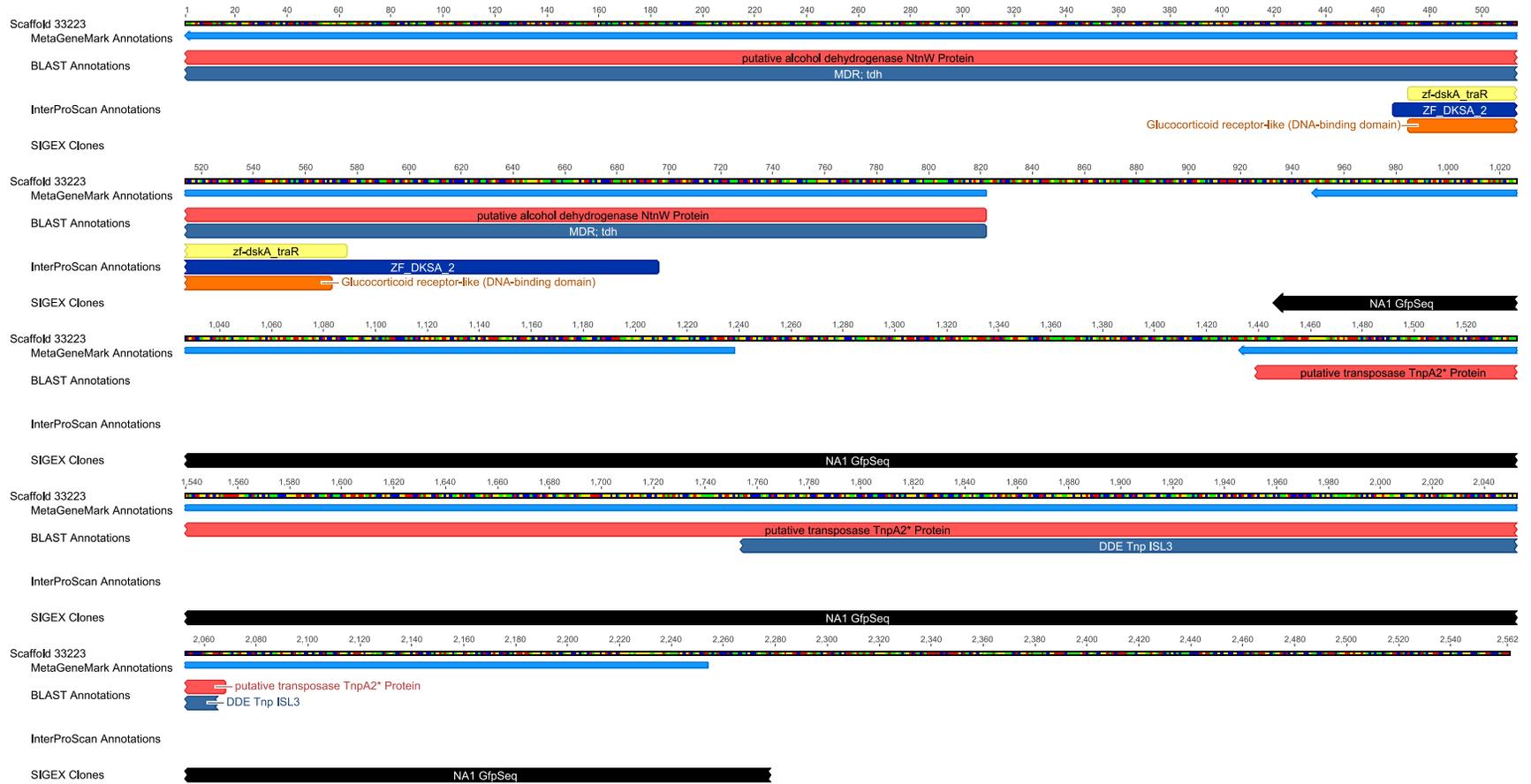
A.10. Contig 23284.

Contig 23284 aligned to LM13 (96.0%), LK16 (99.0%), LN1-GfpSeq (94.9%), LD20 (86.8%), LK13 (96.9%), LE6-GfpSeq (93.4%), LC17 (pADLeft 88.1%, GfpSeq 87.2%).



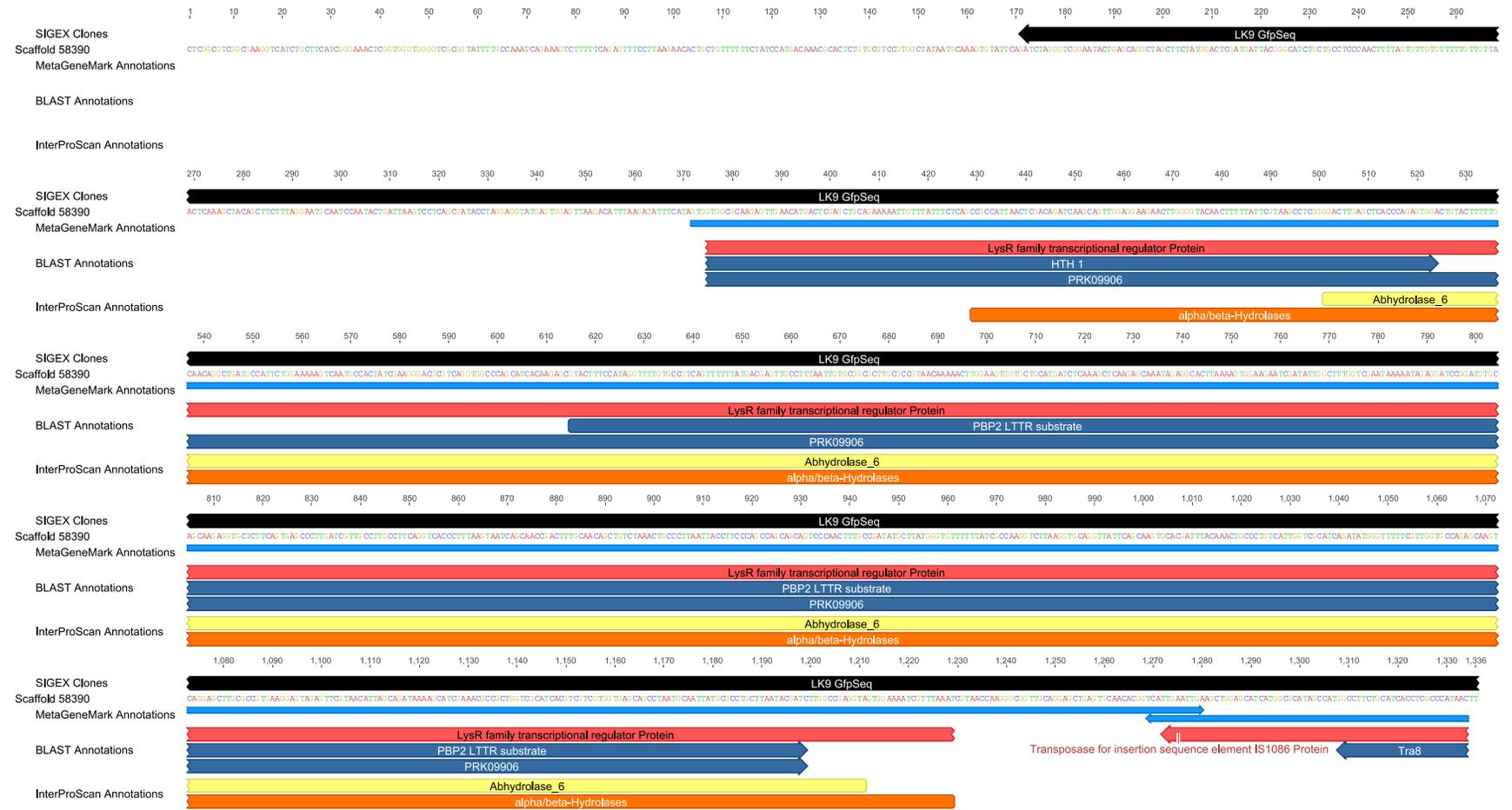
A.11. Contig 33223.

Contig 33223 aligned to NA1-GfpSeq (95.1%).



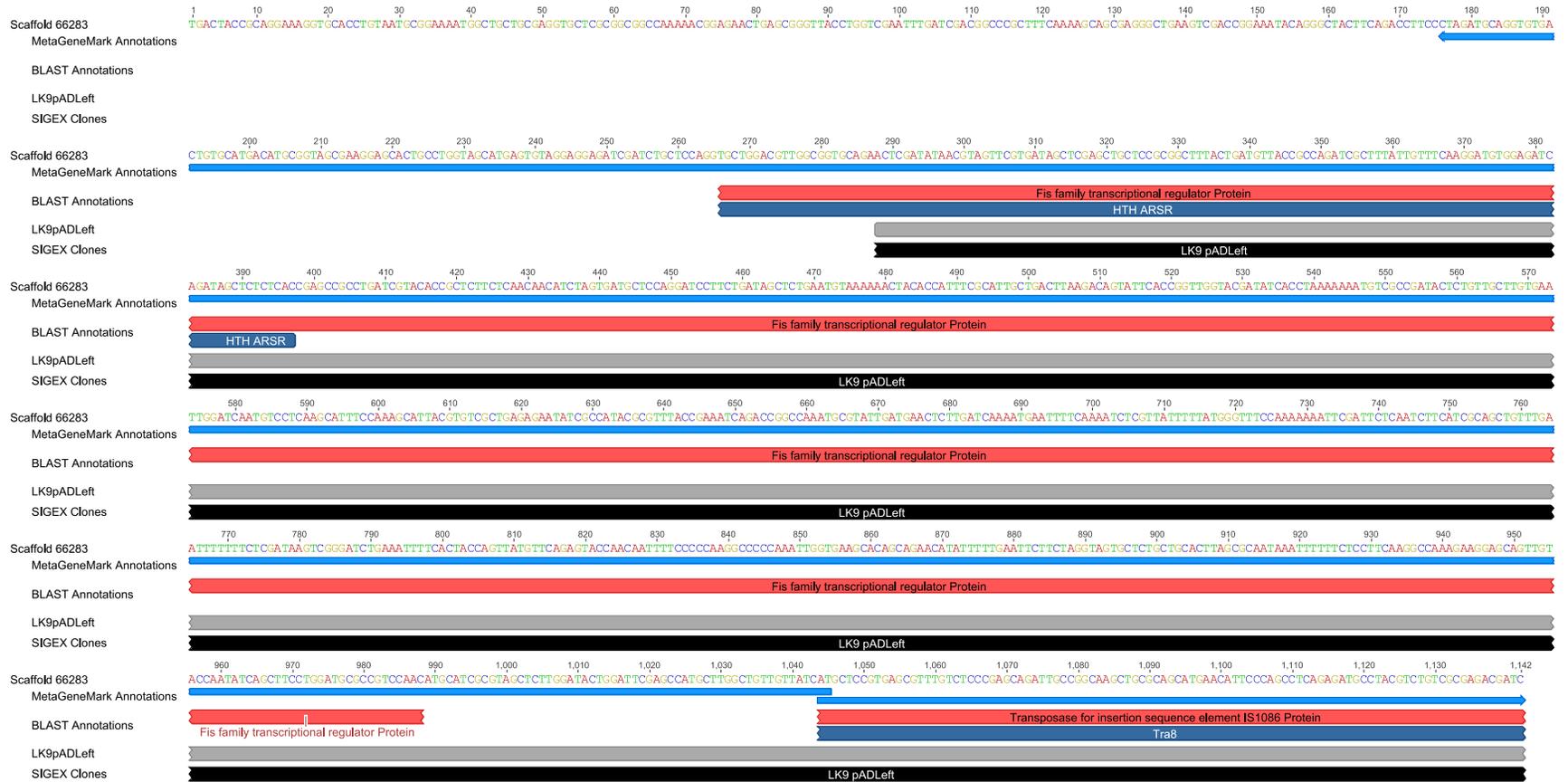
A.12. Contig 58390.

Contig 58390 aligned to LK9-GfpSeq (100%).



A.13. Contig 66283.

Contig 66283 aligned to LK9-pADLeft (99.6%).



Appendix B. Annotation Tables of Contigs with Aromatic-Inducible Genes

The following tables encompass a comprehensive record of the annotations that were applied to NGS-derived contigs which have SIGEX clones mapped to them. Annotations were derived from the DNA sequence by first predicting genes using MetaGeneMark, then using those genes as queries in BLASTp searches against the nr database, and InterProScan searches.

B.1. BLASTp search hits.

BLASTp search hit annotations for genes predicted on contigs used in analysis of SIGEX clones. Search hits are from the *nr* database.

Contig	Query Gene	Hit Accession	Hit Organism	% ID	Taxonomy (following Bacteria; Proteobacteria)
14785	RockBay_242264	ZP_10427245	<i>Pseudomonas</i> sp.	96.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
14785	RockBay_242269	YP_004467422	<i>Alteromonas</i> sp.	36.30	Gammaproteobacteria; Alteromonadales; Alteromonadaceae; <i>Alteromonas</i>
14785	RockBay_242270	BAC53588	<i>Pigmentiphaga</i> sp.	65.50	Betaproteobacteria; Burkholderiales; Alcaligenaceae; <i>Pigmentiphaga</i>
14785	RockBay_242265	YP_002875506	<i>Pseudomonas fluorescens</i>	88.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
14785	RockBay_242267	YP_007399641	<i>Pseudomonas poae</i>	74.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
14785	RockBay_242266	ZP_10594862	<i>Pseudomonas</i> sp.	63.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
14785	RockBay_242268	YP_007399640	<i>Pseudomonas poae</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263468	ZP_10670320	<i>Pseudomonas</i> sp.	99.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263473	ZP_11111383	<i>Pseudomonas mandelii</i>	90.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263471	ZP_11111385	<i>Pseudomonas mandelii</i>	88.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263472	ZP_11111384	<i>Pseudomonas mandelii</i>	88.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263469	ZP_11111386	<i>Pseudomonas mandelii</i>	98.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263474	ZP_11111382	<i>Pseudomonas mandelii</i>	93.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263470	WP_008153585	<i>Pseudomonas</i> sp.	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
18132	RockBay_263475	ZP_11111381	<i>Pseudomonas mandelii</i>	99.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
23284	RockBay_290580	ACV05012	<i>Pseudomonas aeruginosa</i>	94.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
23284	RockBay_290581	YP_006456979	<i>Pseudomonas stutzeri</i>	93.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
23284	RockBay_290579	YP_006456976	<i>Pseudomonas stutzeri</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
23284	RockBay_290577	YP_006456975	<i>Pseudomonas stutzeri</i>	89.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
23284	RockBay_290583	NP_542848	<i>Pseudomonas putida</i>	87.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
23284	RockBay_290584	WP_003450969	<i>Pseudomonas pseudoalcaligenes</i>	86.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i> ; <i>Pseudomonas oleovorans/pseudoalcaligenes</i> group

23284	RockBay_290582	WP_003349068	<i>Pseudomonas syringae</i>	92.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
23284	RockBay_290578	WP_003451984	<i>Pseudomonas pseudoalcaligenes</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas; Pseudomonas oleovorans/pseudoalcaligenes group
23284	RockBay_290585	WP_005749725	<i>Pseudomonas amygdali</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
243	RockBay_25505	YP_003673793	<i>Methylotenera versatilis</i>	75.90	Betaproteobacteria; Methylophilales; Methylophilaceae; Methylotenera
243	RockBay_25509	YP_314657	<i>Thiobacillus denitrificans</i>	71.00	Betaproteobacteria; Hydrogenophilales; Hydrogenophilaceae; Thiobacillus
243	RockBay_25502	YP_005027486	<i>Dechlorosoma suillum</i>	81.70	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Azospira
243	RockBay_25480	WP_008938741	<i>Marinobacter santoriniensis</i>	52.80	Gammaproteobacteria; Alteromonadales; Alteromonadaceae; Marinobacter
243	RockBay_25524	YP_283801	<i>Dechloromonas aromatica</i>	73.20	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Dechloromonas
243	RockBay_25486	YP_283585	<i>Dechloromonas aromatica</i>	62.50	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Dechloromonas
243	RockBay_25483	YP_003165790	<i>Candidatus Accumulibacter</i>	72.40	Betaproteobacteria; Candidatus Accumulibacter
243	RockBay_25482	YP_005028859	<i>Dechlorosoma suillum</i>	63.40	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Azospira
243	RockBay_25514	YP_001352884	<i>Janthinobacterium</i> sp.	73.30	Betaproteobacteria; Burkholderiales; Oxalobacteraceae; Janthinobacterium
243	RockBay_25520	YP_005028515	<i>Dechlorosoma suillum</i>	80.30	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Azospira
243	RockBay_25492	ZP_10382070	<i>Sulfuricella denitrificans</i>	54.90	Betaproteobacteria; Hydrogenophilales; Hydrogenophilaceae; Sulfuricella
243	RockBay_25507	YP_003673795	<i>Methylotenera versatilis</i>	51.10	Betaproteobacteria; Methylophilales; Methylophilaceae; Methylotenera
243	RockBay_25522	YP_007551347	<i>Azoarcus</i> sp.	79.70	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Azoarcus
243	RockBay_25484	ZP_03699520	<i>Pseudogulbenkiania ferrooxidans</i>	80.70	Betaproteobacteria; Neisseriales; Neisseriaceae; Pseudogulbenkiania
243	RockBay_25517	YP_001602164	<i>Gluconacetobacter diazotrophicus</i>	57.70	Alphaproteobacteria; Rhodospirillales; Acetobacteraceae; Gluconacetobacter
243	RockBay_25485	YP_283632	<i>Dechloromonas aromatica</i>	50.10	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Dechloromonas
243	RockBay_25519	YP_283803	<i>Dechloromonas aromatica</i>	65.50	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Dechloromonas
243	RockBay_25495	YP_521681	<i>Rhodoferax ferrireducens</i>	80.10	Betaproteobacteria; Burkholderiales; Comamonadaceae; Albidiferax
243	RockBay_25481	ZP_08274435	Oxalobacteraceae bacterium	57.20	Betaproteobacteria; Burkholderiales; Oxalobacteraceae
243	RockBay_25506	YP_003673794	<i>Methylotenera versatilis</i>	61.90	Betaproteobacteria; Methylophilales; Methylophilaceae; Methylotenera
243	RockBay_25478	YP_003167592	<i>Candidatus Accumulibacter</i>	66.00	Betaproteobacteria; Candidatus Accumulibacter
243	RockBay_25504	YP_934022	<i>Azoarcus</i> sp.	31.80	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Azoarcus
243	RockBay_25510	WP_004332332	<i>Thauera linaloolentis</i>	66.00	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Thauera
243	RockBay_25512	YP_522086	<i>Rhodoferax ferrireducens</i>	73.40	Betaproteobacteria; Burkholderiales; Comamonadaceae; Albidiferax
243	RockBay_25516	YP_902845	<i>Pelobacter propionicus</i>	37.90	Deltaproteobacteria; Desulfuromonadales; Pelobacteraceae; Pelobacter
243	RockBay_25521	YP_160545	<i>Aromatoleum aromaticum</i>	82.50	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Aromatoleum

243	RockBay_25498	YP_001170962	<i>Pseudomonas stutzeri</i>	77.20	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
243	RockBay_25493	ZP_10380478	<i>Sulfuricella denitrificans</i>	61.30	Betaproteobacteria; Hydrogenophilales; Hydrogenophilaceae; Sulfuricella
243	RockBay_25491	YP_003885762	<i>Cyanothece</i> sp.	35.10	Bacteria; Cyanobacteria; Oscillatoriophyceae; Chroococcales; Cyanothece
243	RockBay_25508	YP_001234988	<i>Acidiphilium cryptum</i>	46.40	Alphaproteobacteria; Rhodospirillales; Acetobacteraceae; Acidiphilium
243	RockBay_25490	WP_007510257	<i>Rhodanobacter</i> sp.	55.20	Gammaproteobacteria; Xanthomonadales; Xanthomonadaceae; Rhodanobacter
243	RockBay_25494	YP_521682	<i>Rhodoferax ferrireducens</i>	53.30	Betaproteobacteria; Burkholderiales; Comamonadaceae; Albidiferax
243	RockBay_25496	YP_521680	<i>Rhodoferax ferrireducens</i>	64.00	Betaproteobacteria; Burkholderiales; Comamonadaceae; Albidiferax
243	RockBay_25497	YP_315338	<i>Thiobacillus denitrificans</i>	70.50	Betaproteobacteria; Hydrogenophilales; Hydrogenophilaceae; Thiobacillus
243	RockBay_25489	YP_004847839	<i>Pseudogulbenkiana</i> sp.	75.40	Betaproteobacteria; Neisseriales; Neisseriaceae; Pseudogulbenkiana
243	RockBay_25503	YP_005027487	<i>Dechlorosoma suillum</i>	92.00	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Azospira
243	RockBay_25511	ZP_08772527	<i>Thiocapsa marina</i>	74.00	Gammaproteobacteria; Chromatiales; Chromatiaceae; Thiocapsa
243	RockBay_25525	WP_008481570	<i>Beggiatoa</i> sp.	37.80	Gammaproteobacteria; Thiotrichales; Thiotrichaceae; Beggiatoa
243	RockBay_25513	YP_285996	<i>Dechloromonas aromatica</i>	76.60	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Dechloromonas
243	RockBay_25515	YP_006048844	<i>Rhodospirillum rubrum</i>	57.60	Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; Rhodospirillum
243	RockBay_25477	YP_003522750	<i>Sideroxydans lithotrophicus</i>	93.20	Betaproteobacteria; Gallionellales; Gallionellaceae; Sideroxydans
243	RockBay_25499	YP_283808	<i>Dechloromonas aromatica</i>	56.60	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Dechloromonas
243	RockBay_25501	WP_006221722	<i>Achromobacter piechaudii</i>	61.60	Betaproteobacteria; Burkholderiales; Alcaligenaceae; Achromobacter
243	RockBay_25487	YP_005026571	<i>Dechlorosoma suillum</i>	84.70	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Azospira
243	RockBay_25500	ACN22627	uncultured bacterium	78.50	Bacteria; environmental samples
243	RockBay_25488	ZP_08504949	<i>Methyloversatilis universalis</i>	48.60	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Methyloversatilis
243	RockBay_25523	YP_160547	<i>Aromatoleum aromaticum</i>	72.30	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Aromatoleum
243	RockBay_25518	YP_286710	<i>Dechloromonas aromatica</i>	58.20	Betaproteobacteria; Rhodocyclales; Rhodocyclaceae; Dechloromonas
3075	RockBay_115639	ACO92382	<i>Pseudomonas putida</i>	84.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115651	ZP_10148962	<i>Pseudomonas</i> sp.	65.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115648	YP_003452219	<i>Azospirillum</i> sp.	59.00	Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; Azospirillum
3075	RockBay_115649	YP_003452218	<i>Azospirillum</i> sp.	55.70	Alphaproteobacteria; Rhodospirillales; Rhodospirillaceae; Azospirillum
3075	RockBay_115643	ACO92377	<i>Pseudomonas putida</i>	99.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115645	ACO92375	<i>Pseudomonas putida</i>	92.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115638	ACO92381	<i>Pseudomonas putida</i>	91.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115646	ACO92374	<i>Pseudomonas putida</i>	99.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas

3075	RockBay_115650	YP_003607681	Burkholderia sp.	46.50	Betaproteobacteria; Burkholderiales; Burkholderiaceae; Burkholderia
3075	RockBay_115647	ACO92380	Pseudomonas putida	99.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115644	ACO92376	Pseudomonas putida	93.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115640	ACO92383	Pseudomonas putida	99.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3075	RockBay_115642	ACO92378	Pseudomonas putida	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117048	ZP_10635276	Pseudomonas sp.	93.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117074	ZP_10636063	Pseudomonas sp.	93.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117043	ZP_10641574	Pseudomonas sp.	95.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117050	ZP_11112323	Pseudomonas mandelii	97.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117051	WP_008043471	Pseudomonas sp.	71.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117053	ZP_11112325	Pseudomonas mandelii	90.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117062	ZP_10638996	Pseudomonas sp.	87.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117059	WP_008045228	Pseudomonas sp.	95.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117046	ZP_10635274	Pseudomonas sp.	92.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117070	ZP_10699138	Pseudomonas sp.	95.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117081	ZP_10641821	Pseudomonas sp.	96.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117037	YP_004352605	Pseudomonas brassicacearum	90.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117045	ZP_10635273	Pseudomonas sp.	93.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117047	ZP_10638982	Pseudomonas sp.	98.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117067	ZP_11112338	Pseudomonas mandelii	88.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117060	ZP_10636304	Pseudomonas sp.	93.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117057	WP_007946808	Pseudomonas sp.	95.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117054	ZP_10635281	Pseudomonas sp.	90.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117066	ZP_10641869	Pseudomonas sp.	92.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117058	WP_008029057	Pseudomonas sp.	94.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117063	ZP_11112334	Pseudomonas mandelii	89.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117084	WP_008034968	Pseudomonas sp.	94.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117040	ZP_10633660	Pseudomonas sp.	93.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
3148	RockBay_117052	WP_008043472	Pseudomonas sp.	79.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas

3148	RockBay_117065	WP_008035075	<i>Pseudomonas</i> sp.	91.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117041	WP_008034293	<i>Pseudomonas</i> sp.	97.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117055	WP_007943924	<i>Pseudomonas</i> sp.	99.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117064	ZP_11112335	<i>Pseudomonas mandelii</i>	76.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117072	ZP_10636065	<i>Pseudomonas</i> sp.	99.20	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117069	ZP_10654702	<i>Pseudomonas</i> sp.	97.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117044	ZP_10638979	<i>Pseudomonas</i> sp.	83.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117077	ZP_11112352	<i>Pseudomonas mandelii</i>	93.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117038	ZP_11112310	<i>Pseudomonas mandelii</i>	99.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117083	WP_007984791	<i>Pseudomonas</i> sp.	99.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117042	ZP_10598296	<i>Pseudomonas</i> sp.	90.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117071	ZP_10673866	<i>Pseudomonas</i> sp.	98.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117061	ZP_10638995	<i>Pseudomonas</i> sp.	95.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117073	ZP_10636064	<i>Pseudomonas</i> sp.	73.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117075	ZP_10670039	<i>Pseudomonas</i> sp.	96.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117082	WP_003184256	<i>Pseudomonas fluorescens</i>	97.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117080	ZP_11112355	<i>Pseudomonas mandelii</i>	84.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117078	ZP_10596712	<i>Pseudomonas</i> sp.	99.20	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117039	ZP_10633659	<i>Pseudomonas</i> sp.	87.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117076	ZP_10653544	<i>Pseudomonas</i> sp.	63.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117085	ZP_10596705	<i>Pseudomonas</i> sp.	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117056	WP_008029049	<i>Pseudomonas</i> sp.	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117049	ZP_11112322	<i>Pseudomonas mandelii</i>	85.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3148	RockBay_117068	YP_007396413	<i>Pseudomonas poae</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
33223	RockBay_330226	AAC38358	<i>Pseudomonas</i> sp.	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
33223	RockBay_330228	AAF23984	<i>Pseudomonas</i> sp.	92.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127619	YP_005884801	<i>Marinobacter adhaerens</i>	66.70	Gammaproteobacteria; Alteromonadales; Alteromonadaceae; <i>Marinobacter</i>
3721	RockBay_127588	WP_003082491	<i>Pseudomonas aeruginosa</i>	47.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>

3721	RockBay_127580	WP_003406071	<i>Pseudomonas syringae</i>	82.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127617	ZP_10994472	<i>Pseudomonas fuscovaginae</i>	92.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127616	YP_005938337	<i>Pseudomonas stutzeri</i>	45.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127586	ZP_19203401	<i>Pseudomonas</i> sp.	84.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127603	WP_007970831	<i>Pseudomonas</i> sp.	84.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127618	NP_863106	<i>Pseudomonas putida</i>	76.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127584	YP_002892604	<i>Tolomonas auensis</i>	80.30	Gammaproteobacteria; Aeromonadales; Aeromonadaceae; <i>Tolomonas</i>
3721	RockBay_127622	ZP_19212538	<i>Pseudomonas putida</i>	82.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127581	AGH85959	<i>Ralstonia solanacearum</i>	57.10	Betaproteobacteria; Burkholderiales; Burkholderiaceae; <i>Ralstonia</i>
3721	RockBay_127591	YP_004379227	<i>Pseudomonas mendocina</i>	60.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127601	ZP_10557216	<i>Pantoea</i> sp.	61.10	Gammaproteobacteria; Enterobacteriales; Enterobacteriaceae; <i>Pantoea</i>
3721	RockBay_127585	WP_006894058	<i>Methylobacter tundripaludum</i>	73.90	Gammaproteobacteria; Methylococcales; Methylococcaceae; <i>Methylobacter</i>
3721	RockBay_127587	YP_004475573	<i>Pseudomonas fulva</i>	76.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127597	YP_007242137	<i>Pseudomonas stutzeri</i>	57.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127600	YP_002362926	<i>Methylocella silvestris</i>	49.10	Alphaproteobacteria; Rhizobiales; Beijerinckiaceae; <i>Methylocella</i>
3721	RockBay_127583	YP_003167582	<i>Candidatus Accumulibacter</i>	57.10	Betaproteobacteria; <i>Candidatus Accumulibacter</i>
3721	RockBay_127614	YP_004713956	<i>Pseudomonas stutzeri</i>	73.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127602	WP_003411083	<i>Pseudomonas syringae</i>	59.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127615	ZP_10708092	<i>Pseudomonas</i> sp.	68.20	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127623	YP_709325	<i>Pseudomonas putida</i>	95.20	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127599	ZP_10673017	<i>Pseudomonas</i> sp.	66.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127613	WP_003450504	<i>Pseudomonas pseudoalcaligenes</i>	62.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i> ; <i>Pseudomonas</i> oleovorans/pseudoalcaligenes group
3721	RockBay_127582	ZP_18875393	<i>Pseudomonas chlororaphis</i>	64.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127605	AGH87151	<i>Ralstonia solanacearum</i>	73.00	Betaproteobacteria; Burkholderiales; Burkholderiaceae; <i>Ralstonia</i>
3721	RockBay_127608	YP_006536112	<i>Pseudomonas putida</i>	80.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127607	YP_006536111	<i>Pseudomonas putida</i>	80.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127604	YP_006389795	<i>Pseudomonas putida</i>	69.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127592	WP_003450823	<i>Pseudomonas pseudoalcaligenes</i>	81.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i> ; <i>Pseudomonas</i> oleovorans/pseudoalcaligenes group
3721	RockBay_127593	YP_001186605	<i>Pseudomonas mendocina</i>	78.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>

3721	RockBay_127609	YP_006536113	<i>Pseudomonas putida</i>	76.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127610	YP_006389807	<i>Pseudomonas putida</i>	79.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127606	WP_008013356	<i>Pseudomonas</i> sp.	71.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127611	AAM18547	uncultured bacterium	88.60	Bacteria; environmental samples
3721	RockBay_127589	WP_003109220	<i>Pseudomonas aeruginosa</i>	56.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127620	YP_002798088	<i>Azotobacter vinelandii</i>	82.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Azotobacter</i>
3721	RockBay_127625	WP_003448357	<i>Pseudomonas pseudoalcaligenes</i>	93.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i> ; <i>Pseudomonas oleovorans/pseudoalcaligenes</i> group
3721	RockBay_127624	ZP_10704174	<i>Pseudomonas</i> sp.	90.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127598	ZP_19212574	<i>Pseudomonas putida</i>	92.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127594	YP_001186606	<i>Pseudomonas mendocina</i>	93.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127612	YP_534831	<i>Pseudomonas putida</i>	83.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127595	WP_003451732	<i>Pseudomonas pseudoalcaligenes</i>	80.20	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i> ; <i>Pseudomonas oleovorans/pseudoalcaligenes</i> group
3721	RockBay_127596	YP_294916	<i>Ralstonia eutropha</i>	69.60	Betaproteobacteria; Burkholderiales; Burkholderiaceae; <i>Cupriavidus</i>
3721	RockBay_127590	ZP_11258118	<i>Pseudomonas</i> sp.	78.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
3721	RockBay_127621	YP_002798089	<i>Azotobacter vinelandii</i>	90.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Azotobacter</i>
58390	RockBay_397300	YP_005090855	<i>Oceanimonas</i> sp.	72.30	Gammaproteobacteria; Aeromonadales; Aeromonadaceae; <i>Oceanimonas</i>
58390	RockBay_397301	WP_003461919	<i>Pseudomonas pseudoalcaligenes</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i> ; <i>Pseudomonas oleovorans/pseudoalcaligenes</i> group
5976	RockBay_161267	YP_622756	<i>Burkholderia cenocepacia</i>	76.70	Betaproteobacteria; Burkholderiales; Burkholderiaceae; <i>Burkholderia</i> ; <i>Burkholderia cepacia</i> complex
5976	RockBay_161266	YP_622757	<i>Burkholderia cenocepacia</i>	68.20	Betaproteobacteria; Burkholderiales; Burkholderiaceae; <i>Burkholderia</i> ; <i>Burkholderia cepacia</i> complex
5976	RockBay_161268	YP_622755	<i>Burkholderia cenocepacia</i>	70.70	Betaproteobacteria; Burkholderiales; Burkholderiaceae; <i>Burkholderia</i> ; <i>Burkholderia cepacia</i> complex
5976	RockBay_161264	YP_004715165	<i>Pseudomonas stutzeri</i>	99.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
5976	RockBay_161265	YP_622758	<i>Burkholderia cenocepacia</i>	79.90	Betaproteobacteria; Burkholderiales; Burkholderiaceae; <i>Burkholderia</i> ; <i>Burkholderia cepacia</i> complex
6160	RockBay_163609	NP_863095	<i>Pseudomonas putida</i>	86.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163615	NP_863102	<i>Pseudomonas putida</i>	91.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163617	NP_863106	<i>Pseudomonas putida</i>	95.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163612	NP_863098	<i>Pseudomonas putida</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163616	NP_863103	<i>Pseudomonas putida</i>	100.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>

6160	RockBay_163602	NP_863090	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163614	NP_863101	<i>Pseudomonas putida</i>	96.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163613	NP_863100	<i>Pseudomonas putida</i>	94.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163603	NP_863091	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163598	NP_863088	<i>Pseudomonas putida</i>	93.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163606	NP_863093	<i>Pseudomonas putida</i>	94.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163611	NP_863097	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163596	NP_863086	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163600	NP_863089	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163599	YP_001354474	<i>Janthinobacterium</i> sp.	93.80	Betaproteobacteria; Burkholderiales; Oxalobacteraceae; <i>Janthinobacterium</i>
6160	RockBay_163597	NP_863087	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163607	NP_863094	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163608	NP_943111	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163610	NP_863096	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163604	NP_863092	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
6160	RockBay_163605	NP_943108	<i>Pseudomonas putida</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
66283	RockBay_413155	YP_004931645	<i>Pseudoxanthomonas spadix</i>	29.20	Gammaproteobacteria; Xanthomonadales; Xanthomonadaceae; <i>Pseudoxanthomonas</i>
66283	RockBay_413156	WP_003465026	<i>Pseudomonas pseudoalcaligenes</i>	100.0 0	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i> ; <i>Pseudomonas</i> <i>oleovorans/pseudoalcaligenes</i> group
9794	RockBay_202206	ZP_10622512	<i>Pseudomonas</i> sp.	97.40	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
9794	RockBay_202201	WP_008064201	<i>Pseudomonas</i> sp.	93.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
9794	RockBay_202199	WP_008064198	<i>Pseudomonas</i> sp.	81.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
9794	RockBay_202205	WP_008020685	<i>Pseudomonas</i> sp.	96.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
9794	RockBay_202204	ZP_10622510	<i>Pseudomonas</i> sp.	98.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>
9794	RockBay_202197	ZP_10667796	<i>Pseudomonas</i> sp.	78.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; <i>Pseudomonas</i>

9794	RockBay_202200	WP_008064199	Pseudomonas sp.	89.90	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202207	ZP_10622513	Pseudomonas sp.	95.80	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202210	ZP_10622515	Pseudomonas sp.	96.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202211	ZP_10622516	Pseudomonas sp.	94.10	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202202	ZP_10622508	Pseudomonas sp.	92.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202208	ZP_10622514	Pseudomonas sp.	99.50	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202198	ZP_10623566	Pseudomonas sp.	85.70	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202209	ZP_10622514	Pseudomonas sp.	97.60	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202203	ZP_10636222	Pseudomonas sp.	99.00	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas
9794	RockBay_202196	ZP_10646039	Pseudomonas sp.	92.30	Gammaproteobacteria; Pseudomonadales; Pseudomonadaceae; Pseudomonas

B.2. BLAST region annotations

BLAST region annotations for genes predicted on contigs used in analysis of SIGEX clones. Search hits are from the NCBI *nr* database and the conserved domain database (CDD).

Contig	Region Name	Hit			CDD ID (% Sim)	Region Description (Hit Accession)
		Start	End	Length		
14785	COG3519	1	1212	1212	CDD:33322 (96.15)	Type VI protein secretion system component VasA [Intracellular trafficking, secretion, and vesicular transport] (ZP_10427245)
14785	DUF879	1	1200	1200	CDD:203356 (96.11)	Bacterial protein of unknown function (DUF879); pfam05947 (ZP_10427245)
14785	LysR	4784	5488	705	CDD:30928 (65.9)	Transcriptional regulator [Transcription]; COG0583 (BAC53588)
14785	PBP2 DntR NahR LinR like	4784	5374	591	CDD:176148 (67.04)	The C-terminal substrate binding domain of LysR-type transcriptional regulators that are involved in the catabolism of dinitrotoluene, naphthalene and gamma-hexachlorohexane; contains the type 2 periplasmic binding fold; cd08459 (BAC53588)
14785	COG3518	1239	1700	462	CDD:226049 (96.01)	Predicted component of the type VI protein secretion system [Intracellular trafficking, secretion, and vesicular transport] (YP_002875506)
14785	DUF4285	2173	2535	363	CDD:222551 (73.98)	Domain of unknown function (DUF4285); pfam14113 (YP_007399641)
14785	ZnMc	3472	3639	168	CDD:213077 (22.59)	Zinc-dependent metalloprotease. This super-family of metalloproteases contains two major branches, the astacin-like proteases and the adamalysin/reprolysin-like proteases. Both branches have wide phylogenetic distribution, and contain sub-families, which...; cl00064 (YP_004467422)
14785	HTH 1	5465	5488	24	CDD:201021 (82.22)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (BAC53588)
18132	narG	37	2133	2097	CDD:162434 (99.13)	respiratory nitrate reductase, alpha subunit; TIGR01580 (ZP_10670320)
18132	NarQ	6859	8520	1662	CDD:33640 (99.21)	Signal transduction histidine kinase, nitrate/nitrite-specific [Signal transduction mechanisms]; COG3850 (ZP_11111383)
18132	PRK15034	3884	5179	1296	CDD:184994 (98.85)	nitrate/nitrite transport protein NarU; Provisional (ZP_11111385)
18132	NarK	5253	6410	1158	CDD:32405 (98.83)	Nitrate/nitrite transporter [Inorganic ion transport and metabolism]; COG2223 (ZP_11111384)
18132	MFS	5637	6410	774	CDD:119392 (100)	The Major Facilitator Superfamily (MFS) is a large and diverse group of secondary transporters that includes uniporters, symporters, and antiporters. MFS proteins facilitate the transport across cytoplasmic or internal membranes of a variety of...; cd06174 (ZP_11111384)
18132	narG	2139	2906	768	CDD:162434 (98.8)	respiratory nitrate reductase, alpha subunit; TIGR01580 (ZP_11111386)
18132	PRK10651	8574	9191	618	CDD:182619 (100)	transcriptional regulator NarL; Provisional (ZP_11111382)
18132	Crp	9314	9901	588	CDD:31008 (99.39)	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases [Signal transduction mechanisms]; COG0664 (ZP_11111381)

18132	MopB CT Nitrate-R-NarG-like	64	486	423	CDD:30308 (99.08)	Respiratory nitrate reductase A (NarGHI), alpha chain (NarG) and related proteins. Under anaerobic conditions in the presence of nitrate, <i>E. coli</i> synthesizes the cytoplasmic membrane-bound quinol-nitrate oxidoreductase (NarGHI), which reduces nitrate to...; cd02776 (ZP_10670320)
18132	Molybdopterin-Binding	1243	1626	384	CDD:209095 (100)	Molybdopterin-Binding (MopB) domain of the MopB superfamily of proteins, a large, diverse, heterogeneous superfamily of enzymes that, in general, bind molybdopterin as a cofactor. The MopB domain is found in a wide variety of molybdenum- and...; cl09928 (ZP_10670320)
18132	CAP ED	9332	9679	348	CDD:28920 (100)	effector domain of the CAP family of transcription factors; members include CAP (or cAMP receptor protein (CRP)), which binds cAMP, FNR (fumarate and nitrate reduction), which uses an iron-sulfur cluster to sense oxygen) and CooA, a heme containing CO...; cd00038 (ZP_11111381)
18132	REC	8580	8924	345	CDD:29071 (100)	Signal receiver domain; originally thought to be unique to bacteria (CheY, OmpR, NtrC, and PhoB), now recently identified in eukaryotes ETR1 <i>Arabidopsis thaliana</i> ; this domain receives the signal from the sensor partner in a two-component systems; cd00156 (ZP_11111382)
18132	PilJ	6859	7161	303	CDD:205851 (97.48)	Type IV pili methyl-accepting chemotaxis transducer N-term; pfam13675 (ZP_11111383)
18132	HATPase c	8257	8511	255	CDD:28956 (100)	Histidine kinase-like ATPases; This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerases, heat shock protein HSP90, phytochrome-like ATPases and DNA mismatch repair proteins; cd00075 (ZP_11111383)
18132	MopB Nitrate-R-NarG-like	2679	2906	228	CDD:73319 (100)	Respiratory nitrate reductase A (NarGHI), alpha chain (NarG) and related proteins. Under anaerobic conditions in the presence of nitrate, <i>E. coli</i> synthesizes the cytoplasmic membrane-bound quinol-nitrate oxidoreductase (NarGHI), which reduces nitrate to...; cd02750 (ZP_11111386)
18132	HisKA 3	7924	8121	198	CDD:203743 (99.69)	Histidine kinase; pfam07730 (ZP_11111383)
18132	Molybdopterin-Binding	2139	2309	171	CDD:209095 (96.91)	Molybdopterin-Binding (MopB) domain of the MopB superfamily of proteins, a large, diverse, heterogeneous superfamily of enzymes that, in general, bind molybdopterin as a cofactor. The MopB domain is found in a wide variety of molybdenum- and...; cl09928 (ZP_11111386)
18132	LuxR C like	9009	9179	171	CDD:99777 (100)	C-terminal DNA-binding domain of LuxR-like proteins. This domain contains a helix-turn-helix motif and binds DNA. Proteins belonging to this group are response regulators; some act as transcriptional activators, others as transcriptional repressors. Many...; cd06170 (ZP_11111382)
18132	HAMP	7294	7437	144	CDD:100122 (100)	Histidine kinase, Adenylyl cyclase, Methyl-accepting protein, and Phosphatase (HAMP) domain. HAMP is a signaling domain which occurs in a wide variety of signaling proteins, many of which are bacterial. The HAMP domain consists of two alpha helices...; cd06225 (ZP_11111383)
18132	Molybdopterin-Binding	1996	2133	138	CDD:209095 (100)	Molybdopterin-Binding (MopB) domain of the MopB superfamily of proteins, a large, diverse, heterogeneous superfamily of enzymes that, in general, bind molybdopterin as a cofactor. The MopB domain is found in a wide variety of molybdenum- and...; cl09928 (ZP_10670320)
18132	HTH CRP	9791	9901	111	CDD:128696 (100)	helix_turn_helix, cAMP Regulatory protein; smart00419 (ZP_11111381)
18132	Molybdopterin-Binding	550	642	93	CDD:209095 (100)	Molybdopterin-Binding (MopB) domain of the MopB superfamily of proteins, a large, diverse, heterogeneous superfamily of enzymes that, in general, bind molybdopterin as a cofactor. The MopB domain is found in a wide variety of molybdenum- and...; cl09928 (ZP_10670320)
23284	salicylate mono	2225	3409	1185	CDD:132263 (99.05)	salicylate 1-monooxygenase; TIGR03219 (ACV05012)
23284	PRK06475	2261	3403	1143	CDD:180582 (99.01)	salicylate hydroxylase; Provisional (ACV05012)

23284	LysR	3666	4550	885	CDD:223656 (96.24)	Transcriptional regulator [Transcription]; COG0583 (YP_006456979)
23284	PBP2 DntR NahR LinR like	3942	4544	603	CDD:176148 (94.58)	The C-terminal substrate binding domain of LysR-type transcriptional regulators that are involved in the catabolism of dinitrotoluene, naphthalene and gamma-hexachlorohexane; contains the type 2 periplasmic binding fold; cd08459 (YP_006456979)
23284	catechol 2 3	855	1403	549	CDD:234146 (100)	catechol 2,3 dioxygenase; TIGR03211 (YP_006456976)
23284	OH muco semi DH	1	462	462	CDD:132260 (100)	2-hydroxymuconic semialdehyde dehydrogenase; TIGR03216 (YP_006456975)
23284	ALDH F8 HMSADH	1	405	405	CDD:143412 (100)	Human aldehyde dehydrogenase family 8 member A1-like; cd07093 (YP_006456975)
23284	2 3 CTD N	1044	1403	360	CDD:176686 (100)	N-terminal domain of catechol 2,3-dioxygenase; cd07265 (YP_006456976)
23284	rve	5090	5383	294	CDD:201381 (80.42)	Integrase core domain; pfam00665 (NP_542848)
23284	rve 3	4976	5176	201	CDD:205859 (93.6)	Integrase core domain; pfam13683 (NP_542848)
23284	HTH 1	3672	3848	177	CDD:215735 (100)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (YP_006456979)
23284	HTH Tnp 1	5914	6081	168	CDD:248217 (100)	Transposase; cl17663 (WP_003450969)
23284	2 3 CTD C	855	986	132	CDD:176667 (100)	C-terminal domain of catechol 2,3-dioxygenase; cd07243 (YP_006456976)
23284	DDE Tnp IS66 C	4612	4704	93	CDD:205990 (86.25)	IS66 C-terminal element; pfam13817 (WP_003349068)
23284	Glo EDI BRP like	569	634	66	CDD:246679 (100)	This domain superfamily is found in a variety of structurally related metalloproteins, including the type I extradiol dioxygenases, glyoxalase I and a group of antibiotic resistance proteins; cl14632 (WP_003451984)
23284	DDE Tnp IS66	4720	4761	42	CDD:217337 (100)	Transposase IS66 family; pfam03050 (WP_003349068)
243	AcrB	32410	35505	3096	CDD:31183 (80.06)	Cation/multidrug efflux pump [Defense mechanisms]; COG0841 (YP_003673793)
243	pepN	39038	41761	2724	CDD:184453 (65.38)	aminopeptidase N; Provisional; PRK14015 (YP_314657)
243	GluZincin	39041	41725	2685	CDD:209905 (65.17)	Peptidase Gluzincin family (thermolysin-like proteinases, TLPs) includes peptidases M1, M2, M3, M4, M13, M32 and M36 (fungalsins); cl14813 (YP_314657)
243	PRK15399	28009	30162	2154	CDD:185297 (87.63)	lysine decarboxylase LdcC; Provisional (YP_005027486)
243	metG	58029	60098	2070	CDD:178889 (71.96)	methionyl-tRNA synthetase; Reviewed; PRK00133 (YP_283801)
243	PRK10917	14049	16049	2001	CDD:182836 (61.74)	ATP-dependent DNA helicase RecG; Provisional (YP_283585)
243	recQ	9400	11205	1806	CDD:130456 (70.68)	ATP-dependent DNA helicase RecQ; TIGR01389 (YP_003165790)
243	PhoX	7371	9122	1752	CDD:33024 (62.09)	Predicted phosphatase [General function prediction only]; COG3211 (YP_005028859)
243	MsbA rel	45369	47042	1674	CDD:131259 (76.78)	ABC transporter, permease/ATP-binding protein; TIGR02204 (YP_001352884)
243	pyrG	53536	55140	1605	CDD:180047 (81.87)	CTP synthetase; Validated; PRK05380 (YP_005028515)
243	Peptidase M17	20012	21412	1401	CDD:48344 (50.1)	Cytosol aminopeptidase family, N-terminal and catalytic domains. Family M17 contains zinc- and manganese-dependent exopeptidases (EC 3.4.11.1), including leucine aminopeptidase. They catalyze removal of amino acids from the N-terminus of a protein and...; cd00433 (ZP_10382070)

243	outer NodT	36846	38198	1353	CDD:162557 (59.5)	efflux transporter, outer membrane factor (OMF) lipoprotein, NodT family; TIGR01845 (YP_003673795)
243	OKR DC 1	28489	29793	1305	CDD:201705 (92.28)	Orn/Lys/Arg decarboxylase, major domain; pfam01276 (YP_005027486)
243	eno	56049	57329	1281	CDD:234617 (86.21)	enolase; Provisional; PRK00077 (YP_007551347)
243	enolase	56064	57278	1215	CDD:239429 (86.58)	Enolase: Enolases are homodimeric enzymes that catalyse the reversible dehydration of 2-phospho-D-glycerate to phosphoenolpyruvate as part of the glycolytic and gluconeogenesis pathways. The reaction is facilitated by the presence of metal ions; cd03313 (YP_007551347)
243	OAT like; PRK05964	49458	50624	1167	CDD:180329; CDD:99735 (64.07)	Acetyl ornithine aminotransferase family. This family belongs to pyridoxal phosphate (PLP)-dependent aspartate aminotransferase superfamily (fold I). The major groups in this CD correspond to ornithine aminotransferase, acetylorithine aminotransferase; cd00610; adenosylmethionine--8-amino-7-oxononanoate transaminase; Provisional (YP_001602164)
243	PRK06814	12871	13995	1125	CDD:180708 (49.15)	acylglycerophosphoethanolamine acyltransferase; Provisional (YP_283632)
243	MetRS core	58035	59129	1095	CDD:173907 (86.16)	catalytic core domain of methioninyl-tRNA synthetases; cd00814 (YP_283801)
243	PRK11670	52090	53106	1017	CDD:183270 (67.2)	antiporter inner membrane protein; Provisional (YP_283803)
243	RND mfp	35705	36679	975	CDD:162505 (65.84)	RND family efflux transporter, MFP subunit; TIGR01730 (YP_003673794)
243	leuO	31295	32194	900	CDD:181918 (25.2)	leucine transcriptional activator; Reviewed; PRK09508 (YP_934022)
243	LysR	41899	42774	876	CDD:223656 (72.16)	Transcriptional regulator [Transcription]; COG0583 (WP_004332332)
243	NPD like	11477	12343	867	CDD:73392 (84.8)	2-Nitropropane dioxygenase (NPD), one of the nitroalkane oxidizing enzyme families, catalyzes oxidative denitrification of nitroalkanes to their corresponding carbonyl compounds and nitrites. NDP is a member of the NAD(P)H-dependent flavin oxidoreductase...; cd04730 (YP_004848376)
243	SBP bac 8	5734	6564	831	CDD:205594 (60.18)	Bacterial extracellular solute-binding protein; pfam13416 (ZP_08274435)
243	COG1741	43592	44419	828	CDD:31927 (76.63)	Pirin-related protein [General function prediction only] (YP_522086)
243	COG0679	640	1464	825	CDD:31023 (67.34)	Predicted permeases [General function prediction only] (YP_003167592)
243	PRK05198	55191	55988	798	CDD:179961 (82.15)	2-dehydro-3-deoxyphosphooctonate aldolase; Provisional (YP_160545)
243	DAHP synth 1	55203	55994	792	CDD:189723 (82.63)	DAHP synthetase I family; pfam00793 (YP_160545)
243	PflA	23314	24102	789	CDD:31373 (82.97)	Pyruvate-formate lyase-activating enzyme [Posttranslational modification, protein turnover, chaperones]; COG1180 (YP_521681)
243	AfuA	5788	6561	774	CDD:32025 (60.74)	ABC-type Fe3+ transport system, periplasmic component [Inorganic ion transport and metabolism]; COG1840 (ZP_08274435)
243	MEMO like	21703	22470	768	CDD:153373 (63.03)	Memo (mediator of ErbB2-driven cell motility) is co-precipitated with the C terminus of ErbB2, a protein involved in cell motility; cd07361 (ZP_10380478)
243	ABC membrane	45474	46235	762	CDD:201380 (74.53)	ABC transporter transmembrane region; pfam00664 (YP_001352884)
243	CTGs	53542	54300	759	CDD:48377 (90.09)	CTP synthetase (CTPs) is a two-domain protein, which consists of an N-terminal synthetase domain and C-terminal glutaminase domain. The enzymes hydrolyze the amide bond of glutamine to ammonia and glutamate at the glutaminase domains and transfer nascent...; cd03113 (YP_005028515)

243	GATase1 CTP Synthase	54400	55140	741	CDD:153217 (75.06)	Type 1 glutamine amidotransferase (GATase1) domain found in Cytidine Triphosphate Synthetase; cd01746 (YP_005028515)
243	EAL	1878	2603	726	CDD:238923 (56.66)	EAL domain. This domain is found in diverse bacterial signaling proteins. It is called EAL after its conserved residues and is also known as domain of unknown function 2 (DUF2). The EAL domain has been shown to stimulate degradation of a second...; cd01948 (WP_008938741)
243	TS Pyrimidine HMase	25784	26437	654	CDD:58645 (78.37)	Thymidylate synthase and pyrimidine hydroxymethylase: Thymidylate synthase (TS) and deoxycytidylate hydroxymethylase (dCMP-HMase) are homologs that catalyze analogous alkylation of C5 of pyrimidine nucleotides. Both enzymes are involved in the...; cd00351 (YP_001170962)
243	AAA 26	38374	39024	651	CDD:222178 (36.43)	AAA domain; pfam13500 (YP_001234988)
243	bioD	38374	39021	648	CDD:234625 (36.85)	dithiobiotin synthetase; Reviewed; PRK00090 (YP_001234988)
243	P-loop NTPase	46395	47042	648	CDD:213113 (84.23)	P-loop containing Nucleoside Triphosphate Hydrolases; cl09099 (YP_001352884)
243	MRP-like	52189	52815	627	CDD:73300 (77.42)	MRP (Multiple Resistance and pH adaptation) is a member of the Fer4_NifH superfamily. Like the other members of the superfamily, MRP contains a ATP-binding domain at the N-termini. It is found in bacteria as a membrane-spanning protein and functions...; cd02037 (YP_283803)
243	PBP2 DntR NahR LinR like	31286	31903	618	CDD:176148 (19.03)	The C-terminal substrate binding domain of LysR-type transcriptional regulators that are involved in the catabolism of dinitrotoluene, naphthalene and gamma-hexachlorohexane; contains the type 2 periplasmic binding fold; cd08459 (YP_934022)
243	PBP2 CrgA like	41911	42507	597	CDD:176114 (69.13)	The C-terminal substrate binding domain of LysR-type transcriptional regulator CrgA and its related homologs, contains the type 2 periplasmic binding domain; cd08422 (WP_004332332)
243	AMMECR1	22577	23167	591	CDD:202019 (46.02)	AMMECR1; pfam01871 (YP_521682)
243	COG2199	2649	3191	543	CDD:225109 (52.99)	c-di-GMP synthetase (diguanylate cyclase, GGDEF domain) [Signal transduction mechanisms] (WP_008938741)
243	FMN red	42882	43424	543	CDD:212217 (73.15)	NADPH-dependent FMN reductase; cl00438 (ZP_08772527)
243	Radical SAM	23440	23970	531	CDD:100105 (85.15)	Radical SAM superfamily. Enzymes of this family generate radicals by combining a 4Fe-4S cluster and S-adenosylmethionine (SAM) in close proximity. They are characterized by a conserved CxxxCxxC motif, which coordinates the conserved iron-sulfur cluster; cd01335 (YP_521681)
243	MogA MoaB	17619	18116	498	CDD:58167 (74.59)	MogA_MoaB family. Members of this family are involved in biosynthesis of the molybdenum cofactor (MoCF) an essential cofactor of a diverse group of redox enzymes. MoCF biosynthesis is an evolutionarily conserved pathway present in eubacteria, archaea; cd00886 (YP_004847839)
243	Ycel	44550	45041	492	CDD:197935 (80.58)	Ycel-like domain; smart00867 (YP_285996)
243	PRTases typel	25133	25612	480	CDD:206754 (77.29)	Phosphoribosyl transferase (PRT)-type I domain; cd06223 (YP_315338)
243	GGDEF	2655	3131	477	CDD:143635 (52.07)	Diguanylate-cyclase (DGC) or GGDEF domain; cd01949 (WP_008938741)
243	Abhydrolase 5	24407	24883	477	CDD:205024 (69.33)	Alpha/beta hydrolase family; pfam12695 (YP_521680)
243	FhlA	3510	3980	471	CDD:225113 (40.09)	FOG: GAF domain [Signal transduction mechanisms]; COG2203 (WP_008938741)
243	GGDEF	19006	19461	456	CDD:143635 (34.14)	Diguanylate-cyclase (DGC) or GGDEF domain; cd01949 (YP_003885762)

243	COG3108	47224	47664	441	CDD:32922 (58.34)	Uncharacterized protein conserved in bacteria [Function unknown] (YP_006048844)
243	DEXDc	9487	9924	438	CDD:28927 (77.2)	DEAD-like helicases superfamily. A diverse family of proteins involved in ATP-dependent RNA or DNA unwinding. This domain contains the ATP-binding region; cd00046 (YP_003165790)
243	pgi	26699	27133	435	CDD:178917 (47.78)	glucose-6-phosphate isomerase; Reviewed; PRK00179 (YP_283808)
243	Peptidase M15 3	47224	47658	435	CDD:120464 (59.39)	Peptidase M15; cl01194 (YP_006048844)
243	GAF	3537	3965	429	CDD:216590 (40.34)	GAF domain; pfam01590 (WP_008938741)
243	HELICc	14334	14753	420	CDD:28960 (73.1)	Helicase superfamily c-terminal domain; associated with DEXDc-, DEAD-, and DEAH-box proteins, yeast initiation factor 4A, Ski2p, and Hepatitis C virus NS3 helicases; this domain is found in a wide variety of helicases and helicase related proteins; may...; cd00079 (YP_283585)
243	YkuD	48232	48651	420	CDD:202749 (35.46)	L,D-transpeptidase catalytic domain; pfam03734 (YP_902845)
243	FolA	27511	27924	414	CDD:223340 (58.43)	Dihydrofolate reductase [Coenzyme metabolism]; COG0262 (WP_006221722)
243	LabA like	3	413	411	CDD:199895 (94.63)	LabA_like proteins; cd06167 (YP_003522750)
243	DHFR	27511	27921	411	CDD:238127 (58.16)	Dihydrofolate reductase (DHFR). Reduces 7,8-dihydrofolate to 5,6,7,8-tetrahydrofolate with NADPH as a cofactor. This is an essential step in the biosynthesis of deoxythymidine phosphate since 5,6,7,8-tetrahydrofolate is required to regenerate 5; cd00209 (WP_006221722)
243	Anticodon Ia Met	59157	59564	408	CDD:153411 (48.33)	Anticodon-binding domain of methionyl tRNA synthetases; cd07957 (YP_283801)
243	DEXDc	14832	15236	405	CDD:28927 (63.57)	DEAD-like helicases superfamily. A diverse family of proteins involved in ATP-dependent RNA or DNA unwinding. This domain contains the ATP-binding region; cd00046 (YP_283585)
243	OKR DC 1 C	28024	28416	393	CDD:112521 (84.68)	Orn/Lys/Arg decarboxylase, C-terminal domain; pfam03711 (YP_005027486)
243	HELICc	10009	10389	381	CDD:28960 (80.36)	Helicase superfamily c-terminal domain; associated with DEXDc-, DEAD-, and DEAH-box proteins, yeast initiation factor 4A, Ski2p, and Hepatitis C virus NS3 helicases; this domain is found in a wide variety of helicases and helicase related proteins; may...; cd00079 (YP_003165790)
243	OKR DC 1 N	29809	30177	369	CDD:146376 (76.5)	Orn/Lys/Arg decarboxylase, N-terminal domain; pfam03709 (YP_005027486)
243	SIS PGI 2	26699	27058	360	CDD:88411 (48.78)	Phosphoglucose isomerase (PGI) contains two SIS (Sugar ISomerase) domains. This classification is based on the alignment of the second SIS domain. PGI is a multifunctional enzyme which as an intracellular dimer catalyzes the reversible isomerization of...; cd05016 (YP_283808)
243	tRNA bind EcMetRS like	59775	60098	324	CDD:48402 (84.46)	tRNA-binding-domain-containing Escherichia coli methionyl-tRNA synthetase (EcMetRS)-like proteins. This family includes EcMetRS and Aquifex aeolicus Trbp111 (AaTrbp111). This domain has general tRNA binding properties. MetRS aminoacylates methionine...; cd02800 (YP_283801)
243	HlyD 3	36110	36427	318	CDD:205615 (68.82)	HlyD family secretion protein; pfam13437 (YP_003673794)
243	YjgF YER057c UK114 family	16213	16527	315	CDD:100004 (84.6)	YjgF, YER057c, and UK114 belong to a large family of proteins present in bacteria, archaea, and eukaryotes with no definitive function. The conserved domain is similar in structure to chorismate mutase but there is no sequence similarity and no...; cd00448 (YP_005026571)
243	Pirin C	44108	44419	312	CDD:203319 (71.01)	Pirin C-terminal cupin domain; pfam05726 (YP_522086)

243	PAS	3174	3482	309	CDD:238075 (50.61)	PAS domain; PAS motifs appear in archaea, eubacteria and eukarya. Probably the most surprising identification of a PAS domain was that in EAG-like K ⁺ -channels. PAS domains have been found to bind ligands, and to act as sensors for light and oxygen in...; cd00130 (WP_008938741)
243	Pirin	43640	43945	306	CDD:202346 (83.6)	Pirin; pfam02678 (YP_522086)
243	PAS 9	3171	3470	300	CDD:222120 (49.37)	PAS domain; pfam13426 (WP_008938741)
243	RQC	10606	10881	276	CDD:198024 (79.96)	This DNA-binding domain is found in the RecQ helicase among others and has a helix-turn-helix structure; smart00956 (YP_003165790)
243	trimeric dUTPase	30452	30709	258	CDD:143638 (96.76)	Trimeric dUTP diphosphatases; cd07557 (YP_005027487)
243	RecG wedge OBF	15720	15959	240	CDD:72960 (52.37)	RecG_wedge_OBF: A subfamily of OB folds corresponding to the OB fold found in the N-terminal (wedge) domain of Escherichia coli RecG. RecG is a branched-DNA-specific helicase, which catalyzes the interconversion of a DNA replication fork to a...; cd04488 (YP_283585)
243	DUF59	52885	53109	225	CDD:202026 (39.02)	Domain of unknown function DUF59; pfam01883 (YP_283803)
243	HRDC	11002	11205	204	CDD:201312 (62.22)	HRDC domain; pfam00570 (YP_003165790)
243	ftsB	57422	57616	195	CDD:179156 (70.94)	cell division protein FtsB; Reviewed; PRK00888 (YP_160547)
243	SPOR	16656	16841	186	CDD:113793 (52.94)	Sporulation related domain; pfam05036 (ZP_08504949)
243	HTH 1	31997	32173	177	CDD:201021 (51.83)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (YP_934022)
243	LysM	48721	48849	129	CDD:212030 (30.3)	Lysine Motif is a small domain involved in binding peptidoglycan; cd00118 (YP_902845)
243	HTH 1	42646	42768	123	CDD:215735 (81.63)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (WP_004332332)
243	Rel-Spo like	27210	27302	93	CDD:245818 (86.75)	RelA- and SpoT-like ppGpp Synthetases and Hydrolases, catalytic domain; cl11966 (ACN22627)
243	sensory box	4035	4040	6	CDD:232884 (88.89)	PAS domain S-box; TIGR00229 (WP_008938741)
3075	COG1033	3	1829	1827	CDD:31236 (66.8)	Predicted exporters of the RND superfamily [General function prediction only] (ZP_10148962)
3075	DUF1302	4612	6144	1533	CDD:191660 (60.19)	Protein of unknown function (DUF1302); pfam06980 (YP_003452219)
3075	HcaE	10597	11769	1173	CDD:34257 (99.81)	Phenylpropionate dioxygenase and related ring-hydroxylating dioxygenases, large terminal subunit [Inorganic ion transport and metabolism / General function prediction only]; COG4638 (ACO92377)
3075	DUF1329	3246	4340	1095	CDD:148578 (56.91)	Protein of unknown function (DUF1329); pfam07044 (YP_003452218)
3075	gentsiate 1 2	8799	9800	1002	CDD:131325 (99.14)	gentsiate 1,2-dioxygenase; TIGR02272 (ACO92375)
3075	COG4447	2003	2974	972	CDD:34127 (44.2)	Uncharacterized protein related to plant photosystem II stability/assembly factor [General function prediction only] (YP_003607681)
3075	PRK07609	7771	8733	963	CDD:181058 (99.11)	CDP-6-deoxy-delta-3,4-glucoseen reductase; Validated (ACO92374)
3075	COG3547	16325	17284	960	CDD:33349 (88.63)	Transposase and inactivated derivatives [DNA replication, recombination, and repair] (ACO92381)
3075	leuO	6750	7625	876	CDD:181918 (98.92)	leucine transcriptional activator; Reviewed; PRK09508 (ACO92380)

3075	MA	14809	15579	771	CDD:197627 (100)	Methyl-accepting chemotaxis-like domains (chemotaxis sensory transducer); smart00283 (ACO92382)
3075	MhpD	9810	10502	693	CDD:30528 (100)	2-keto-4-pentenoate hydratase/2-oxohepta-3-ene-1,7-dioic acid hydratase (catechol pathway) [Secondary metabolites biosynthesis, transport, and catabolism]; COG0179 (ACO92376)
3075	RHO alpha C ahdA1c-like	11131	11793	663	CDD:176889 (100)	C-terminal catalytic domain of the large/alpha subunit (ahdA1c) of a ring-hydroxylating dioxygenase from <i>Sphingomonas</i> sp. strain P2 and related proteins; cd08880 (ACO92377)
3075	O2ase reductase like	8083	8733	651	CDD:99784 (98.68)	The oxygenase reductase FAD/NADH binding domain acts as part of the multi-component bacterial oxygenases which oxidize hydrocarbons using oxygen as the oxidant. Electron transfer is from NADH via FAD (in the oxygenase reductase) and an [2Fe-2S]...; cd06187 (ACO92374)
3075	maiA	12824	13459	636	CDD:162274 (98.75)	maleylacetoacetate isomerase; TIGR01262 (ACO92383)
3075	FAA hydrolase	9888	10502	615	CDD:201859 (100)	Fumarylacetoacetate (FAA) hydrolase family; pfam01557 (ACO92376)
3075	PBP2 DntR NahR LinR like	6738	7340	603	CDD:176148 (98.45)	The C-terminal substrate binding domain of LysR-type transcriptional regulators that are involved in the catabolism of dinitrotoluene, naphthalene and gamma-hexachlorohexane; contains the type 2 periplasmic binding fold; cd08459 (ACO92380)
3075	MCP signal	14956	15393	438	CDD:206779 (100)	Methyl-accepting chemotaxis protein (MCP), signaling domain; cd11386 (ACO92382)
3075	ring hydroxylating dioxygenases beta	11852	12271	420	CDD:29629 (100)	Ring hydroxylating dioxygenase beta subunit. This subunit has a similar structure to NTF-2, Ketosteroid isomerase and scytalone dehydratase. The degradation of aromatic compounds by aerobic bacteria frequently begins with the dihydroxylation of the...; cd00667 (ACO92378)
3075	DEDD Tnp IS110	16880	17278	399	CDD:201852 (76.16)	Transposase; pfam01548 (ACO92381)
3075	Rieske	10675	11061	387	CDD:207253 (100)	Rieske domain; a [2Fe-2S] cluster binding domain commonly found in Rieske non-heme iron oxygenase (RO) systems such as naphthalene and biphenyl dioxygenases, as well as in plant/cyanobacterial chloroplast b6f and mitochondrial cytochrome bc(1) complexes; cl00938 (ACO92377)
3075	GST C Zeta	13091	13450	360	CDD:198300 (97.84)	C-terminal, alpha helical domain of Class Zeta Glutathione S-transferases; cd03191 (ACO92383)
3075	Rieske RO ferredoxin	12294	12587	294	CDD:58551 (100)	Rieske non-heme iron oxygenase (RO) family, Rieske ferredoxin component; composed of the Rieske ferredoxin component of some three-component RO systems including biphenyl dioxygenase (BPDO) and carbazole 1,9a-dioxygenase (CARDO). The RO family comprise a...; cd03528 (ACO92379)
3075	fer2	7771	8016	246	CDD:29262 (100)	2Fe-2S iron-sulfur cluster binding domain. Iron-sulfur proteins play an important role in electron transfer processes and in various enzymatic reactions. The family includes plant and algal ferredoxins, which act as electron carriers in photosynthesis...; cd00207 (ACO92374)
3075	Transposase 20	16439	16666	228	CDD:202223 (100)	Transposase IS116/IS110/IS902 family; pfam02371 (ACO92381)
3075	GST N Zeta	12821	13045	225	CDD:48591 (100)	GST_N family, Class Zeta subfamily; GSTs are cytosolic dimeric proteins involved in cellular detoxification by catalyzing the conjugation of glutathione (GSH) with a wide range of endogenous and xenobiotic alkylating agents, including carcinogens; cd03042 (ACO92383)
3075	Cupin 2	9045	9251	207	CDD:203791 (100)	Cupin domain; pfam07883 (ACO92375)
3075	HTH 1	7431	7610	180	CDD:201021 (100)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (ACO92380)
3075	HAMP	14611	14751	141	CDD:197640 (100)	HAMP (Histidine kinases, Adenylyl cyclases, Methyl binding proteins, Phosphatases) domain; smart00304 (ACO92382)

3148	DEAH box HrpA	12214	15861	3648	CDD:162629 (99.2)	ATP-dependent helicase HrpA; TIGR01967 (ZP_10635276)
3148	PtrB	42525	44558	2034	CDD:31956 (96.24)	Protease II [Amino acid transport and metabolism]; COG1770 (ZP_10636063)
3148	DUF3418	12211	13980	1770	CDD:204772 (98.68)	Domain of unknown function (DUF3418); pfam11898 (ZP_10635276)
3148	MdlB	5064	6809	1746	CDD:31327 (96.2)	ABC-type multidrug transport system, ATPase and permease components [Defense mechanisms]; COG1132 (ZP_10641574)
3148	COG4579	16341	18041	1701	CDD:34217 (98.06)	Isocitrate dehydrogenase kinase/phosphatase [Signal transduction mechanisms] (ZP_11112323)
3148	aceK	16341	18035	1695	CDD:179509 (98.05)	bifunctional isocitrate dehydrogenase kinase/phosphatase protein; Validated; PRK02946 (ZP_11112323)
3148	PutA	21302	22780	1479	CDD:31216 (96.1)	NAD-dependent aldehyde dehydrogenases [Energy production and conversion]; COG1012 (ZP_11112325)
3148	COG3333	30622	32082	1461	CDD:33142 (99.68)	Uncharacterized protein conserved in bacteria [Function unknown] (ZP_10638996)
3148	ALDH KGSADH	21317	22636	1320	CDD:143447 (96.53)	Alpha-Ketoglutaric Semialdehyde Dehydrogenase; cd07129 (ZP_11112325)
3148	COG3825	39218	40393	1176	CDD:33618 (99.42)	Uncharacterized protein conserved in bacteria [Function unknown] (ZP_10699138)
3148	NAT SF	7784	8902	1119	CDD:213096 (96.84)	N-Acyltransferase superfamily: Various enzymes that characteristically catalyze the transfer of an acyl group to a substrate; cl00357 (ZP_10635273)
3148	PRK07515	10705	11820	1116	CDD:181012 (99.2)	3-oxoacyl-(acyl carrier protein) synthase III; Reviewed (ZP_10638982)
3148	KAS III	10714	11817	1104	CDD:29417 (99.19)	Ketoacyl-acyl carrier protein synthase III (KASIII) initiates the elongation in type II fatty acid synthase systems. It is found in bacteria and plants. Elongation of fatty acids in the type II systems occurs by Claisen condensation of malonyl-acyl...; cd00830 (ZP_10638982)
3148	DUF482	7784	8881	1098	CDD:146791 (96.78)	Protein of unknown function, DUF482; pfam04339 (ZP_10635273)
3148	GguC	22857	23840	984	CDD:33596 (97.81)	Uncharacterized protein conserved in bacteria [Function unknown]; COG3802 (ZP_10635281)
3148	COG3181	29116	30033	918	CDD:32994 (98.07)	Uncharacterized protein conserved in bacteria [Function unknown] (ZP_10636304)
3148	CysK	36116	37033	918	CDD:30381 (98.64)	Cysteine synthase [Amino acid transport and metabolism]; COG0031 (ZP_10641869)
3148	CBS like	36116	37006	891	CDD:107204 (98.6)	CBS_like: This subgroup includes Cystathionine beta-synthase (CBS) and Cysteine synthase. CBS is a unique heme-containing enzyme that catalyzes a pyridoxal 5'-phosphate (PLP)-dependent condensation of serine and homocysteine to give cystathionine; cd01561 (ZP_10641869)
3148	LysR	2617	3468	852	CDD:30928 (99.31)	Transcriptional regulator [Transcription]; COG0583 (ZP_10633660)
3148	LysR	32299	33150	852	CDD:30928 (99.01)	Transcriptional regulator [Transcription]; COG0583 (ZP_11112334)
3148	NMT1 2	29236	30024	789	CDD:212369 (98.18)	NMT1-like family; cl15260 (ZP_10636304)
3148	COG0121	41086	41835	750	CDD:30470 (99.19)	Predicted glutamine amidotransferase [General function prediction only] (ZP_10636065)
3148	ABC membrane	5121	5867	747	CDD:207103 (95.94)	ABC transporter transmembrane region; cl00549 (ZP_10641574)
3148	YafJ	41089	41835	747	CDD:48481 (99.18)	Glutamine amidotransferases class-II (Gn-AT)_YafJ-type. YafJ is a glutamine amidotransferase-like protein of unknown function found in prokaryotes, eukaryotes and archaea. YafJ has a conserved structural fold similar to those of other class II...; cd01908 (ZP_10636065)

3148	ABCC MsbA	6078	6791	714	CDD:73010 (97.57)	MsbA is an essential ABC transporter, closely related to eukaryotic MDR proteins. ABC transporters are a large family of proteins involved in the transport of a wide variety of different compounds, like sugars, ions, peptides, and more complex organic...; cd03251 (ZP_10641574)
3148	PRK05420	7072	7752	681	CDD:180067 (98.45)	aquaporin Z; Provisional (ZP_10638979)
3148	Peptidase S9	43908	44561	654	CDD:201156 (96.61)	Prolyl oligopeptidase family; pfam00326 (ZP_10636063)
3148	PRK00170	1474	2070	597	CDD:178912 (99.22)	azoreductase; Reviewed (ZP_11112310)
3148	LysR substrate	2875	3468	594	CDD:202651 (99.01)	LysR substrate binding domain; pfam03466 (ZP_10633660)
3148	PBP2 LTTR substrate	32533	33126	594	CDD:209302 (98.57)	The substrate binding domain of LysR-type transcriptional regulators (LTTRs), a member of the type 2 periplasmic binding fold protein superfamily; cl11398 (ZP_11112334)
3148	PRK10903	4346	4891	546	CDD:182824 (99.03)	peptidyl-prolyl <i>cis-trans</i> isomerase A (rotamase A); Provisional (ZP_10598296)
3148	FAA hydrolase	22893	23435	543	CDD:201859 (99.29)	Fumarylacetoacetate (FAA) hydrolase family; pfam01557 (ZP_10635281)
3148	DUF2937	40553	41047	495	CDD:151599 (98.7)	Protein of unknown function (DUF2937); pfam11157 (ZP_10673866)
3148	cyclophilin	4433	4888	456	CDD:213082 (98.85)	cyclophilin: cyclophilin-type peptidylprolyl <i>cis-trans</i> isomerases. This family contains eukaryotic, bacterial and archeal proteins which exhibit a peptidylprolyl <i>cis-trans</i> isomerases activity (PPlase, Rotamase) and in addition bind the...; cl00197 (ZP_10598296)
3148	AdoMet MTases	45737	46192	456	CDD:213457 (98.6)	S-adenosylmethionine-dependent methyltransferases (SAM or AdoMet-MTase), class I; AdoMet-MTases are enzymes that use S-adenosyl-L-methionine (SAM or AdoMet) as a substrate for methyltransfer, creating the product S-adenosyl-L-homocysteine (AdoHcy); cl16911 (ZP_11112352)
3148	PRK05170	49265	49705	441	CDD:235356 (97.95)	hypothetical protein; Provisional (WP_003184256)
3148	TctB	30134	30571	438	CDD:203614 (93.99)	Tripartite tricarboxylate transporter TctB family; pfam07331 (ZP_10638995)
3148	AAA	38495	38917	423	CDD:197690 (99.59)	ATPases associated with a variety of cellular activities; smart00382 (ZP_10654702)
3148	AAA	38495	38914	420	CDD:99707 (99.59)	The AAA+ (ATPases Associated with a wide variety of cellular Activities) superfamily represents an ancient group of ATPases belonging to the ASCE (for additional strand, catalytic E) division of the P-loop NTPase fold. The ASCE division also includes ABC; cd00009 (ZP_10654702)
3148	DEXDc	15433	15849	417	CDD:28927 (100)	DEAD-like helicases superfamily. A diverse family of proteins involved in ATP-dependent RNA or DNA unwinding. This domain contains the ATP-binding region; cd00046 (ZP_10635276)
3148	DUF748	37584	37994	411	CDD:191261 (93.68)	Domain of Unknown Function (DUF748); pfam05359 (ZP_11112338)
3148	HELICc	14860	15267	408	CDD:28960 (100)	Helicase superfamily c-terminal domain; associated with DEXDc-, DEAD-, and DEAH-box proteins, yeast initiation factor 4A, Ski2p, and Hepatitis C virus NS3 helicases; this domain is found in a wide variety of helicases and helicase related proteins; may...; cd00079 (ZP_10635276)
3148	Crp	44836	45237	402	CDD:31008 (96.69)	cAMP-binding proteins - catabolite gene activator and regulatory subunit of cAMP-dependent protein kinases [Signal transduction mechanisms]; COG0664 (ZP_10670039)
3148	CAP ED	44836	45180	345	CDD:28920 (96.2)	effector domain of the CAP family of transcription factors; members include CAP (or cAMP receptor protein (CRP)), which binds cAMP, FNR (fumarate and nitrate reduction), which uses an iron-sulfur cluster to sense oxygen) and CooA, a heme containing CO...; cd00038 (ZP_10670039)

3148	PRK11295	46403	46729	327	CDD:183079 (100)	hypothetical protein; Provisional (ZP_10596712)
3148	OB NTP bind	14014	14313	300	CDD:191824 (98.48)	Oligonucleotide/oligosaccharide-binding (OB)-fold; pfam07717 (ZP_10635276)
3148	EamA	33580	33873	294	CDD:144477 (93.35)	EamA-like transporter family; pfam00892 (ZP_11112335)
3148	HA2	14431	14703	273	CDD:190974 (100)	Helicase associated domain (HA2); pfam04408 (ZP_10635276)
3148	YcgL	51841	52062	222	CDD:147380 (100)	YcgL domain; pfam05166 (ZP_10596705)
3148	DEAH box HrpA	15864	16067	204	CDD:162629 (100)	ATP-dependent helicase HrpA; TIGR01967 (ZP_11112322)
3148	HisKA	287	478	192	CDD:119399 (100)	Histidine Kinase A (dimerization/phosphoacceptor) domain; Histidine Kinase A dimers are formed through parallel association of 2 domains creating 4-helix bundles; usually these domains contain a conserved His residue and are activated via...; cd00082 (YP_004352605)
3148	DUF2892	47571	47762	192	CDD:204593 (92.62)	Protein of unknown function (DUF2892); pfam11127 (ZP_11112355)
3148	HTH 1	2620	2796	177	CDD:201021 (100)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (ZP_10633660)
3148	HNHc	46454	46621	168	CDD:28969 (100)	HNH nucleases; HNH endonuclease signature which is found in viral, prokaryotic, and eukaryotic proteins. The alignment includes members of the large group of homing endonucleases, yeast intron 1 protein, MutS, as well as bacterial colicins, pyocins, and...; cd00085 (ZP_10596712)
3148	HTH 1	32299	32445	147	CDD:201021 (100)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (ZP_11112334)
3148	HATPase c	2	121	120	CDD:28956 (94.97)	Histidine kinase-like ATPases; This family includes several ATP-binding proteins for example: histidine kinase, DNA gyrase B, topoisomerases, heat shock protein HSP90, phytochrome-like ATPases and DNA mismatch repair proteins; cd00075 (YP_004352605)
33223	MDR; tdh	1	822	822	CDD:180054; CDD:211475 (100)	L-threonine 3-dehydrogenase; Validated; PRK05396; Medium chain reductase/dehydrogenase (MDR)/zinc-dependent alcohol dehydrogenase-like family; cl16912 (AAC38358)
33223	DDE Tnp ISL3	1754	2065	312	CDD:201887 (97.46)	Transposase; pfam01610 (AAF23984)
3721	sulP	13,199	14,827	1,629	CDD:162054 (80.71)	high affinity sulphate transporter 1; TIGR00815 (YP_294916)
3721	PRK03562	<4978	6,561	>1584	CDD:235131 (58.74)	glutathione-regulated potassium-efflux system protein KefC; Provisional (WP_003109220)
3721	OH muco semi DH	37,240	38,688	1,449	CDD:132260 (95.89)	2-hydroxymuconic semialdehyde dehydrogenase; TIGR03216 (YP_002798089)
3721	GltB	6,668	8,080	1,413	CDD:30418 (83.97)	Glutamate synthase domain 2 [Amino acid transport and metabolism]; COG0069 (ZP_11258118)
3721	ALDH F8 HMSADH	37,297	38,682	1,386	CDD:143412 (95.87)	Human aldehyde dehydrogenase family 8 member A1-like; cd07093 (YP_002798089)
3721	GltS FMN	6,809	8,071	1,263	CDD:73370 (84.77)	Glutamate synthase (GltS) FMN-binding domain. GltS is a complex iron-sulfur flavoprotein that catalyzes the reductive synthesis of L-glutamate from 2-oxoglutarate and L-glutamine via intramolecular channelling of ammonia, a reaction in the plant, yeast...; cd02808 (ZP_11258118)
3721	PRK01388	11,789	13,033	1,245	CDD:234949 (83.92)	arginine deiminase; Provisional (WP_003451732)
3721	salicylate mono	29,905	>31056	>1152	CDD:132263 (88.78)	salicylate 1-monooxygenase; TIGR03219 (YP_534831)
3721	PRK06475	29,926	>31056	>1131	CDD:180582 (89.03)	salicylate hydroxylase; Provisional (YP_534831)

3721	MFS	911	1,993	1,083	CDD:119392 (78.34)	The Major Facilitator Superfamily (MFS) is a large and diverse group of secondary transporters that includes uniporters, symporters, and antiporters. MFS proteins facilitate the transport across cytoplasmic or internal membranes of a variety of...; cd06174 (ZP_18875393)
3721	MFS 1	920	1,942	1,023	CDD:191813 (79.08)	Major Facilitator Superfamily; pfam07690 (ZP_18875393)
3721	PRK02102	10,690	11,694	1,005	CDD:179366 (93.79)	ornithine carbamoyltransferase; Validated (YP_001186606)
3721	2A0115	<3134	32,338	>990	CDD:233175 (81.3)	benzoate transport; TIGR00895 (WP_003450504)
3721	PRK08300	39,602	40,537	936	CDD:181366 (90.31)	acetaldehyde dehydrogenase; Validated (ZP_10704174)
3721	KefB	<4978	5,892	>915	CDD:223551 (64.68)	Kef-type K ⁺ transport systems, membrane components [Inorganic ion transport and metabolism]; COG0475 (WP_003109220)
3721	catechol 2 3	36,293	37,204	912	CDD:234146 (83.77)	catechol 2,3 dioxygenase; TIGR03211 (YP_002798088)
3721	PRK12454	9,774	10,670	897	CDD:183535 (87.24)	carbamate kinase-like carbamoyl phosphate synthetase; Reviewed (YP_001186605)
3721	AAK CK	9,774	10,667	894	CDD:58601 (87.2)	AAK_CK: Carbamate kinase (CK) catalyzes both the ATP-phosphorylation of carbamate and carbamoyl phosphate (CP) utilization with the production of ATP from ADP and CP. Both CK (this CD) and nonhomologous CP synthetase synthesize carbamoyl phosphate, an...; cd04235 (YP_001186605)
3721	lucif BA3436	26,618	27,511	894	CDD:163333 (77.44)	luciferase-type oxidoreductase, BA3436 family; TIGR03571 (YP_006536113)
3721	LysR	22,934	>23815	>882	CDD:223656 (79.23)	Transcriptional regulator [Transcription]; COG0583 (AGH87151)
3721	PRK08195	40,567	>41448	>882	CDD:181282 (94.32)	4-hydroxy-2-oxovalerate/4-hydroxy-2-oxopentanoic acid aldolase.; Validated (WP_003448357)
3721	PRK08163	<1691	17,785	>873	CDD:181262 (96.98)	salicylate hydroxylase; Provisional (ZP_19212574)
3721	PRK14997	27,719	28,582	864	CDD:184959 (81.81)	LysR family transcriptional regulator; Provisional (YP_006389807)
3721	NADB Rossmann	<1691	17,773	>861	CDD:214164 (96.95)	Rossmann-fold NAD(P)(+)-binding proteins; cl09931 (ZP_19212574)
3721	LysR	18,124	18,981	858	CDD:30928 (66.82)	Transcriptional regulator [Transcription]; COG0583 (ZP_10673017)
3721	LysR	28,780	>29619	>840	CDD:30928 (87.59)	Transcriptional regulator [Transcription]; COG0583 (AAM18547)
3721	Sulfate transp	13,694	14,497	804	CDD:144493 (78.69)	Sulfate transporter family; pfam00916 (YP_294916)
3721	DRE TIM HOA	40,576	41,358	783	CDD:163681 (94.06)	4-hydroxy-2-oxovalerate aldolase, N-terminal catalytic TIM barrel domain; cd07943 (WP_003448357)
3721	Abhydrolase 6	21,857	22,600	744	CDD:205026 (75.2)	Alpha/beta hydrolase family; pfam12697 (YP_006389795)
3721	PPK2 P aer	8,876	9,562	687	CDD:213852 (88.16)	polyphosphate kinase 2, PA0141 family; TIGR03707 (WP_003450823)
3721	BphB TodD	19,182	19,865	684	CDD:132368 (47.39)	cis-2,3-dihydrobiphenyl-2,3-diol dehydrogenase; TIGR03325 (YP_002362926)
3721	carb red sniffer like SDR c	19,194	19,868	675	CDD:187586 (48.85)	carbonyl reductase sniffer-like, classical (c) SDRs; cd05325 (YP_002362926)
3721	lolE	24,986	25,660	675	CDD:31279 (81.79)	Sugar phosphate isomerases/epimerases [Carbohydrate transport and metabolism]; COG1082 (YP_006536111)

3721	OprD	<3332 5	33,975	>651	CDD:202689 (74.07)	outer membrane porin, OprD family; pfam03573 (ZP_10708092)
3721	ChrR	2,062	>2709	>648	CDD:33600 (58.4)	Transcriptional activator [Transcription]; COG3806 (YP_003167582)
3721	PBP2 DntR NahR LinR like	28,786	29,388	603	CDD:176148 (87.87)	The C-terminal substrate binding domain of LysR-type transcriptional regulators that are involved in the catabolism of dinitrotoluene, naphthalene and gamma-hexachlorohexane; contains the type 2 periplasmic binding fold; cd08459 (AAM18547)
3721	PBP2 CrgA like 8	23,201	23,791	591	CDD:176166 (79.41)	The C-terminal substrate binding domain of an uncharacterized LysR-type transcriptional regulator CrgA-like, contains the type 2 periplasmic binding fold; cd08477 (AGH87151)
3721	PBP2 CrgA like 7	27,992	28,582	591	CDD:176165 (79.63)	The C-terminal substrate binding domain of an uncharacterized LysR-type transcriptional regulator CrgA-like, contains the type 2 periplasmic binding fold; cd08476 (YP_006389807)
3721	PBP2 CrgA like 6	18,139	18,720	582	CDD:176164 (63.78)	The C-terminal substrate binding domain of an uncharacterized LysR-type transcriptional regulator CrgA-like, contains the type 2 periplasmic binding fold; cd08475 (ZP_10673017)
3721	Abhydrolase 2	20,635	21,210	576	CDD:216940 (64.56)	Phospholipase/Carboxylesterase; pfam02230 (WP_003411083)
3721	Flavin utilizing monooxygenases 1	<2692	27,469	>549	CDD:213112 (77.04)	Flavin-utilizing monooxygenases; cl07892 (YP_006536113)
3721	AP endonuc 2	25,956	26,492	537	CDD:201692 (81.11)	Xylose isomerase-like TIM barrel; pfam01261 (YP_006536112)
3721	OTCace	10,699	11,226	528	CDD:215776 (91.26)	Aspartate/ornithine carbamoyltransferase, Asp/Orn binding domain; pfam00185 (YP_001186606)
3721	HAD 2	15,575	16,102	528	CDD:205597 (55.71)	Haloacid dehalogenase-like hydrolase; pfam13419 (YP_007242137)
3721	OprD	32,798	>33301	>504	CDD:202689 (69.68)	outer membrane porin, OprD family; pfam03573 (YP_004713956)
3721	FAA hydrolase 1	<3908	39,581	>501	CDD:245608 (97.54)	Fumarylacetoacetate (FAA) hydrolase family; cl11421 (YP_709325)
3721	PRK11460 5	<2072	21,216	>492	CDD:183144 (64.67)	putative hydrolase; Provisional (WP_003411083)
3721	AcetDehyd-dimer	39,995	40,456	462	CDD:150078 (91.83)	Prokaryotic acetaldehyde dehydrogenase, dimerisation; pfam09290 (ZP_10704174)
3721	2 3 CTD C	36,716	37,144	429	CDD:176667 (85.35)	C-terminal domain of catechol 2,3-dioxygenase; cd07243 (YP_002798088)
3721	OTCace N	11,248	11,673	426	CDD:217204 (96.19)	Aspartate/ornithine carbamoyltransferase, carbamoyl-P binding domain; pfam02729 (YP_001186606)
3721	DUF4197	<4074	4,451	>378	CDD:206023 (84.47)	Protein of unknown function (DUF4197); pfam13852 (YP_004475573)
3721	2 3 CTD N	36,293	36,658	366	CDD:176686 (78.45)	N-terminal domain of catechol 2,3-dioxygenase; cd07265 (YP_002798088)
3721	Semialdehyde dh	39,617	39,961	345	CDD:197927 (92.83)	Semialdehyde dehydrogenase, NAD binding domain; smart00859 (ZP_10704174)
3721	TrkA N	5,998	6,330	333	CDD:216949 (61.49)	TrkA-N domain; pfam02254 (WP_003109220)
3721	STAS SulP like sulfate transporter	13,208	13,531	324	CDD:132913 (77.93)	Sulphate Transporter and Anti-Sigma factor antagonist domain of SulP-like sulfate transporters, plays a role in the function and regulation of the transport activity, proposed general NTP binding function; cd07042 (YP_294916)

3721	MFS	<3134 9	>31660	>312	CDD:119392 (96.18)	The Major Facilitator Superfamily (MFS) is a large and diverse group of secondary transporters that includes uniporters, symporters, and antiporters. MFS proteins facilitate the transport across cytoplasmic or internal membranes of a variety of...; cd06174 (WP_003450504)
3721	Cupin 7	2,110	2,382	273	CDD:193446 (70.89)	ChrR Cupin-like domain; pfam12973 (YP_003167582)
3721	DUF167	3,400	3,657	258	CDD:242111 (83.93)	Uncharacterized ACR, YggU family COG1872; cl00811 (WP_006894058)
3721	Sulfate tra GLY	14,582	14,830	249	CDD:205965 (88.41)	Sulfate transporter N-terminal domain with GLY motif; pfam13792 (YP_294916)
3721	HAD like	15,806	16,048	243	CDD:119389 (54)	Haloacid dehalogenase-like hydrolases. The haloacid dehalogenase-like (HAD) superfamily includes L-2-haloacid dehalogenase, epoxide hydrolase, phosphoserine phosphatase, phosphomannomutase, phosphoglycolate phosphatase, P-type ATPase, and many others; cd01427 (YP_007242137)
3721	FAA hydrolase	38,825	>39061	>237	CDD:245608 (83.66)	Fumarylacetoacetate (FAA) hydrolase family; cl11421 (ZP_19212538)
3721	DctP; SBP bac 7	<3490 1	>35125	>225	CDD:210106; CDD:31825 (43.6)	Bacterial extracellular solute-binding protein, family 7; cl15441; TRAP-type C4-dicarboxylate transport system, periplasmic component [Carbohydrate transport and metabolism]; COG1638 (YP_005938337)
3721	ABM	20,050	20,265	216	CDD:202845 (65.87)	Antibiotic biosynthesis monooxygenase; pfam03992 (ZP_10557216)
3721	salicylate mono	35,599	>35811	>213	CDD:132263 (69.73)	salicylate 1-monooxygenase; TIGR03219 (NP_863106)
3721	FlIB	2,958	>3161	>204	CDD:112504 (80.05)	Flagellin N-methylase; pfam03692 (YP_002892604)
3721	Cupin 7	<2500	2,697	>198	CDD:193446 (46.27)	ChrR Cupin-like domain; pfam12973 (YP_003167582)
3721	KefB	4,787	>4981	>195	CDD:223551 (67.78)	Kef-type K ⁺ transport systems, membrane components [Inorganic ion transport and metabolism]; COG0475 (WP_003082491)
3721	UbiH	35,617	>35811	>195	CDD:30999 (78.69)	2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases [Coenzyme metabolism / Energy production and conversion]; COG0654 (NP_863106)
3721	PRK03562	4,790	>4981	>192	CDD:235131 (68.52)	glutathione-regulated potassium-efflux system protein KefC; Provisional (WP_003082491)
3721	HTH 1	22,937	23,116	180	CDD:215735 (83.67)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (AGH87151)
3721	HTH 1	27,728	27,904	177	CDD:201021 (89.86)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (YP_006389807)
3721	HTH 1	18,817	18,981	165	CDD:201021 (86.99)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (ZP_10673017)
3721	DUF4197	3,890	>4036	>147	CDD:206023 (87.14)	Protein of unknown function (DUF4197); pfam13852 (ZP_19203401)
3721	HTH 1	29,482	>29619	>138	CDD:201021 (93.04)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (AAM18547)
3721	KdpD	<1	>102	>102	CDD:145711 (76.33)	Osmosensitive K ⁺ channel His kinase sensor domain; pfam02702 (WP_003406071)
3721	DmpG comm	41,362	>41448	>87	CDD:149094 (100)	DmpG-like communication domain; pfam07836 (WP_003448357)
3721	PFDH like; fdhA non GSH	<3524 1	>35318	>78	CDD:163032; CDD:176242 (95.49)	<i>Pseudomonas putida</i> aldehyde-dismutating formaldehyde dehydrogenase (PFDH); cd08282; formaldehyde dehydrogenase, glutathione-independent; TIGR02819 (ZP_10994472)
3721	fer2	<3617 0	36,220	>51	CDD:29262 (89.77)	2Fe-2S iron-sulfur cluster binding domain. Iron-sulfur proteins play an important role in electron transfer processes and in various enzymatic reactions. The family includes plant and algal ferredoxins, which act as electron carriers in photosynthesis...; cd00207 (YP_005884801)

58390	PRK09906	375	1199	825	CDD:182137 (74.69)	DNA-binding transcriptional regulator HcaR; Provisional (YP_005090855)
58390	PBP2 LTRR substrate	615	1199	585	CDD:209302 (73.56)	The substrate binding domain of LysR-type transcriptional regulators (LTRRs), a member of the type 2 periplasmic binding fold protein superfamily; cl11398 (YP_005090855)
58390	HTH 1	375	524	150	CDD:201021 (86.18)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (YP_005090855)
58390	Tra8	1308	1334	27	CDD:225382 (100)	Transposase and inactivated derivatives, IS30 family [DNA replication, recombination, and repair]; COG2826 (WP_003461919)
5976	COG3039	249	1019	771	CDD:32853 (99.24)	Transposase and inactivated derivatives, IS5 family [DNA replication, recombination, and repair] (YP_004715165)
5976	DUF772	213	437	225	CDD:203284 (100)	Transposase domain (DUF772); pfam05598 (YP_004715165)
5976	DEAD	6290	6478	189	CDD:201124 (87.38)	DEAD/DEAH box helicase; pfam00270 (YP_622756)
5976	DDE 4	828	995	168	CDD:211471 (100)	DDE superfamily endonuclease; cl15789 (YP_004715165)
5976	HELICc	7892	8017	126	CDD:28960 (97.63)	Helicase superfamily c-terminal domain; associated with DEXDc-, DEAD-, and DEAH-box proteins, yeast initiation factor 4A, Ski2p, and Hepatitis C virus NS3 helicases; this domain is found in a wide variety of helicases and helicase related proteins; may...; cd00079 (YP_622756)
6160	OH muco semi DH	14261	15709	1449	CDD:132260 (100)	2-hydroxymuconic semialdehyde dehydrogenase; TIGR03216 (NP_863102)
6160	ALDH F8 HMSADH	14267	15652	1386	CDD:143412 (100)	Human aldehyde dehydrogenase family 8 member A1-like; cd07093 (NP_863102)
6160	salicylate mono	17595	18797	1203	CDD:132263 (100)	salicylate 1-monooxygenase; TIGR03219 (NP_863106)
6160	UbiH	17613	18767	1155	CDD:30999 (100)	2-polyprenyl-6-methoxyphenol hydroxylase and related FAD-dependent oxidoreductases [Coenzyme metabolism / Energy production and conversion]; COG0654 (NP_863106)
6160	PRK08195	10609	11616	1008	CDD:181282 (100)	4-hydroxy-2-oxovalerate/4-hydroxy-2-oxopentanoic acid aldolase.; Validated (NP_863098)
6160	catechol 2 3	15756	16667	912	CDD:163184 (100)	catechol 2,3 dioxygenase; TIGR03211 (NP_863103)
6160	PRK11139	3141	4034	894	CDD:182990 (100)	DNA-binding transcriptional activator GcvA; Provisional (NP_863090)
6160	DRE TIM HOA	10816	11607	792	CDD:163681 (100)	4-hydroxy-2-oxovalerate aldolase, N-terminal catalytic TIM barrel domain; cd07943 (NP_863098)
6160	biphenyl bphD	13390	14160	771	CDD:132386 (100)	2-hydroxy-6-oxo-6-phenylhexa-2,4-dienoate hydrolase; TIGR03343 (NP_863101)
6160	catechol dmpE	12578	13345	768	CDD:132264 (100)	2-oxopent-4-enoate hydratase; TIGR03220 (NP_863100)
6160	fabG	4366	5109	744	CDD:180133 (100)	3-ketoacyl-(acyl-carrier-protein) reductase; Provisional; PRK05565 (NP_863091)
6160	MA	7763	8485	723	CDD:197627 (100)	Methyl-accepting chemotaxis-like domains (chemotaxis sensory transducer); smart00283 (NP_863095)
6160	SDR c	4387	5103	717	CDD:212491 (100)	classical (c) SDRs; cd05233 (NP_863091)
6160	PBP2 GcdR TrpI HvrB AmpR like	3162	3764	603	CDD:176123 (100)	The C-terminal substrate domain of LysR-type GcdR, TrpI, HvrB and beta-lactamase regulators, and that of other closely related homologs; contains the type 2 periplasmic binding fold; cd08432 (NP_863090)
6160	catechol dmpH	9814	10410	597	CDD:132262 (100)	4-oxalocrotonate decarboxylase; TIGR03218 (NP_863097)
6160	MCP signal	7895	8419	525	CDD:206779 (100)	Methyl-accepting chemotaxis protein (MCP), signaling domain; cd11386 (NP_863095)

6160	4HB MCP 1	8780	9292	513	CDD:193205 (100)	Four helix bundle sensory module for signal transduction; pfam12729 (NP_863095)
6160	2 3 CTD C	15819	16244	426	CDD:176667 (100)	C-terminal domain of catechol 2,3-dioxygenase; cd07243 (NP_863103)
6160	SdiA-regulated	3	410	408	CDD:203557 (100)	SdiA-regulated; pfam06977 (NP_863086)
6160	SdiA-regulated	3	386	384	CDD:197380 (100)	SdiA-regulated; cd09971 (NP_863086)
6160	2 3 CTD N	16302	16667	366	CDD:176686 (100)	N-terminal domain of catechol 2,3-dioxygenase; cd07265 (NP_863103)
6160	HTH	6499	6849	351	CDD:213080 (100)	Helix-turn-helix domains; cl00088 (NP_863094)
6160	Phage-tail 3	2050	2391	342	CDD:205728 (88.89)	Putative phage tail protein; pfam13550 (YP_001354474)
6160	DksA	628	960	333	CDD:31920 (100)	DnaK suppressor protein [Signal transduction mechanisms]; COG1734 (NP_863087)
6160	Tra8	6952	7281	330	CDD:32654 (100)	Transposase and inactivated derivatives, IS30 family [DNA replication, recombination, and repair]; COG2826 (NP_943111)
6160	rve	5840	6163	324	CDD:201381 (100)	Integrase core domain; pfam00665 (NP_863093)
6160	USP Like	1041	1337	297	CDD:30165 (100)	Usp: Universal stress protein family. The universal stress protein Usp is a small cytoplasmic bacterial protein whose expression is enhanced when the cell is exposed to stress agents. Usp enhances the rate of cell survival during prolonged exposure to...; cd00293 (NP_863088)
6160	COG2963	6538	6825	288	CDD:32783 (100)	Transposase and inactivated derivatives [DNA replication, recombination, and repair] (NP_863094)
6160	DmpG comm	10615	10812	198	CDD:149094 (100)	DmpG-like communication domain; pfam07836 (NP_863098)
6160	HTH 1	3846	4034	189	CDD:201021 (100)	Bacterial regulatory helix-turn-helix protein, lysR family; pfam00126 (NP_863090)
6160	4Oxalocrotonate Tautomerase	9577	9750	174	CDD:29603 (100)	4-Oxalocrotonate Tautomerase: Catalyzes the isomerization of unsaturated ketones. The structure is a homo-hexamer that is arranged as a trimer of dimers. The hexamer contains six active sites, each formed by residues from three monomers, two from one...; cd00491 (NP_863096)
6160	HAMP	8555	8695	141	CDD:100122 (100)	Histidine kinase, Adenylyl cyclase, Methyl-accepting protein, and Phosphatase (HAMP) domain. HAMP is a signaling domain which occurs in a wide variety of signaling proteins, many of which are bacterial. The HAMP domain consists of two alpha helices...; cd06225 (NP_863095)
6160	Esterase lipase	13819	13941	123	CDD:211462 (100)	Esterases and lipases (includes fungal lipases, cholinesterases, etc.) These enzymes act on carboxylic esters (EC: 3.1.1.-). The catalytic apparatus involves three residues (catalytic triad): a serine, a glutamate or aspartate and a histidine. These...; cl12031 (NP_863101)
6160	HTH 21	6221	6307	87	CDD:205456 (100)	HTH-like domain; pfam13276 (NP_863093)
66283	HTH ARSR	266	397	132	CDD:28974 (32.74)	Arsenical Resistance Operon Repressor and similar prokaryotic, metal regulated homodimeric repressors. ARSR subfamily of helix-turn-helix bacterial transcription regulatory proteins (winged helix topology). Includes several proteins that appear to...; cd00090 (YP_004931645)
66283	Tra8	1044	1142	99	CDD:225382 (100)	Transposase and inactivated derivatives, IS30 family [DNA replication, recombination, and repair]; COG2826 (WP_003465026)
9794	AcoR	11578	13389	1812	CDD:33094 (97.94)	Transcriptional activator of acetoin/glycerol metabolism [Secondary metabolites biosynthesis, transport, and catabolism / Transcription]; COG3284 (ZP_10622512)

9794	PRK05347	8566	9936	1371	CDD:180031 (98.82)	glutaminyl-tRNA synthetase; Provisional (ZP_10622510)
9794	MalK	13763	14809	1047	CDD:33631 (99.21)	ABC-type sugar transport systems, ATPase components [Carbohydrate transport and metabolism]; COG3839 (ZP_10622513)
9794	AFG1 ATPase	461	1324	864	CDD:112768 (85.59)	AFG1-like ATPase; pfam03969 (ZP_10667796)
9794	LplB	15967	16791	825	CDD:33938 (99.93)	ABC-type polysaccharide transport system, permease component [Carbohydrate transport and metabolism]; COG4209 (ZP_10622515)
9794	UgpE	16895	17614	720	CDD:30744 (100)	ABC-type sugar transport system, permease component [Carbohydrate transport and metabolism]; COG0395 (ZP_10622516)
9794	MPP YbbF-LpxH	6869	7516	648	CDD:163641 (98.04)	Escherichia coli YbbF/LpxH and related proteins, metallophosphatase domain; cd07398 (ZP_10622508)
9794	ABC Carb Solutes like	13769	14407	639	CDD:73018 (99.25)	ABC Carbohydrate and Solute Transporters-like subgroup. This family is comprised of proteins involved in the transport of apparently unrelated solutes and proteins specific for di- and oligosaccharides and polyols. ABC transporters are a large family...; cd03259 (ZP_10622513)
9794	MalK	14857	15453	597	CDD:33631 (99.9)	ABC-type sugar transport systems, ATPase components [Carbohydrate transport and metabolism]; COG3839 (ZP_10622514)
9794	ABC Carb Solutes like	14866	15453	588	CDD:73018 (99.9)	ABC Carbohydrate and Solute Transporters-like subgroup. This family is comprised of proteins involved in the transport of apparently unrelated solutes and proteins specific for di- and oligosaccharides and polyols. ABC transporters are a large family...; cd03259 (ZP_10622514)
9794	tRNA-synt 1c C	9295	9870	576	CDD:202825 (99.9)	tRNA synthetases class I (E and Q), anti-codon binding domain; pfam03950 (ZP_10622510)
9794	TM PBP2	17018	17569	552	CDD:119394 (100)	Transmembrane subunit (TM) found in Periplasmic Binding Protein (PBP)-dependent ATP-Binding Cassette (ABC) transporters which generally bind type 2 PBPs. These types of transporters consist of a PBP, two TMs, and two cytoplasmic ABC ATPase subunits, and...; cd06261 (ZP_10622516)
9794	AAA	11986	12471	486	CDD:99707 (100)	The AAA+ (ATPases Associated with a wide variety of cellular Activities) superfamily represents an ancient group of ATPases belonging to the ASCE (for additional strand, catalytic E) division of the P-loop NTPase fold. The ASCE division also includes ABC; cd00009 (ZP_10622512)
9794	GlnRS core	8872	9303	432	CDD:185676 (99.25)	catalytic core domain of glutaminyl-tRNA synthetase; cd00807 (ZP_10622510)
9794	TM PBP2	16366	16770	405	CDD:119394 (99.86)	Transmembrane subunit (TM) found in Periplasmic Binding Protein (PBP)-dependent ATP-Binding Cassette (ABC) transporters which generally bind type 2 PBPs. These types of transporters consist of a PBP, two TMs, and two cytoplasmic ABC ATPase subunits, and...; cd06261 (ZP_10622515)
9794	HTH	1885	2238	354	CDD:213080 (93.56)	Helix-turn-helix domains; cl00088 (ZP_10623566)
9794	MalK	15581	15919	339	CDD:33631 (96.01)	ABC-type sugar transport systems, ATPase components [Carbohydrate transport and metabolism]; COG3839 (ZP_10622514)
9794	HTH MARR	1903	2205	303	CDD:197670 (92.56)	helix_turn_helix multiple antibiotic resistance protein; smart00347 (ZP_10623566)
9794	PRK10791	7702	7995	294	CDD:182734 (99.22)	peptidyl-prolyl <i>cis-trans</i> isomerase B (rotamase B); Provisional (ZP_10636222)
9794	cyclophilin EcCYP like	7702	7989	288	CDD:29391 (99.2)	cyclophilin_EcCYP_like: cyclophilin-type A-like peptidylprolyl <i>cis-trans</i> isomerase (PPIase) domain similar to the cytosolic E. coli cyclophilin A and Streptomyces antibioticus SanCyp18. Compared to the archetypal cyclophilin Human cyclophilin A, these...; cd01920 (ZP_10636222)

9794	TOBE 2	14585	14800	216	CDD:207413 (100)	TOBE domain; cl01440 (ZP_10622513)
9794	TOBE 2	15698	15907	210	CDD:203932 (100)	TOBE domain; pfam08402 (ZP_10622514)
9794	HTH 8	11578	11691	114	CDD:202485 (100)	Bacterial regulatory protein, Fis family; pfam02954 (ZP_10622512)
9794	nt trans	8566	8643	78	CDD:212170 (95.68)	nucleotidyl transferase superfamily; cl00015 (ZP_10622510)

B.3. BLASTp protein annotations.

BLASTp Protein hits for genes predicted on contigs used in analysis of SIGEX clones. Search hits are from the nr database.

Contig	Protein Name	Hit			Hit Accession (% Similarity to Protein)
		Start	End	Length	
14785	type VI secretion protein TssF	1	1212	1212	ZP_10427245 (96.15)
14785	hypothetical protein	3208	4164	957	YP_004467422 (30.08)
14785	putative LysR-type transcriptional regulator Protein	4784	5488	705	BAC53588 (65.9)
14785	hypothetical protein	1212	1724	513	YP_002875506 (95.47)
14785	hypothetical protein	2164	2601	438	YP_007399641 (73.29)
14785	hypothetical protein	1721	2110	390	ZP_10594862 (60.3)
14785	type VI secretion system effector Hcp1 Protein	2686	2757	72	YP_007399640 (100)
18132	respiratory nitrate reductase, alpha subunit Protein	1	2133	2133	ZP_10670320 (99.15)
18132	protein NarX	6754	8553	1800	ZP_11111383 (99.27)
18132	nitrite extrusion protein 2	3884	5179	1296	ZP_11111385 (98.85)
18132	protein NarK	5196	6410	1215	ZP_11111384 (95.3)
18132	nitrate reductase A subunit alpha Protein	2139	2906	768	ZP_11111386 (98.8)
18132	transcriptional regulator NarL Protein	8553	9197	645	ZP_11111382 (100)
18132	respiratory nitrate reductase, alpha subunit Protein	3079	3705	627	WP_008153585 (100)
18132	protein DnrS	9287	9901	615	ZP_11111381 (99.41)
23284	salicylate hydroxylase Protein	2186	3496	1311	ACV05012 (98.57)
23284	naphthalene degradation LysR-family transcriptional activator Protein	3651	4550	900	YP_006456979 (96.31)
23284	catechol 2,3-dioxygenase Protein	855	1403	549	YP_006456976 (100)
23284	2-hydroxymuconic semialdehyde dehydrogenase Protein	1	471	471	YP_006456975 (100)
23284	transposase subunit B Protein	4922	5383	462	NP_542848 (83.22)
23284	transposase IS3/IS911 family protein	5776	6081	306	WP_003450969 (100)
23284	putative transposase Protein	4612	4761	150	WP_003349068 (91.37)

23284	Catechol 2,3-dioxygenase Protein	509	634	126	WP_003451984 (100)
23284	transposase Protein	6452	6565	114	WP_005749725 (100)
243	acriflavin resistance protein	32410	35550	3141	YP_003673793 (79.4)
243	aminopeptidase Protein	39038	41761	2724	YP_314657 (65.38)
243	arginine/lysine/ornithine decarboxylase Protein	27997	30231	2235	YP_005027486 (88.06)
243	GAF and PAS/PAC sensor-containing diguanylate cyclase/phosphodiesterase Protein	1860	4040	2181	WP_008938741 (50.69)
243	methionyl-tRNA synthetase Protein	58026	60098	2073	YP_283801 (72)
243	ATP-dependent DNA helicase RecG Protein	14049	16049	2001	YP_283585 (61.74)
243	ATP-dependent DNA helicase RecQ Protein	9400	11205	1806	YP_003165790 (70.68)
243	putative phosphatase Protein	7371	9122	1752	YP_005028859 (62.09)
243	multidrug ABC transporter ATPase and permease Protein	45369	47096	1728	YP_001352884 (76.66)
243	CTP synthase Protein	53536	55176	1641	YP_005028515 (80.73)
243	hypothetical protein	20012	21436	1425	ZP_10382070 (50)
243	NodT family RND efflux system outer membrane lipoprotein	36813	38204	1392	YP_003673795 (58.77)
243	enolase Protein	56049	57329	1281	YP_007551347 (86.21)
243	2-nitropropane dioxygenase NPD Protein	11429	12670	1242	YP_004848376 (82.51)
243	adenosylmethionine-8-amino-7-oxononanoate aminotransferase Protein	49458	50624	1167	YP_001602164 (64.07)
243	phospholipid/glycerol acyltransferase Protein	12871	13995	1125	YP_283632 (49.15)
243	hypothetical protein	52024	53115	1092	YP_283803 (67.53)
243	radical SAM family protein	23200	24267	1068	YP_521681 (80.66)
243	Spermidine/putrescine-binding periplasmic protein	5665	6729	1065	ZP_08274435 (54.65)
243	RND family efflux transporter MFP subunit Protein	35645	36679	1035	YP_003673794 (64.41)
243	auxin efflux carrier Protein	619	1554	936	YP_003167592 (67.68)
243	LysR family transcriptional regulator Protein	31286	32194	909	YP_934022 (25.39)
243	LysR family transcriptional regulator Protein	41899	42786	888	WP_004332332 (72.03)
243	Pirin-like protein	43553	44419	867	YP_522086 (72.99)
243	ErfK/YbiS/YcfS/YnhG family protein	48034	48873	840	YP_902845 (31.26)
243	2-dehydro-3-deoxyphosphooctonate aldolase Protein	55182	56003	822	YP_160545 (81.72)

243	thymidylate synthase Protein	25778	26581	804	YP_001170962 (75.32)
243	hypothetical protein	21697	22479	783	ZP_10380478 (62.82)
243	diguanylate cyclase Protein	19006	19707	702	YP_003885762 (25.32)
243	dethiobiotin synthase Protein	38374	39030	657	YP_001234988 (36.52)
243	putative esterase Protein	18214	18861	648	WP_007510257 (58.79)
243	AMMECR1 domain-containing protein	22529	23167	639	YP_521682 (43.42)
243	dienelactone hydrolase Protein	24332	24955	624	YP_521680 (69.56)
243	hypothetical protein	25046	25669	624	YP_315338 (73.01)
243	molybdopterin biosynthesis protein	17619	18191	573	YP_004847839 (75.58)
243	deoxycytidine triphosphate deaminase Protein	30365	30928	564	YP_005027487 (92.55)
243	FMN-dependent NADH-azoreductase Protein	42882	43424	543	ZP_08772527 (73.15)
243	hypothetical protein	60308	60829	522	WP_008481570 (40.27)
243	Ycel Protein	44547	45047	501	YP_285996 (79.54)
243	twin-arginine translocation pathway signal Protein	47224	47673	450	YP_006048844 (57.88)
243	hypothetical protein	3	443	441	YP_003522750 (93.28)
243	glucose-6-phosphate isomerase Protein	26699	27133	435	YP_283808 (47.78)
243	dihydrofolate reductase Protein	27511	27924	414	WP_006221722 (58.43)
243	endoribonuclease L-PSP Protein	16201	16533	333	YP_005026571 (84.02)
243	conserved hypothetical protein	27186	27422	237	ACN22627 (80.5)
243	Putative sporulation related protein	16656	16874	219	ZP_08504949 (47.55)
243	cell division FtsB ortholog Protein	57422	57616	195	YP_160547 (70.94)
243	hypothetical protein	50989	51153	165	YP_286710 (56.23)
3075	methyl-accepting chemotaxis sensory transducer Protein	13555	15582	2028	ACO92382 (99.72)
3075	transporter Protein	3	1940	1938	ZP_10148962 (66.67)
3075	hypothetical protein	4612	6192	1581	YP_003452219 (59.31)
3075	hypothetical protein	3234	4508	1275	YP_003452218 (60.19)
3075	salicylate-5-hydroxylase large oxygenase component Protein	10543	11796	1254	ACO92377 (99.82)
3075	gentisate 1,2-dioxygenase Protein	8763	9803	1041	ACO92375 (99.17)

3075	transposase IS116/IS110/IS902 family Protein	16274	17296	1023	ACO92381 (88.72)
3075	salicylate 5-hydroxylase ferredoxin reductase component Protein	7771	8748	978	ACO92374 (99.12)
3075	BNR repeat-containing protein	2000	2974	975	YP_003607681 (44.36)
3075	LysR-type transcriptional regulator Protein	6732	7634	903	ACO92380 (98.95)
3075	fumarylpyruvate hydrolase Protein	9810	10502	693	ACO92376 (100)
3075	maleylpyruvate isomerase Protein	12815	13462	648	ACO92383 (98.77)
3075	salicylate-5-hydroxylase small oxygenase component Protein	11852	12271	420	ACO92378 (100)
3148	ATP-dependent helicase HrpA Protein	12205	15861	3657	ZP_10635276 (99.2)
3148	protease II Protein	42513	44564	2052	ZP_10636063 (96.07)
3148	ABC-type multidrug transport system, ATPase and permease component Protein	4995	6824	1830	ZP_10641574 (96.21)
3148	bifunctional isocitrate dehydrogenase kinase/phosphatase protein	16326	18041	1716	ZP_11112323 (98.08)
3148	methyl-accepting chemotaxis protein	18467	20068	1602	WP_008043471 (91.99)
3148	NADP-dependent fatty aldehyde dehydrogenase Protein	21224	22801	1578	ZP_11112325 (95.79)
3148	hypothetical protein	30622	32148	1527	ZP_10638996 (98.5)
3148	sugar phosphate permease Protein	27306	28643	1338	WP_008045228 (98.05)
3148	hypothetical protein	9188	10501	1314	ZP_10635274 (98.16)
3148	hypothetical protein	39218	40393	1176	ZP_10699138 (99.42)
3148	hypothetical protein	48017	49192	1176	ZP_10641821 (96.49)
3148	histidine kinase, Hybrid Protein	2	1168	1167	YP_004352605 (97.11)
3148	hypothetical protein	7784	8902	1119	ZP_10635273 (96.84)
3148	3-oxoacyl-(acyl-carrier-protein) synthase III	10705	11823	1119	ZP_10638982 (99.2)
3148	hypothetical protein	37170	38237	1068	ZP_11112338 (93.76)
3148	hypothetical protein	29023	30036	1014	ZP_10636304 (97.51)
3148	pyridoxal phosphate biosynthesis protein	25356	26357	1002	WP_007946808 (96.53)
3148	hypothetical protein	22848	23840	993	ZP_10635281 (97.83)
3148	cysteine synthase A Protein	36071	37042	972	ZP_10641869 (98.72)
3148	putative TIM-barrel fold metal-dependent hydrolase Protein	26357	27310	954	WP_008029057 (93.15)
3148	LysR family transcriptional regulator Protein	32221	33159	939	ZP_11112334 (98.26)

3148	phosphoglycerate dehydrogenase-like oxidoreductase Protein	50731	51660	930	WP_008034968 (98.57)
3148	transcriptional regulator Protein	2599	3519	921	ZP_10633660 (99.36)
3148	putative permease Protein	20253	21128	876	WP_008043472 (97.09)
3148	aspartyl/asparaginyl beta-hydroxylase-like dioxygenase Protein	34912	35787	876	WP_008035075 (97.39)
3148	putative hydrolase or acyltransferase of alpha/beta superfamily Protein	3537	4349	813	WP_008034293 (97.77)
3148	enolase superfamily enzyme related to L-alanine-DL-glutamate epimerase Protein	23874	24671	798	WP_007943924 (99.57)
3148	hypothetical protein	33574	34368	795	ZP_11112335 (95.63)
3148	putative glutamine amidotransferase Protein	41062	41835	774	ZP_10636065 (99.21)
3148	MoxR-like ATPase Protein	38495	39202	708	ZP_10654702 (96.27)
3148	MIP family channel protein	7066	7758	693	ZP_10638979 (98.47)
3148	spermidine synthase Protein	45632	46315	684	ZP_11112352 (99.07)
3148	ACP phosphodiesterase Protein	1474	2070	597	ZP_11112310 (99.22)
3148	nitroreductase Protein	49993	50583	591	WP_007984791 (99.15)
3148	peptidyl-prolyl <i>cis-trans</i> isomerase (rotamase) - cyclophilin family Protein	4346	4900	555	ZP_10598296 (99.04)
3148	Protein of unknown function (DUF2937)	40508	41047	540	ZP_10673866 (98.81)
3148	Tripartite tricarboxylate transporter TctB family Protein	30092	30622	531	ZP_10638995 (95.05)
3148	hypothetical protein	41869	42363	495	ZP_10636064 (95.99)
3148	cyclic nucleotide-binding protein	44779	45249	471	ZP_10670039 (97.22)
3148	protein of unknown function, UPF0153 family	49265	49711	447	WP_003184256 (97.98)
3148	hypothetical protein	47514	47885	372	ZP_11112355 (94.22)
3148	HNH endonuclease Protein	46367	46735	369	ZP_10596712 (99.41)
3148	hypothetical protein	2238	2549	312	ZP_10633659 (96.72)
3148	hypothetical protein	45252	45539	288	ZP_10653544 (83.2)
3148	hypothetical protein	51784	52065	282	ZP_10596705 (100)
3148	enolase superfamily enzyme related to L-alanine-DL-glutamate epimerase Protein	24793	25035	243	WP_008029049 (100)
3148	ATP-dependent helicase HrpA Protein	15864	16097	234	ZP_11112322 (100)
3148	putative ATP-binding protein	38389	38457	69	YP_007396413 (100)
33223	putative alcohol dehydrogenase NtnW Protein	1	822	822	AAC38358 (100)

33223	putative transposase TnpA2* Protein	1439	2068	630	AAF23984 (95.44)
3721	sulfate anion transporter Protein	<13163	>14842	>1680	YP_294916 (81.02)
3721	glutamate synthase Protein	6,605	>8206	>1602	ZP_11258118 (82.75)
3721	glutathione-regulated potassium-efflux system protein KefB	<4978	>6561	>1584	WP_003109220 (58.74)
3721	2-hydroxymuconic semialdehyde dehydrogenase Protein	37,231	38,688	1,458	YP_002798089 (95.92)
3721	Arginine deiminase Protein	11,783	13,033	1,251	WP_003451732 (84.01)
3721	salicylate hydroxylase NahG Protein	29,833	>31056	>1224	YP_534831 (88.12)
3721	benzoate MFS transporter BenK Protein	<31349	32,533	>1185	WP_003450504 (78.14)
3721	transporter, major facilitator family Protein	<899	>1993	>1095	ZP_18875393 (78.59)
3721	ornithine carbamoyltransferase Protein	10,687	11,694	1,008	YP_001186606 (93.8)
3721	luciferase family protein	26,594	27,583	990	YP_006536113 (74.03)
3721	luciferase family oxidoreductase, group 1 Protein	23,897	24,877	981	WP_008013356 (74.9)
3721	monooxygenase FAD-binding protein	<16913	>17857	>945	ZP_19212574 (97.2)
3721	acetaldehyde dehydrogenase (acetylating) Protein	39,602	40,537	936	ZP_10704174 (90.31)
3721	carbamate kinase Protein	9,747	10,670	924	YP_001186605 (85.73)
3721	Catechol 2,3 dioxygenase, XylE Protein	36,284	37,204	921	YP_002798088 (83.91)
3721	putative hydrolase signal peptide protein	<21839	>22744	>906	YP_006389795 (71.83)
3721	transcriptional regulator Protein	<18118	19,014	>897	ZP_10673017 (67.04)
3721	4-hydroxy-2-oxovalerate aldolase Protein	40,552	>41448	>897	WP_003448357 (94.1)
3721	Transcriptional regulator, LysR family Protein	22,928	>23815	>888	AGH87151 (79.38)
3721	LysR family transcriptional regulator Protein	<27719	>28594	>876	YP_006389807 (81.37)
3721	LysR-type transcriptional regulator NahR Protein	28,780	>29619	>840	AAM18547 (87.59)
3721	Hypothetical protein	8,825	9,646	822	WP_003450823 (83.89)
3721	xylose isomerase domain-containing protein	<24947	25,741	>795	YP_006536111 (81.45)
3721	xylose isomerase domain-containing protein	<25773	26,567	>795	YP_006536112 (80.72)
3721	short-chain dehydrogenase/reductase SDR Protein	19,167	19,868	702	YP_002362926 (47.02)
3721	Phospholipase/Carboxylesterase Protein	<20629	>21285	>657	WP_003411083 (60.8)
3721	outer membrane porin, OprD family Protein	<33325	33,978	>654	ZP_10708092 (74.2)

3721	anti-ECFsigma factor ChrR Protein	<2059	>2709	>651	YP_003167582 (58.1)
3721	haloacid dehalogenase superfamily protein	<15572	>16165	>594	YP_007242137 (56.77)
3721	benzoate-specific porin Protein	<32783	>33301	>519	YP_004713956 (69.56)
3721	2-oxopent-4-enoate hydratase Protein	<39081	39,584	>504	YP_709325 (97.44)
3721	hypothetical protein	<4074	>4496	>423	YP_004475573 (84.19)
3721	Hypothetical protein	<311	688	>378	AGH85959 (54.16)
3721	hypothetical protein	<8384	8,746	>363	YP_004379227 (56.65)
3721	hypothetical protein	20,038	20,388	351	ZP_10557216 (68.13)
3721	UPF0235 protein yggU	<3382	>3675	>294	WP_006894058 (81.98)
3721	2-oxopent-4-enoate hydratase Protein	38,813	>39061	>249	ZP_19212538 (82.2)
3721	salicylate hydroxylase Protein	<35584	>35811	>228	NP_863106 (70.44)
3721	TRAP dicarboxylate family transporter subunit DctP Protein	<34901	>35125	>225	YP_005938337 (43.6)
3721	hypothetical protein	2,949	>3161	>213	YP_002892604 (80.8)
3721	glutathione-regulated potassium-efflux system protein KefB	4,787	>4981	>195	WP_003082491 (67.78)
3721	hypothetical protein, 4-oxalocrotonate tautomerase	<21532	21,720	>189	WP_007970831 (82.56)
3721	hypothetical protein	3,884	>4036	>153	ZP_19203401 (85.6)
3721	osmosensitive K+ channel signal transduction histidine kinase Protein	<1	>102	>102	WP_003406071 (76.33)
3721	formaldehyde dehydrogenase Protein	<35241	>35318	>78	ZP_10994472 (95.49)
3721	chloroplast-type ferredoxin-like protein	<36170	>36241	>72	YP_005884801 (65.65)
58390	LysR family transcriptional regulator Protein	375	1229	855	YP_005090855 (74.9)
58390	Transposase for insertion sequence element IS1086 Protein	1272	1334	63	WP_003461919 (100)
5976	helicase-like protein	4475	8671	4197	YP_622756 (76.59)
5976	hypothetical protein	2001	4469	2469	YP_622757 (67.7)
5976	hypothetical protein	8676	10538	1863	YP_622755 (73.02)
5976	DDE-type transposase Protein	66	1043	978	YP_004715165 (98.99)
5976	N-6 DNA methylase Protein	1204	1635	432	YP_622758 (79.71)
6160	methyl-accepting chemotaxis protein	7712	9292	1581	NP_863095 (100)
6160	2-hydroxymuconic semialdehyde dehydrogenase Protein	14261	15718	1458	NP_863102 (100)

6160	salicylate hydroxylase Protein	17517	18815	1299	NP_863106 (100)
6160	4-hydroxy-2-ketovalerate aldolase Protein	10591	11628	1038	NP_863098 (100)
6160	catechol 2,3-dioxygenase Protein	15756	16676	921	NP_863103 (100)
6160	putative LysR-type transcriptional regulator Protein	3141	4052	912	NP_863090 (100)
6160	2-hydroxymuconic semialdehyde hydrolase Protein	13372	14250	879	NP_863101 (100)
6160	2-oxypent-4-pentenoate hydratase Protein	12575	13357	783	NP_863100 (100)
6160	putative oxidoreductase Protein	4357	5127	771	NP_863091 (100)
6160	hypothetical protein	1041	1739	699	NP_863088 (100)
6160	putative transposase Protein	5834	6499	666	NP_863093 (100)
6160	4-oxalocrotonate decarboxylase Protein	9811	10410	600	NP_863097 (100)
6160	hypothetical protein	3	578	576	NP_863086 (100)
6160	hypothetical protein	2400	2957	558	NP_863089 (100)
6160	hypothetical protein	2050	2427	378	YP_001354474 (89.91)
6160	hypothetical protein	628	984	357	NP_863087 (100)
6160	putative transposase Protein	6499	6849	351	NP_863094 (100)
6160	putative transposase Protein	6940	7281	342	NP_943111 (100)
6160	4-oxalocrotonate tautomerase Protein	9565	9753	189	NP_863096 (100)
6160	hypothetical protein	5204	5344	141	NP_863092 (100)
6160	hypothetical protein	5348	5482	135	NP_943108 (100)
66283	Fis family transcriptional regulator Protein	266	988	723	YP_004931645 (17.02)
66283	Transposase for insertion sequence element IS1086 Protein	1044	1142	99	WP_003465026 (100)
9794	transcriptional activator of acetoin/glycerol metabolism Protein	11572	13401	1830	ZP_10622512 (97.2)
9794	drug resistance transporter, EmrB/QacA subfamily Protein	5139	6665	1527	WP_008064201 (99.27)
9794	efflux transporter, outer membrane factor lipoprotein, NodT family	2283	3749	1467	WP_008064198 (96.44)
9794	cysteinyl-tRNA synthetase Protein	9957	11336	1380	WP_008020685 (99.71)
9794	glutaminyl-tRNA synthetase Protein	8566	9939	1374	ZP_10622510 (98.82)
9794	putative ATPase Protein	446	1639	1194	ZP_10667796 (78.05)
9794	multidrug resistance efflux pump Protein	3822	4967	1146	WP_008064199 (96.52)

9794	ATPase component of ABC-type sugar transporter Protein	13763	14854	1092	ZP_10622513 (99.24)
9794	permease component of ABC-type sugar transporter Protein	15958	16821	864	ZP_10622515 (99.93)
9794	ABC-type sugar transport system, permease component Protein	16832	17614	783	ZP_10622516 (100)
9794	UDP-2,3-diacetylglucosamine hydrolase Protein	6806	7522	717	ZP_10622508 (97.92)
9794	ATPase component of ABC-type sugar transporter Protein	14857	15453	597	ZP_10622514 (99.9)
9794	transcriptional regulator Protein	1813	2283	471	ZP_10623566 (93.22)
9794	ATPase component of ABC-type sugar transporter Protein	15581	15958	378	ZP_10622514 (96.4)
9794	peptidyl-prolyl <i>cis-trans</i> isomerase (rotamase) - cyclophilin family Protein	7702	8004	303	ZP_10636222 (99.24)
9794	hypothetical protein	3	80	78	ZP_10646039 (86.84)

B.4. InterProScan Annotations.

InterProScan Annotations for genes predicted on contigs used in analysis of SIGEX clones.

Contig	Database	InterProScan Name & Information	Start	End	Length	Database Entry ID
243	Gene3D	unintegrated	49566	50871	1306	G3DSA:1.20.1600.10
243	Gene3D	unintegrated	45465	46646	1125	G3DSA:1.20.1250.20
243	Gene3D	unintegrated	52072	53088	1014	G3DSA:3.40.50.300
243	Gene3D	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	56487	57401	915	G3DSA:3.40.640.10
243	Gene3D	Rossmann-like alpha/beta/alpha sandwich fold	31280	32140	861	G3DSA:3.40.50.620
243	Gene3D	unintegrated	5797	6603	807	G3DSA:3.40.190.10
243	Gene3D	Pyridoxal phosphate-dependent transferase, major region, subdomain 1	36996	37784	789	G3DSA:3.40.640.10
243	Gene3D	Aldolase-type TIM barrel	48160	48831	672	G3DSA:3.20.20.70
243	Gene3D	Rossmann-like alpha/beta/alpha sandwich fold	55281	55934	642	G3DSA:3.40.50.620
243	Gene3D	Porin domain	38371	39018	648	G3DSA:2.40.160.10
243	Gene3D	Xylose isomerase-like, TIM barrel domain	22532	23173	642	G3DSA:3.20.20.150
243	Gene3D	unintegrated	13459	14049	591	G3DSA:3.40.50.360
243	Gene3D	unintegrated	32656	33168	513	G3DSA:3.40.190.10
243	Gene3D	unintegrated	29728	30204	477	G3DSA:3.90.180.10
243	Gene3D	Aldehyde dehydrogenase, N-terminal	7194	7664	471	G3DSA:3.40.605.10
243	Gene3D	unintegrated	20603	21043	441	G3DSA:2.40.440.10
243	Gene3D	RmlC-like jelly roll fold	26714	27136	423	G3DSA:2.60.120.10
243	Gene3D	RmlC-like jelly roll fold	9385	9801	417	G3DSA:2.60.120.10
243	Gene3D	unintegrated	58029	58436	408	G3DSA:3.30.420.10
243	Gene3D	Endoribonuclease L-PSP/chorismate mutase-like	23161	23523	363	G3DSA:3.30.1330.40
243	Gene3D	unintegrated	11780	12115	336	G3DSA:3.40.50.2020
243	Gene3D	unintegrated	43871	44206	336	G3DSA:3.30.420.10
243	Gene3D	unintegrated	628	954	327	G3DSA:3.90.800.10

243	Gene3D	NAD(P)-binding domain	29422	29727	306	G3DSA:3.40.50.720
243	Gene3D	unintegrated	35753	36040	288	G3DSA:3.30.700.20
243	Gene3D	Pyridoxal phosphate-dependent transferase, major region, subdomain 2	37788	38066	279	G3DSA:3.90.1150.10
243	Gene3D	Protein of unknown function DUF167	47476	47754	279	G3DSA:3.30.1200.10
243	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	25493	25750	258	G3DSA:1.10.10.10
243	Gene3D	unintegrated	54007	54261	255	G3DSA:3.40.50.740
243	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	24329	24571	243	G3DSA:1.10.10.10
243	Gene3D	Glutamyl/glutaminyl-tRNA synthetase, class Ib, alpha-bundle domain	1114	1341	228	G3DSA:1.10.1160.10
243	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	32413	32634	222	G3DSA:1.10.10.10
243	Gene3D	unintegrated	54262	54441	180	G3DSA:3.90.55.10
243	Gene3D	unintegrated	36041	36205	165	G3DSA:3.30.1490.150
243	Gene3D	unintegrated	53536	53700	165	G3DSA:3.40.228.10
243	Gene3D	Rossmann-like alpha/beta/alpha sandwich fold	958	1113	156	G3DSA:3.40.50.620
243	Gene3D	unintegrated	21092	21220	129	G3DSA:3.10.350.10
243	Gene3D	Homeodomain-like	4857	4967	111	G3DSA:1.10.10.60
3075	Gene3D	unintegrated	3897	4568	672	G3DSA:3.40.50.300
3075	Gene3D	Xylose isomerase-like, TIM barrel domain	9858	10472	615	G3DSA:3.20.20.150
3075	Gene3D	NAD(P)-binding domain	7206	7634	429	G3DSA:3.40.50.720
3075	Gene3D	unintegrated	8799	9215	417	G3DSA:3.40.50.1010
3075	Gene3D	unintegrated	16577	16957	375	G3DSA:3.40.190.10
3075	Gene3D	unintegrated	14071	14451	375	G3DSA:3.40.190.10
3075	Gene3D	unintegrated	6915	7205	291	G3DSA:3.30.360.10
3075	Gene3D	unintegrated	840	1103	264	G3DSA:3.50.50.60
3075	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	13555	13812	258	G3DSA:1.10.10.10
3075	Gene3D	Nucleic acid-binding, OB-fold	7897	8061	165	G3DSA:2.40.50.140
3148	Gene3D	unintegrated	10516	11775	1113	G3DSA:1.20.1250.20
3148	Gene3D	Aldehyde dehydrogenase, N-terminal	41035	41832	798	G3DSA:3.40.605.10
3148	Gene3D	Aldolase-type TIM barrel	21989	22768	780	G3DSA:3.20.20.70

3148	Gene3D	Fumarylacetoacetase, C-terminal-related	20259	21032	774	G3DSA:3.90.850.10
3148	Gene3D	unintegrated	35176	35865	690	G3DSA:3.40.50.300
3148	Gene3D	unintegrated	48602	49222	621	G3DSA:3.40.50.1820
3148	Gene3D	unintegrated	36434	37033	600	G3DSA:3.40.50.300
3148	Gene3D	Molybdopterin binding domain	39230	39793	564	G3DSA:3.40.980.10
3148	Gene3D	unintegrated	27225	27653	429	G3DSA:1.10.390.10
3148	Gene3D	START-like domain	4352	4762	411	G3DSA:3.30.530.20
3148	Gene3D	unintegrated	44779	45177	399	G3DSA:3.40.1280.10
3148	Gene3D	unintegrated	2947	3324	378	G3DSA:3.40.50.300
3148	Gene3D	unintegrated	32224	32523	300	G3DSA:2.40.100.10
3148	Gene3D	unintegrated	33844	34143	300	G3DSA:3.40.190.10
3148	Gene3D	Nucleic acid-binding, OB-fold	8411	8701	291	G3DSA:2.40.50.140
3148	Gene3D	Histidine kinase-like ATPase, ATP-binding domain	6474	6752	279	G3DSA:3.30.565.10
3148	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	46406	46663	258	G3DSA:1.10.10.10
3148	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	34171	34419	249	G3DSA:1.10.10.10
3148	Gene3D	Sporulation-related domain	428	616	189	G3DSA:3.30.70.1070
3148	Gene3D	unintegrated	17277	17390	114	G3DSA:3.40.228.10
3148	Gene3D	unintegrated	40919	41023	105	G3DSA:3.40.50.300
3148	Gene3D	unintegrated	9743	9832	90	G3DSA:2.40.50.100
5976	Gene3D	unintegrated	78	896	819	G3DSA:1.10.3720.10
5976	Gene3D	unintegrated	1207	1686	480	G3DSA:3.40.50.300
6160	Gene3D	unintegrated	14768	15709	942	G3DSA:1.20.1560.10
6160	Gene3D	unintegrated	8021	8911	891	G3DSA:3.20.20.120
6160	Gene3D	unintegrated	3198	4043	846	G3DSA:3.10.180.10
6160	Gene3D	unintegrated	4366	5211	807	G3DSA:3.10.180.10
6160	Gene3D	RmIC-like jelly roll fold	2313	2906	594	G3DSA:2.60.120.10
6160	Gene3D	unintegrated	18252	18809	558	G3DSA:2.70.40.10
6160	Gene3D	unintegrated	12007	12552	546	G3DSA:3.10.180.10

6160	Gene3D	Fumarylacetoacetase, C-terminal-related	12854	13342	489	G3DSA:3.90.850.10
6160	Gene3D	unintegrated	8915	9283	369	G3DSA:3.30.390.10
6160	Gene3D	unintegrated	6236	6493	258	G3DSA:3.30.70.900
9794	Gene3D	NAD(P)-binding domain	15964	16719	756	G3DSA:3.40.50.720
9794	Gene3D	RmlC-like jelly roll fold	9981	10730	750	G3DSA:2.60.120.10
9794	Gene3D	unintegrated	5427	6167	741	G3DSA:3.40.50.300
9794	Gene3D	unintegrated	8845	9357	513	G3DSA:3.40.190.10
9794	Gene3D	Porin domain	6935	7429	495	G3DSA:2.40.160.10
9794	Gene3D	Dihydrofolate reductase-like domain	1208	1693	486	G3DSA:3.40.430.10
9794	Gene3D	unintegrated	12970	13404	435	G3DSA:3.40.50.2300
9794	Gene3D	NAD(P)-binding domain	16832	17263	432	G3DSA:3.40.50.720
9794	Gene3D	unintegrated	17264	17686	423	G3DSA:3.30.360.10
9794	Gene3D	Hedgehog signalling/DD-peptidase zinc-binding domain	2502	2849	348	G3DSA:3.30.1380.10
9794	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	8578	8817	240	G3DSA:1.10.10.10
9794	Gene3D	Winged helix-turn-helix transcription repressor DNA-binding domain	12796	12969	174	G3DSA:1.10.10.10
14785	Gene3D	unintegrated	3190	3810	621	G3DSA:3.40.50.1820
14785	Gene3D	Nitroreductase-like	1146	1691	546	G3DSA:3.40.109.10
14785	Gene3D	unintegrated	310	804	495	G3DSA:3.80.30.10
18132	Gene3D	unintegrated	6757	7914	1116	G3DSA:1.20.1250.20
18132	Gene3D	unintegrated	4334	5179	846	G3DSA:3.40.50.300
18132	Gene3D	unintegrated	3560	4321	762	G3DSA:3.40.50.880
18132	Gene3D	HAD-like domain	8556	9158	603	G3DSA:3.40.50.1000
18132	Gene3D	unintegrated	6087	6401	315	G3DSA:3.40.50.10490
18132	Gene3D	Homeodomain-like	1864	2112	249	G3DSA:1.10.10.60
18132	Gene3D	Phosphoglucose isomerase, C-terminal	5967	6086	120	G3DSA:1.10.1390.10
23284	Gene3D	Fumarylacetoacetase, C-terminal-related	3651	4247	597	G3DSA:3.90.850.10
23284	Gene3D	unintegrated	3122	3427	306	G3DSA:3.30.70.1430
58390	Gene3D	unintegrated	408	1235	828	G3DSA:3.40.50.1820

243	HAMAP	Azoreductase_type1; NADH-azoreductase, FMN-dependent	13456	14049	594	MF_01216
243	HAMAP	UPF0235; Protein of unknown function DUF167	47464	47730	267	MF_00634
3075	HAMAP	Ac_ald_DH_ac; Acetaldehyde dehydrogenase	6663	7634	972	MF_01657
3075	HAMAP	BioD; Dethiobiotin synthase BioD	3873	4571	699	MF_00336
3148	HAMAP	KDO8P_synth; 3-deoxy-8-phosphooctulonate synthase	21986	22780	795	MF_00056
3148	HAMAP	FtsB; Cell division protein FtsB	18467	18742	276	MF_00599
3148	HAMAP	UPF0745; YcgL domain	30092	30352	261	MF_01866
6160	HAMAP	Enolase; Enolase	8036	9292	1257	MF_00318
6160	HAMAP	dCTP_deaminase; Deoxycytidine triphosphate deaminase	18252	18815	564	MF_00146
9794	HAMAP	Ac_ald_DH_ac; Acetaldehyde dehydrogenase	16832	17767	936	MF_01657
18132	HAMAP	PyrG; CTP synthase	3554	5176	1623	MF_01227
243	Panther	NITRATE REDUCTASE ALPHA CHAIN; NITRATE, FROMATE, IRON DEHYDROGENASE; unintegrated	53536	55614	2079	PTHR11615; PTHR11615:SF35
243	Panther	OUTER MEMBRANE CATION EFFLUX PROTEIN; OUTER MEMBRANE CU/AG CATION EFFLUX PROTEIN; unintegrated	49404	50922	1519	PTHR30203; PTHR30203:SF0
243	Panther	MAJOR FACILITATOR SUPERFAMILY DOMAIN-CONTAINING PROTEIN-RELATED; SUBFAMILY NOT NAMED; unintegrated	45399	46649	1251	PTHR24003; PTHR24003:SF115
243	Panther	ADENOSYLMETHIONINE-8-AMINO-7-OXONONANOATE AMINOTRANSFERASE; AMINOTRANSFERASE CLASS III	36849	38081	1233	PTHR11986; PTHR11986:SF8
243	Panther	SPERMIDINE/PUTRESCINE-BINDING PERIPLASMIC PROTEIN; unintegrated	5662	6705	1044	PTHR30222
243	Panther	HTH-TYPE TRANSCRIPTIONAL REGULATOR LEUO-RELATED; unintegrated	24851	25750	900	PTHR30118
243	Panther	TRANSCRIPTIONAL DUAL REGULATOR HCAR-RELATED; unintegrated	32413	33282	870	PTHR30346
243	Panther	UNCHARACTERIZED; unintegrated	20393	21241	849	PTHR30582; PTHR30582:SF0
243	Panther	ALCOHOL DEHYDROGENASE RELATED; L-THREONINE 3-DEHYDROGENASE; Alcohol dehydrogenase superfamily, zinc-type; unintegrated	29419	30231	813	PTHR11695; PTHR11695:SF285
243	Panther	FAMILY NOT NAMED; unintegrated	11450	12133	684	PTHR31299
243	Panther	HYDROXYPYRUVATE ISOMERASE; unintegrated	22505	23173	669	PTHR12110
243	Panther	FMN-DEPENDENT NADH-AZOREDUCTASE; NAD(P)H OXIDOREDUCTASE-RELATED; NADH-azoreductase, FMN-dependent; unintegrated	13459	14049	591	PTHR10204; PTHR10204:SF4
243	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	9385	9969	585	PTHR24567; PTHR24567:SF13
243	Panther	ALDEHYDE DEHYDROGENASE-RELATED; unintegrated	7194	7664	471	PTHR11699

243	Panther	FERRICHRONE IRON RECEPTOR-RELATED; TRANSPOSASE-RELATED; unintegrated	58026	58460	435	PTHR32552; PTHR32552:SF1
243	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	26717	27148	432	PTHR24567; PTHR24567:SF13
243	Panther	AMMECR1 HOMOLOG; SUBFAMILY NOT NAMED; AMMECR1; unintegrated	35807	36226	420	PTHR13016; PTHR13016:SF0
243	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	5157	5561	405	PTHR32071; PTHR32071:SF5
243	Panther	TRANSLATION INITIATION INHIBITOR; YjgF/Yer057p/UK114 family	23158	23535	378	PTHR11803
243	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	48190	48531	342	PTHR32332; PTHR32332:SF0
243	Panther	FAMILY NOT NAMED; unintegrated	44090	44203	114	PTHR24559
3075	Panther	MONOOXYGENASE; MONOXYGENASE; unintegrated	741	1955	1215	PTHR13789; PTHR13789:SF11
3075	Panther	HTH-TYPE TRANSCRIPTIONAL REGULATOR LEUO-RELATED; unintegrated	13555	14454	900	PTHR30118
3075	Panther	HTH-TYPE TRANSCRIPTIONAL REGULATOR LEUO-RELATED; unintegrated	16583	17296	714	PTHR30118
3075	Panther	HYDROXYPYRUVATE ISOMERASE; unintegrated	9825	10472	648	PTHR12110
3075	Panther	DETHIOBIOTIN SYNTHETASE; SUBFAMILY NOT NAMED; unintegrated	3966	4571	606	PTHR21343; PTHR21343:SF0
3075	Panther	GENERAL SECRETION PATHWAY PROTEIN G; unintegrated	6028	6234	207	PTHR30093
3148	Panther	SENSOR HISTIDINE KINASE NARQ-RELATED; TWO COMPONENT SIGNAL TRANSDUCTION PROTEIN-RELATED; unintegrated	4995	6761	1767	PTHR24423; PTHR24423:SF119
3148	Panther	MULTIDRUG RESISTANCE PROTEIN MDTA; unintegrated	9401	10387	987	PTHR30469
3148	Panther	HTH-TYPE TRANSCRIPTIONAL REGULATOR; unintegrated	33523	34413	891	PTHR30537; PTHR30537:SF0
3148	Panther	2-DEHYDRO-3-DEOXYPHOSPHOCTONATE ALDOLASE; PHOSPHO-2-DEHYDRO-3-DEOXYHEPTONATE ALDOLASE; 3-deoxy-8-phosphooctulonate synthase; unintegrated	21989	22771	783	PTHR21057; PTHR21057:SF2
3148	Panther	ACID HYDRATASE; unintegrated	20253	21032	780	PTHR30143
3148	Panther	NITRATE REDUCTASE ALPHA CHAIN; NITRATE, FROMATE, IRON DEHYDROGENASE; unintegrated	17274	18041	768	PTHR11615; PTHR11615:SF35
3148	Panther	ALKANESULFONATE MONOOXYGENASE-RELATED; unintegrated	7093	7816	724	PTHR30011
3148	Panther	FAMILY NOT NAMED; PUTATIVE ABC TRANSPORTER ATP-BINDING SUBUNIT; unintegrated	35149	35868	720	PTHR24220; PTHR24220:SF157
3148	Panther	DIENELACTONE HYDROLASE; unintegrated	48599	49228	630	PTHR17630

3148	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	36446	37042	597	PTHR24220; PTHR24220:SF151
3148	Panther	GEPHYRIN; MOLYBDOPTERIN BIOSYNTHESIS PROTEIN; unintegrated	39257	39790	534	PTHR10192; PTHR10192:SF0
3148	Panther	RNA METHYLTRANSFERASE; unintegrated	44779	45168	390	PTHR12029
3148	Panther	PEPTIDYL-PROLYL CIS-TRANS ISOMERASE; PEPTIDYL-PROLYL CIS-TRANS ISOMERASE B, PPIB; unintegrated	32227	32523	297	PTHR11071; PTHR11071:SF29
3148	Panther	POTASSIUM/PROTON ANTIporter-RELATED; unintegrated	50107	50280	174	PTHR16254
5976	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	90	929	840	PTHR32371; PTHR32371:SF0
5976	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	1381	1698	318	PTHR24695; PTHR24695:SF77
6160	Panther	FAMILY NOT NAMED; unintegrated	13988	15718	1731	PTHR24221
6160	Panther	ENOLASE; Enolase	8009	9292	1284	PTHR11902
6160	Panther	ACID HYDRATASE; unintegrated	12854	13345	492	PTHR30143
9794	Panther	MRP-RELATED NUCLEOTIDE-BINDING PROTEIN; NUCLEOTIDE-BINDING PROTEIN NBP35(YEAST)-RELATED; unintegrated	5148	6221	1074	PTHR23264; PTHR23264:SF4
9794	Panther	ACETALDEHYDE DEHYDROGENASE; Acetaldehyde dehydrogenase	16832	17761	930	PTHR21123
9794	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	8566	9471	906	PTHR32459; PTHR32459:SF0
9794	Panther	PIRIN-RELATED; SUBFAMILY NOT NAMED; Pirin; unintegrated	9963	10817	855	PTHR13903; PTHR13903:SF0
9794	Panther	PROTEIN MEMO1; UPF0103/Mediator of ErbB2-driven cell motility (Memo-related); unintegrated	3768	4562	795	PTHR11060; PTHR11060:SF0
9794	Panther	FAMILY NOT NAMED; SUBFAMILY NOT NAMED; unintegrated	15967	16716	750	PTHR24316; PTHR24316:SF5
9794	Panther	RESPONSE REGULATOR OF TWO-COMPONENT SYSTEM; unintegrated	12799	13404	606	PTHR26402
9794	Panther	DIHYDROFOLATE REDUCTASE; unintegrated	1202	1675	474	PTHR11549
14785	Panther	ACYL-PROTEIN THIOESTERASE 1 (LYSOPHOSPHOLIPASE 1); LYSOPHOSPHOLIPASE-RELATED; unintegrated	3193	3813	621	PTHR10655; PTHR10655:SF6
14785	Panther	NADPH NITROREDUCTASE; unintegrated	1140	1679	540	PTHR23026
14785	Panther	PYRUVATE FORMATE-LYASE-ACTIVATING ENZYME; unintegrated	304	795	492	PTHR30352
18132	Panther	CTP SYNTHASE; SUBFAMILY NOT NAMED; CTP synthase; unintegrated	3548	5179	1632	PTHR11550; PTHR11550:SF0

18132	Panther	MAJOR FACILITATOR SUPERFAMILY DOMAIN-CONTAINING PROTEIN-RELATED; SUBFAMILY NOT NAMED; unintegrated	6754	7908	1155	PTHR24003; PTHR24003:SF36
18132	Panther	HALOACID DEHALOGENASE-LIKE HYDROLASE; unintegrated	8562	9104	543	PTHR12725
18132	Panther	GLUCOSE-6-PHOSPHATE ISOMERASE; SUBFAMILY NOT NAMED; Phosphoglucose isomerase (PGI); unintegrated	5982	6389	408	PTHR11469; PTHR11469:SF0
23284	Panther	ACID HYDRATASE; unintegrated	3651	4250	600	PTHR30143
58390	Panther	ALPHA/BETA HYDROLASE FOLD-CONTAINING PROTEIN; SUBFAMILY NOT NAMED; unintegrated	501	1250	750	PTHR10992; PTHR10992:SF465
243	Pfam	OEP; Outer membrane efflux protein	49599	50859	1192	PF02321
243	Pfam	MFS_1; Major facilitator superfamily	45492	46628	1137	PF07690
243	Pfam	Aminotran_3; Aminotransferase class-III	36906	37910	1005	PF00202
243	Pfam	Molybdopterin; Molybdopterin oxidoreductase	53536	54423	888	PF00384
243	Pfam	NMO; 2-nitropropane dioxygenase, NPD	48187	49056	870	PF03060
243	Pfam	SBP_bac_8; unintegrated	5758	6564	807	PF13416
243	Pfam	tRNA-synt_1c; Glutamyl/glutaminyl-tRNA synthetase, class Ib, catalytic domain	619	1287	669	PF00749
243	Pfam	OprD; Outer membrane porin, bacterial	38380	39021	642	PF03573
243	Pfam	LysR_substrate; LysR, substrate-binding	32641	33243	603	PF03466
243	Pfam	Flavodoxin_2; Flavodoxin-like fold	13462	14049	588	PF02525
243	Pfam	AP_endonuc_2; Xylose isomerase-like, TIM barrel domain	22565	23107	543	PF01261
243	Pfam	DEAD; DNA/RNA helicase, DEAD/DEAH box type, N-terminal	52549	53028	480	PF00270
243	Pfam	Pribosyltran; Phosphoribosyltransferase domain	11507	11983	477	PF00156
243	Pfam	AMMECR1; AMMECR1 domain	35756	36226	471	PF01871
243	Pfam	NAD_binding_9; unintegrated	42304	42756	453	PF13454
243	Pfam	Aldedh; Aldehyde dehydrogenase domain	7227	7664	438	PF00171
243	Pfam	YkuD; L,D-transpeptidase catalytic domain	20603	21022	420	PF03734
243	Pfam	OKR_DC_1_N; Orn/Lys/Arg decarboxylase, N-terminal	56106	56471	366	PF03709
243	Pfam	Ribonuc_L-PSP; YjgF/Yer057p/UK114 family	23179	23529	351	PF01042
243	Pfam	ADH_N; Alcohol dehydrogenase GroES-like	29824	30150	327	PF08240
243	Pfam	rve; Integrase, catalytic core	43880	44206	327	PF00665
243	Pfam	rve; Integrase, catalytic core	58029	58316	288	PF00665

243	Pfam	ADH_zinc_N; Alcohol dehydrogenase, C-terminal	29425	29706	282	PF00107
243	Pfam	cNMP_binding; Cyclic nucleotide-binding domain	9457	9729	273	PF00027
243	Pfam	cNMP_binding; Cyclic nucleotide-binding domain	26810	27073	264	PF00027
243	Pfam	Sulfate_tra_GLY; unintegrated	60368	60610	243	PF13792
243	Pfam	Helicase_C; Helicase, C-terminal	52111	52338	228	PF00271
243	Pfam	DUF167; Protein of unknown function DUF167	47503	47727	225	PF02594
243	Pfam	Usp; UspA	55713	55934	222	PF00582
243	Pfam	DUF772; Transposase InsH, N-terminal	25925	26134	210	PF05598
243	Pfam	HTH_21; HTH-like domain	43646	43825	180	PF13276
243	Pfam	HTH_1; Transcription regulator HTH, LysR	25553	25729	177	PF00126
243	Pfam	HTH_1; Transcription regulator HTH, LysR	24341	24508	168	PF00126
243	Pfam	HTH_1; Transcription regulator HTH, LysR	32413	32562	150	PF00126
243	Pfam	DUF4197; Protein of unknown function DUF4197	44481	44621	141	PF13852
243	Pfam	LysM; Peptidoglycan-binding lysin domain	21092	21220	129	PF01476
243	Pfam	HTH_8; DNA binding HTH domain, Fis-type	4857	4961	105	PF02954
243	Pfam	FliB; Uncharacterised protein family UPF0153	30758	30859	102	PF03692
3075	Pfam	Mem_trans; Auxin efflux carrier	2129	3013	849	PF03547
3075	Pfam	AAA_26; unintegrated	3960	4565	606	PF13500
3075	Pfam	LysR_substrate; LysR, substrate-binding	13840	14442	603	PF03466
3075	Pfam	LysR_substrate; LysR, substrate-binding	16586	17185	600	PF03466
3075	Pfam	NYN; NYN domain, limkain-b1-type	8802	9215	414	PF01936
3075	Pfam	Semialdehyde_dh; Semialdehyde dehydrogenase, NAD-binding	7281	7619	339	PF01118
3075	Pfam	AcetDehyd-dimer; Acetaldehyde dehydrogenase, C-terminal	6918	7244	327	PF09290
3075	Pfam	AP_endonuc_2; Xylose isomerase-like, TIM barrel domain	10113	10364	252	PF01261
3075	Pfam	TOBE_2; Transport-associated OB, type 2	7888	8097	210	PF08402
3075	Pfam	FAD_binding_3; Monooxygenase, FAD-binding	867	1067	201	PF01494
3075	Pfam	HTH_1; Transcription regulator HTH, LysR	13576	13752	177	PF00126
3075	Pfam	HNH; HNH endonuclease	11918	12064	147	PF01844

3075	Pfam	NAD_binding_8; unintegrated	1836	1943	108	PF13450
3075	Pfam	N_methyl_2; Prokaryotic N-terminal methylation site	6166	6231	66	PF13544
3148	Pfam	DUF879; Type VI secretion system, VCA0110	22632	23828	1197	PF05947
3148	Pfam	Peptidase_M1; Peptidase M1, membrane alanine aminopeptidase, N-terminal	27423	28493	1071	PF01433
3148	Pfam	MFS_1; Major facilitator superfamily	10804	11778	975	PF07690
3148	Pfam	HlyD_2; unintegrated	9389	10276	888	PF12700
3148	Pfam	DAHP_synth_1; DAHP synthetase I/KDSA	22004	22780	777	PF00793
3148	Pfam	Molybdopterin; Molybdopterin oxidoreductase	17271	18041	771	PF00384
3148	Pfam	Bac_luciferase; Luciferase-like domain	7156	7822	667	PF00296
3148	Pfam	FAA_hydrolase; Fumarylacetoacetase, C-terminal	20418	21032	615	PF01557
3148	Pfam	LysR_substrate; LysR, substrate-binding	33529	34143	615	PF03466
3148	Pfam	Abhydrolase_5; unintegrated	48671	49150	480	PF12695
3148	Pfam	MoCF_biosynth; Molybdopterin binding domain	39245	39664	420	PF00994
3148	Pfam	SpoU_methylase; tRNA/rRNA methyltransferase, SpoU	44779	45168	390	PF00588
3148	Pfam	ABC_tran; ABC transporter-like	36539	36898	360	PF00005
3148	Pfam	ABC_tran; ABC transporter-like	35392	35748	357	PF00005
3148	Pfam	GPW_gp25; Anti-sigma factor antagonist, IraD	25920	26249	330	PF04965
3148	Pfam	Phage-tail_3; unintegrated	24336	24662	327	PF13550
3148	Pfam	PilJ; unintegrated	5094	5396	303	PF13675
3148	Pfam	Pro_isomerase; Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain	32230	32526	297	PF00160
3148	Pfam	HATPase_c; Histidine kinase-like ATPase, ATP-binding domain	6492	6752	261	PF02518
3148	Pfam	DivIC; Septum formation initiator	18497	18733	237	PF04977
3148	Pfam	YcgL; YcgL domain	30095	30313	219	PF05166
3148	Pfam	HisKA_3; Signal transduction histidine kinase, subgroup 3, dimerisation and phosphoacceptor domain	6165	6365	201	PF07730
3148	Pfam	HAMP; HAMP linker domain	5481	5675	195	PF00672
3148	Pfam	HTH_1; Transcription regulator HTH, LysR	34225	34413	189	PF00126
3148	Pfam	HTH_1; Transcription regulator HTH, LysR	46412	46591	180	PF00126
3148	Pfam	SPOR; Sporulation-related domain	431	607	177	PF05036

3148	Pfam	Na_H_Exchange; Cation/H+ exchanger	50128	50280	153	PF00999
3148	Pfam	KdpD; Signal transduction histidine kinase, osmosensitive K+ channel sensor, N-terminal	40919	41023	105	PF02702
3148	Pfam	DDE_Tnp_IS66_C; unintegrated	42738	42815	78	PF13817
5976	Pfam	BPD_transp_1; Binding-protein-dependent transport systems inner membrane component	306	896	591	PF00528
5976	Pfam	DDE_Tnp_ISL3; Transposase IS204/IS1001/IS1096/IS1165	1746	2192	447	PF01610
5976	Pfam	AAA; ATPase, AAA-type, core	1204	1602	399	PF00004
6160	Pfam	Enolase_C; Enolase, C-terminal	8027	8875	849	PF00113
6160	Pfam	ABC_membrane; ABC transporter, transmembrane domain	14852	15670	819	PF00664
6160	Pfam	Glyoxalase; Glyoxalase/fosfomycin resistance/dioxygenase domain	3255	4028	681	PF00903
6160	Pfam	Glyoxalase; Glyoxalase/fosfomycin resistance/dioxygenase domain	4381	5154	681	PF00903
6160	Pfam	PPK2; Polyphosphate kinase-2-related	9679	10356	678	PF03976
6160	Pfam	Cupin_7; ChrR-like cupin domain	2316	2903	537	PF12973
6160	Pfam	FAA_hydrolase; Fumarylacetoacetase, C-terminal	12857	13330	474	PF01557
6160	Pfam	Enolase_N; Enolase, N-terminal	8891	9283	393	PF03952
6160	Pfam	ABC_tran; ABC transporter-like	14201	14578	378	PF00005
6160	Pfam	Glyoxalase; Glyoxalase/fosfomycin resistance/dioxygenase domain	12211	12543	333	PF00903
6160	Pfam	dUTPase; DeoxyUTP pyrophosphatase	18267	18572	306	PF00692
6160	Pfam	ABM; Antibiotic biosynthesis monooxygenase	6317	6484	168	PF03992
9794	Pfam	MFS_1; Major facilitator superfamily	14860	15699	840	PF07690
9794	Pfam	Memo; UPF0103/Mediator of ErbB2-driven cell motility (Memo-related)	3789	4544	756	PF01875
9794	Pfam	adh_short_C2; unintegrated	16003	16713	711	PF13561
9794	Pfam	LysR_substrate; LysR, substrate-binding	8833	9453	621	PF03466
9794	Pfam	OprD; Outer membrane porin, bacterial	6938	7441	504	PF03573
9794	Pfam	DHFR_1; Dihydrofolate reductase domain	1208	1675	468	PF00186
9794	Pfam	AcetDehyd-dimer; Acetaldehyde dehydrogenase, C-terminal	17225	17686	462	PF09290
9794	Pfam	Peptidase_M15_2; Protein of unknown function DUF882	2406	2849	444	PF05951
9794	Pfam	Semialdhyde_dh; Semialdehyde dehydrogenase, NAD-binding	16847	17197	351	PF01118
9794	Pfam	Response_reg; Signal transduction response regulator, receiver domain	13066	13401	336	PF00072

9794	Pfam	Pirin_C; Pirin, C-terminal domain	10500	10808	309	PF05726
9794	Pfam	Pirin; Pirin, N-terminal domain	10044	10337	294	PF02678
9794	Pfam	ParA; ATPase-like, ParA/MinD	5769	6011	243	PF10609
9794	Pfam	DUF59; Domain of unknown function DUF59	5157	5378	222	PF01883
9794	Pfam	DUF2892; Protein of unknown function DUF2892	1870	2052	183	PF11127
9794	Pfam	HTH_1; Transcription regulator HTH, LysR	8584	8757	174	PF00126
9794	Pfam	GerE; Transcription regulator LuxR, C-terminal	12805	12972	168	PF00196
9794	Pfam	AAA_31; AAA domain	5445	5603	159	PF13614
14785	Pfam	Abhydrolase_2; Phospholipase/carboxylesterase/thioesterase	3196	3804	609	PF02230
14785	Pfam	DUF2937; Protein of unknown function DUF2937	2113	2601	489	PF11157
14785	Pfam	Radical_SAM; Radical SAM	472	930	459	PF04055
14785	Pfam	Nitroreductase; Nitroreductase-like	1236	1673	438	PF00881
14785	Pfam	Fer4_12; unintegrated	646	951	306	PF13353
18132	Pfam	MFS_1; Major facilitator superfamily	6763	7812	1050	PF07690
18132	Pfam	CTP_synth_N; CTP synthase, N-terminal	4355	5173	819	PF06418
18132	Pfam	GATase; Glutamine amidotransferase type 1	3578	4282	705	PF00117
18132	Pfam	HAD_2; HAD-like domain	8565	9089	525	PF13419
18132	Pfam	PGI; Phosphoglucose isomerase (PGI)	5982	6383	402	PF00342
18132	Pfam	HTH_Tnp_1; Transposase IS3/IS911 family	1888	2133	246	PF01527
23284	Pfam	FAA_hydrolase; Fumarylacetoacetase, C-terminal	3669	4247	579	PF01557
33223	Pfam	zf-dskA_traR; Zinc finger, DksA/TraR C4-type	472	576	105	PF01258
58390	Pfam	Abhydrolase_6; unintegrated	501	1211	711	PF12697
243	PIR	4-aminobutyrate/lysine/ornithine transaminase; unintegrated	36777	37955	1179	PIRSF000521
3075	PIR	Dethiobiotin synthetase; Dethiobiotin synthase BioD	3870	4571	702	PIRSF006755
3148	PIR	Signal transduction histidine kinase, nitrate/nitrite-sensing; Signal transduction histidine kinase, nitrate/nitrite-sensing	4995	6794	1800	PIRSF003167
6160	PIR	Enolase; Enolase	8018	9292	1275	PIRSF001400
9794	PIR	Acetaldehyde dehydrogenase (acetylating); Acetaldehyde dehydrogenase	16832	17767	936	PIRSF015689
9794	PIR	Pirin; Pirin	9984	10820	837	PIRSF006232

9794	PIR	Dihydrofolate reductase; Dihydrofolate reductase	1202	1681	480	PIRSF000194
14785	PIR	Predicted pyruvate-formate lyase-activating enzyme; Pyruvate-formate lyase-activating enzyme, predicted	223	1194	972	PIRSF004869
14785	PIR	Uncharacterised conserved protein, UCP029393 type; Uncharacterised conserved protein UCP029393	2068	2601	534	PIRSF029393
243	PRINTS	SPERMDNBNDNG; Bacterial periplasmic spermidine/putrescine-binding protein	5857	6591	273	PR00909
243	PRINTS	TRNASYNTHCYS; Cysteinyl-tRNA synthetase/mycothiol ligase	31466	32122	189	PR00983
243	PRINTS	UNVRSLSTRESS; Universal stress protein A	55572	55934	165	PR01438
243	PRINTS	HTHLYSR; Transcription regulator HTH, LysR	24374	24472	105	PR00039
243	PRINTS	HTHLYSR; Transcription regulator HTH, LysR	25586	25684	105	PR00039
243	PRINTS	HTHLYSR; Transcription regulator HTH, LysR	32431	32529	105	PR00039
3075	PRINTS	RNGMNOXGNASE; Aromatic-ring hydroxylase-like	855	1952	267	PR00420
3075	PRINTS	HTHLYSR; Transcription regulator HTH, LysR	13621	13719	105	PR00039
3148	PRINTS	ALADIPTASE; Peptidase M1, membrane alanine aminopeptidase, N-terminal	27600	28049	168	PR00756
3148	PRINTS	CSAPPIMRASE; Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain	32269	32385	87	PR00153
3148	PRINTS	HTHLYSR; Transcription regulator HTH, LysR	34270	34368	105	PR00039
6160	PRINTS	ENOLASE; Enolase	8153	9184	273	PR00148
9794	PRINTS	GDHRDH; Glucose/ribitol dehydrogenase	15982	16665	318	PR00081
9794	PRINTS	SDRFAMILY; Short-chain dehydrogenase/reductase SDR	16201	16485	123	PR00080
9794	PRINTS	DHFR; Dihydrofolate reductase	1364	1642	141	PR00070
9794	PRINTS	HTHLUXR; Transcription regulator LuxR, C-terminal	12841	12969	135	PR00038
18132	PRINTS	HADHALOGNASE; Haloacid dehalogenase/epoxide hydrolase	8556	9035	150	PR00413
18132	PRINTS	G6PISOMERASE; Phosphoglucose isomerase (PGI)	6075	6212	144	PR00662
58390	PRINTS	EPOXYHYDRLASE; Epoxide hydrolase-like	579	1226	201	PR00412
58390	PRINTS	ABHYDROLASE; Alpha/beta hydrolase fold-1	579	1109	177	PR00111
3148	ProDom	Q4ZW60_PSEU2_Q4ZW60;; YcgL domain	30092	30331	240	PD030374
243	PROFILE	AMMECR1; AMMECR1 domain	35669	36235	567	PS51112
243	PROFILE	HELICASE_ATP_BIND_1; Helicase, superfamily 1/2, ATP-binding domain	52498	53016	519	PS51192
243	PROFILE	INTEGRASE; Integrase, catalytic core	58026	58457	432	PS50994
243	PROFILE	INTEGRASE; Integrase, catalytic core	43817	44212	396	PS50994

243	PROFILE	CNMP_BINDING_3; Cyclic nucleotide-binding domain	26750	27112	363	PS50042
243	PROFILE	CNMP_BINDING_3; Cyclic nucleotide-binding domain	9400	9714	315	PS50042
243	PROFILE	HTH_LYSR; Transcription regulator HTH, LysR	24323	24496	174	PS50931
243	PROFILE	HTH_LYSR; Transcription regulator HTH, LysR	25562	25735	174	PS50931
243	PROFILE	HTH_CRP_2; Helix-turn-helix motif, Crp-type	9805	9969	165	PS51063
243	PROFILE	HTH_LYSR; Transcription regulator HTH, LysR	32410	32553	144	PS50931
243	PROFILE	PROKAR_LIPOPROTEIN; unintegrated	49386	49460	75	PS51257
243	PROFILE	PROKAR_LIPOPROTEIN; unintegrated	27186	27251	66	PS51257
3075	PROFILE	HTH_LYSR; Transcription regulator HTH, LysR	13570	13743	174	PS50931
3148	PROFILE	ABC_TRANSPORTER_2; ABC transporter-like	35173	35865	693	PS50893
3148	PROFILE	ABC_TRANSPORTER_2; ABC transporter-like	36344	37033	690	PS50893
3148	PROFILE	HIS_KIN; Signal transduction histidine kinase, core	6174	6764	591	PS50109
3148	PROFILE	CSA_PPIASE_2; Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain	32242	32547	306	PS50072
3148	PROFILE	YCGL; YcgL domain	30092	30343	252	PS51648
3148	PROFILE	HTH_LYSR; Transcription regulator HTH, LysR	34237	34431	195	PS50931
3148	PROFILE	HTH_LYSR; Transcription regulator HTH, LysR	46406	46579	174	PS50931
3148	PROFILE	HAMP; HAMP linker domain	5529	5687	159	PS50885
5976	PROFILE	ABC_TM1; Binding-protein-dependent transport systems inner membrane component	261	896	636	PS50928
6160	PROFILE	ABC_TM1F; ABC transporter, integral membrane type 1	14813	15670	858	PS50929
6160	PROFILE	PROKAR_LIPOPROTEIN; unintegrated	16611	16676	66	PS51257
9794	PROFILE	DHFR_2; Dihydrofolate reductase domain	1208	1678	471	PS51330
9794	PROFILE	RESPONSE_REGULATORY; Signal transduction response regulator, receiver domain	13054	13404	351	PS50110
9794	PROFILE	HTH_LUXR_2; Transcription regulator LuxR, C-terminal	12793	12990	198	PS50043
9794	PROFILE	HTH_LYSR; Transcription regulator HTH, LysR	8575	8748	174	PS50931
9794	PROFILE	TAT; Twin-arginine translocation pathway, signal sequence	13763	13912	150	PS51318
18132	PROFILE	MFS; Major facilitator superfamily domain	6754	7929	1176	PS50850
18132	PROFILE	GATASE_TYPE_1; Glutamine amidotransferase type 1	3548	4312	765	PS51273
18132	PROFILE	P_GLUCCOSE_ISOMERASE_3; Phosphoglucose isomerase (PGI)	5961	6410	450	PS51463

23284	PROFILE	PROKAR_LIPOPROTEIN; unintegrated	530	634	105	PS51257
33223	PROFILE	ZF_DKSA_2; Zinc finger, DksA/TraR C4-type	466	696	231	PS51128
243	PROSITE	AA_TRANSFER_CLASS_3; Aminotransferase class-III	37518	37631	114	PS00600
243	PROSITE	CYS_SYNTHASE; Cysteine synthase/cystathionine beta-synthase P-phosphate-binding site	16426	16482	57	PS00901
243	PROSITE	SBP_BACTERIAL_1; Bacterial extracellular solute-binding family 1, conserved site	6235	6288	54	PS01037
243	PROSITE	MOLYBDOPTERIN_PROK_2; Molybdopterin oxidoreductase, prokaryotic, conserved site	54253	54306	54	PS00490
243	PROSITE	ADH_ZINC; Alcohol dehydrogenase, zinc-type, conserved site	30001	30045	45	PS00059
243	PROSITE	OKR_DC_1; Orn/Lys/Arg decarboxylase, major domain	57207	57251	45	PS00703
243	PROSITE	MIP; Major intrinsic protein, conserved site	27369	27395	27	PS00221
3075	PROSITE	PROKAR_NTER_METHYL; Prokaryotic N-terminal methylation site	6157	6219	63	PS00409
3148	PROSITE	CSA_PPIASE_1; Cyclophilin-type peptidyl-prolyl cis-trans isomerase, conserved site	32332	32385	54	PS00170
3148	PROSITE	MOCF_BIOSYNTHESIS_1; Molybdenum cofactor biosynthesis, conserved site	39428	39469	42	PS01078
6160	PROSITE	EXTRADIOL_DIOXYGENAS; Exradiol ring-cleavage dioxygenase, class I /II	3258	3323	66	PS00082
6160	PROSITE	ENOLASE; Enolase, conserved site	8252	8293	42	PS00164
9794	PROSITE	HTH_LUXR_1; Transcription regulator LuxR, C-terminal	12844	12927	84	PS00622
9794	PROSITE	SUGAR_TRANSPORT_2; Sugar transporter, conserved site	15052	15129	78	PS00217
9794	PROSITE	DHFR_1; Dihydrofolate reductase conserved site	1574	1642	69	PS00075
9794	PROSITE	MRP; Mrp, conserved site	5760	5810	51	PS01215
18132	PROSITE	P_GLUCCOSE_ISOMERASE_2; Phosphoglucose isomerase, conserved site	6063	6116	54	PS00174
23284	PROSITE	2FE2S_FER_1; 2Fe-2S ferredoxin, iron-sulphur binding site	506	532	27	PS00197
33223	PROSITE	ZF_DKSA_1; Zinc finger, DksA/TraR C4-type conserved site	487	561	75	PS01102
3148	sig_peptide	signal peptide; MPIRNMRIGLRASLSFAVLASLLVLVGLFGLGQMATLRESA	18467	18574	108	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MSPVDIVRLLSLAAIWGASFLFMRIAPVIGS	20253	20348	96	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MRNAFVRRTSRLFLGCTLIAAGALPALAHA	29023	29112	90	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MLLSYLRLVLFAGLLIGVQVPGFINDYA	40961	41047	87	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MSLFKRSVTELLGTFWLVLGGCGSAVLA	7066	7149	84	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MLKIALVAGSVLFAANLMAATPAKA	4346	4423	78	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MRLASTKTAALCGGLLLAMSVASA	9188	9265	78	SignalP 3.0 HMM

3148	sig_peptide	signal peptide; MARSSASLQLPGAQAQPAAG	28584	28643	60	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MKKFCCVVLAMLPLTAFA	2496	2549	54	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MAVGLRLALAVGLLALSSSVWA	49178	49192	15	SignalP 3.0 HMM
3148	sig_peptide	signal peptide; MTFTIAAKASLLLLFLGSTLYVHL	35776	35787	12	SignalP 3.0 HMM
9794	sig_peptide	signal peptide; MSNNASFTPPSLVLSTIGLSLATFMQVLDTTIA	5139	5237	99	SignalP 3.0 HMM
9794	sig_peptide	signal peptide; MNIKTLRSGLSMVLAMTMAGCASYSGLTTEG	2283	2378	96	SignalP 3.0 HMM
9794	sig_peptide	signal peptide; MILLISDLHLEERPDITRAFLDLLAGRARA	7430	7522	93	SignalP 3.0 HMM
9794	sig_peptide	signal peptide; MATADTTPTAENTPANPDSGKRKFMLLALAVVVALSGAGVWA	3822	3893	72	SignalP 3.0 HMM
9794	sig_peptide	signal peptide; MSMRKLVPLLIYLFLLVPIYWLL	16832	16903	72	SignalP 3.0 HMM
9794	sig_peptide	signal peptide; MNKVQNNKAWWLVPVLLVAFS	15958	16026	69	SignalP 3.0 HMM
9794	sig_peptide	signal peptide; MLAMVVND DAGSLTPRGALRFFASMLAPTAVA	75	80	6	SignalP 3.0 HMM
14785	sig_peptide	signal peptide; MTKRLLILLTSGLLLSANAWA	2099	2110	12	SignalP 3.0 HMM
243	Site	other; interface (dimer of trimers) [polypeptide binding]	20054	21334	147	CDD:48344
243	Site	other; dimer interface [polypeptide binding]	56070	57275	111	CDD:239429
243	Site	active;	58053	58937	30	CDD:173907
243	Site	metal binding site [ion binding]; metal binding site [ion binding]	56172	56984	12	CDD:239429
243	Site	other; FMN binding site [chemical binding]	11495	12220	45	CDD:73392
243	Site	other; substrate binding pocket [chemical binding]	56511	57212	21	CDD:239429
243	Site	active;	53575	54201	33	CDD:48377
243	Site	active; putative ligand binding pocket/active site [active]	21724	22335	21	CDD:153373
243	Site	other; dimerization interface [polypeptide binding]	25823	26404	48	CDD:58645
243	Site	active;	25838	26416	45	CDD:58645
243	Site	other; UTP binding site [chemical binding]	53650	54201	36	CDD:48377
243	Site	other; substrate binding pocket [chemical binding]	31355	31888	21	CDD:176148
243	Site	FeS/SAM binding site; FeS/SAM binding site	23449	23961	48	CDD:100105
243	Site	active;	54580	55092	45	CDD:153217
243	Site	putative effector binding pocket; putative effector binding pocket	41974	42477	24	CDD:176114
243	Site	other; ATP binding site [chemical binding]	46494	46967	27	CDD:72971

243	Site	inhibitor-cofactor binding pocket; inhibitor-cofactor binding pocket	49752	50225	33	CDD:99735
243	Site	other; pyridoxal 5'-phosphate binding site [chemical binding]	49755	50225	24	CDD:99735
243	Site	other; pyridoxal 5'-phosphate binding pocket [chemical binding]	29056	29523	24	CDD:99742
243	Site	other; dimerization interface [polypeptide binding]	31418	31885	87	CDD:176148
243	Site	catalytic triad [active]; catalytic triad [active]	54667	55092	9	CDD:153217
243	Site	other; dimerization interface [polypeptide binding]	42079	42474	72	CDD:176114
243	Site	Catalytic site [active]; Catalytic site [active]	53572	53958	45	CDD:48377
243	Site	Substrate-binding/catalytic site; Substrate-binding/catalytic site	20363	20695	24	CDD:48344
243	Site	other; dimer interface [polypeptide binding]	59775	60098	54	CDD:48402
243	Site	active;	25262	25579	30	CDD:206754
243	Site	other; homotrimer interaction site [polypeptide binding]	16213	16527	78	CDD:100004
243	Site	other; putative metal binding site [ion binding]	99	401	18	CDD:199895
243	Site	putative active site [active]; putative active site [active]	16225	16527	15	CDD:100004
243	Site	other; trimer interface [polypeptide binding]	17838	18116	42	CDD:58167
243	Site	other; folate binding site [chemical binding]	27517	27795	18	CDD:238127
243	Site	other; tRNA binding surface [nucleotide binding]	59160	59429	48	CDD:153411
243	Site	Zn-binding sites [ion binding]; Zn-binding sites [ion binding]	20441	20695	15	CDD:48344
243	Site	anticodon binding site; anticodon binding site	59181	59429	33	CDD:153411
243	Site	generic binding surface II; generic binding surface II	15723	15959	15	CDD:72960
243	Site	other; trimer interface [polypeptide binding]	30461	30682	81	CDD:143638
243	Site	MPT binding site; MPT binding site	17832	18032	24	CDD:58167
243	Site	other; nucleotide binding region [chemical binding]	14469	14654	27	CDD:28960
243	Site	ssDNA binding site; ssDNA binding site	15744	15926	15	CDD:72960
243	Site	putative active site [active]; putative active site [active]	3261	3434	27	CDD:238075
243	Site	other; NADP+ binding site [chemical binding]	27583	27750	24	CDD:238127
243	Site	other; putative tRNA-binding site [nucleotide binding]	59850	60008	18	CDD:48402
243	Site	other; nucleotide binding region [chemical binding]	10105	10260	27	CDD:28960
243	Site	heme pocket [chemical binding]; heme pocket [chemical binding]	3234	3374	21	CDD:238075

243	Site	metal binding site [ion binding]; metal binding site [ion binding]	2889	3020	6	CDD:143635
243	Site	metal binding site [ion binding]; metal binding site [ion binding]	19219	19350	6	CDD:143635
243	Site	active;	2886	3005	30	CDD:143635
243	Site	active;	19216	19335	30	CDD:143635
243	Site	other; putative metal binding site [ion binding]	21829	21933	6	CDD:153373
243	Site	I-site; I-site	2826	2915	6	CDD:143635
243	Site	other; ATP-binding site [chemical binding]	14364	14453	12	CDD:28960
243	Site	other; ATP-binding site [chemical binding]	10276	10362	12	CDD:28960
243	Site	I-site; I-site	19159	19245	6	CDD:143635
243	Site	putative oxyanion hole; putative oxyanion hole	54586	54672	6	CDD:153217
243	Site	active;	30524	30598	27	CDD:143638
243	Site	Zn binding site [ion binding]; Zn binding site [ion binding]	40733	40804	9	CDD:189001
243	Site	other; dimer interface [polypeptide binding]	26723	26764	12	CDD:88411
243	Site	ABC transporter signature motif; ABC transporter signature motif	46797	46826	30	CDD:72971
243	Site	Walker A/P-loop; Walker A/P-loop	46485	46508	24	CDD:72971
243	Site	Walker A motif; Walker A motif	52774	52797	24	CDD:73300
243	Site	H-loop/switch region; H-loop/switch region	46953	46973	21	CDD:72971
243	Site	Walker B; Walker B	46857	46874	18	CDD:72971
243	Site	other; ATP binding site [chemical binding]	9511	9525	15	CDD:28927
243	Site	other; ATP binding site [chemical binding]	15201	15215	15	CDD:28927
243	Site	KMSKS motif; KMSKS motif	59028	59042	15	CDD:173907
243	Site	other; putative Mg ⁺⁺ binding site [ion binding]	9811	9822	12	CDD:28927
243	Site	other; putative Mg ⁺⁺ binding site [ion binding]	14910	14921	12	CDD:28927
243	Site	Q-loop/lid; Q-loop/lid	46620	46631	12	CDD:72971
243	Site	D-loop; D-loop	46881	46892	12	CDD:72971
243	Site	HIGH motif; HIGH motif	58080	58091	12	CDD:173907
243	Site	other; substrate binding site [chemical binding]	12038	12043	6	CDD:73392
243	Site	active; putative catalytic residue [active]	12041	12043	3	CDD:73392

243	Site	catalytic residue [active]; catalytic residue [active]	29056	29058	3	CDD:99742
243	Site	catalytic residue [active]; catalytic residue [active]	50223	50225	3	CDD:99735
3075	Site	other; putative alpha subunit interface [polypeptide binding]	11155	11736	81	CDD:176889
3075	Site	other; FAD binding pocket [chemical binding]	8155	8733	48	CDD:99784
3075	Site	other; substrate binding pocket [chemical binding]	6813	7322	21	CDD:176148
3075	Site	putative active site [active]; putative active site [active]	11182	11637	60	CDD:176889
3075	Site	other; putative substrate binding site [chemical binding]	11182	11628	57	CDD:176889
3075	Site	other; dimerization interface [polypeptide binding]	6876	7319	87	CDD:176148
3075	Site	Fe binding site [ion binding]; Fe binding site [ion binding]	11200	11637	9	CDD:176889
3075	Site	other; dimer interface [polypeptide binding]	14962	15387	171	CDD:206779
3075	Site	inter-subunit interface; inter-subunit interface	11936	12265	54	CDD:29629
3075	Site	other; NAD binding pocket [chemical binding]	8383	8655	21	CDD:99784
3075	Site	other; N-terminal domain interface [polypeptide binding]	13103	13354	21	CDD:198300
3075	Site	other; maleylacetoacetate (MAA) substrate binding site (H site) [chemical binding]	13136	13354	15	CDD:198300
3075	Site	other; C-terminal domain interface [polypeptide binding]	12839	13036	21	CDD:48591
3075	Site	other; GSH binding site (G-site) [chemical binding]	12845	13021	24	CDD:48591
3075	Site	other; dimer interface [polypeptide binding]	13091	13258	63	CDD:198300
3075	Site	other; dimerization interface [polypeptide binding]	14611	14736	51	CDD:100122
3075	Site	catalytic loop [active]; catalytic loop [active]	7861	7983	30	CDD:29262
3075	Site	iron binding site [ion binding]; iron binding site [ion binding]	7873	7983	12	CDD:29262
3075	Site	other; putative CheW interface [polypeptide binding]	15109	15210	102	CDD:206779
3075	Site	iron-sulfur cluster [ion binding]; iron-sulfur cluster [ion binding]	10801	10884	24	CDD:58538
3075	Site	other; [2Fe-2S] cluster binding site [ion binding]	10801	10878	18	CDD:58538
3075	Site	other; phosphate binding motif [ion binding]	8257	8328	18	CDD:99784
3075	Site	other; [2Fe-2S] cluster binding site [ion binding]	12414	12482	12	CDD:58551
3075	Site	other; putative dimer interface [polypeptide binding]	12977	13036	18	CDD:48591
3075	Site	beta-alpha-beta structure motif; beta-alpha-beta structure motif	8368	8397	18	CDD:99784
3075	Site	other; FAD binding motif [chemical binding]	8188	8199	9	CDD:99784

3148	Site	other; dimer interface [polypeptide binding]	21320	22633	81	CDD:143447
3148	Site	other; dimer interface [polypeptide binding]	36119	36994	75	CDD:107204
3148	Site	other; pyridoxal 5'-phosphate binding site [chemical binding]	36137	36913	36	CDD:107204
3148	Site	other; dimer interface [polypeptide binding]	10738	11448	75	CDD:29417
3148	Site	active;	10831	11358	9	CDD:29417
3148	Site	other; ATP binding site [chemical binding]	6186	6674	27	CDD:73010
3148	Site	amphipathic channel; amphipathic channel	7195	7632	24	CDD:29423
3148	Site	catalytic residues [active]; catalytic residues [active]	21899	22324	12	CDD:143447
3148	Site	other; dimerization interface [polypeptide binding]	32578	32967	69	CDD:176102
3148	Site	putative active site [active]; putative active site [active]	41452	41832	18	CDD:48481
3148	Site	Asn-Pro-Ala signature motifs; Asn-Pro-Ala signature motifs	7255	7629	18	CDD:29423
3148	Site	other; dimerization interface [polypeptide binding]	2932	3294	69	CDD:176102
3148	Site	other; ATP binding site [chemical binding]	38495	38848	9	CDD:99707
3148	Site	other; putative dimer interface [polypeptide binding]	41413	41697	15	CDD:48481
3148	Site	other; NADP binding site [chemical binding]	22070	22327	24	CDD:143447
3148	Site	active;	4553	4750	33	CDD:29390
3148	Site	other; dimer interface [polypeptide binding]	302	460	36	CDD:119399
3148	Site	other; nucleotide binding region [chemical binding]	15043	15201	27	CDD:28960
3148	Site	other; ATP-binding site [chemical binding]	14887	15027	12	CDD:28960
3148	Site	other; ATP binding site [chemical binding]	2	103	12	CDD:28956
3148	Site	active;	46529	46621	33	CDD:28969
3148	Site	other; ligand binding site [chemical binding]	45043	45081	15	CDD:28920
3148	Site	ABC transporter signature motif; ABC transporter signature motif	6504	6533	30	CDD:73010
3148	Site	flexible hinge region; flexible hinge region	45139	45165	18	CDD:28920
3148	Site	Walker A/P-loop; Walker A/P-loop	6177	6200	24	CDD:73010
3148	Site	H-loop/switch region; H-loop/switch region	6660	6680	21	CDD:73010
3148	Site	Walker B; Walker B	6564	6581	18	CDD:73010
3148	Site	Walker B motif; Walker B motif	38705	38722	18	CDD:99707

3148	Site	other; ATP binding site [chemical binding]	15808	15822	15	CDD:28927
3148	Site	Q-loop/lid; Q-loop/lid	6312	6323	12	CDD:73010
3148	Site	D-loop; D-loop	6588	6599	12	CDD:73010
3148	Site	other; putative Mg ⁺⁺ binding site [ion binding]	15529	15540	12	CDD:28927
3148	Site	phosphorylation; phosphorylation site [posttranslational modification]	440	442	3	CDD:119399
3148	Site	Mg ²⁺ binding site [ion binding]; Mg ²⁺ binding site [ion binding]	89	91	3	CDD:28956
3148	Site	other; CoA binding pocket [chemical binding]	10921	10923	3	CDD:29417
3148	Site	catalytic residue [active]; catalytic residue [active]	36911	36913	3	CDD:107204
3148	Site	arginine finger; arginine finger	38888	38890	3	CDD:99707
5976	Site	other; ATP-binding site [chemical binding]	7925	8011	12	CDD:28960
6160	Site	other; NAD binding site [chemical binding]	14351	15265	63	CDD:143412
6160	Site	active;	11026	11583	27	CDD:163681
6160	Site	metal binding site [ion binding]; metal binding site [ion binding]	11026	11580	9	CDD:163681
6160	Site	other; NAD(P) binding site [chemical binding]	4396	4935	78	CDD:212491
6160	Site	other; dimer interface [polypeptide binding]	7895	8395	207	CDD:206779
6160	Site	other; substrate binding pocket [chemical binding]	3300	3746	21	CDD:176123
6160	Site	other; tetramer interface [polypeptide binding]	15819	16229	81	CDD:176667
6160	Site	catalytic residues [active]; catalytic residues [active]	14855	15256	12	CDD:143412
6160	Site	other; tetramer interface [polypeptide binding]	16302	16661	18	CDD:176686
6160	Site	active;	15882	16220	33	CDD:176667
6160	Site	Fe binding site [ion binding]; Fe binding site [ion binding]	15882	16220	15	CDD:176667
6160	Site	putative active site [active]; putative active site [active]	24	335	15	CDD:197380
6160	Site	other; Ligand Binding Site [chemical binding]	1041	1331	27	CDD:30165
6160	Site	other; hexamer interface [polypeptide binding]	9583	9741	75	CDD:29603
6160	Site	active;	4696	4839	12	CDD:212491
6160	Site	other; dimerization interface [polypeptide binding]	8558	8695	57	CDD:100122
6160	Site	active site 2 [active]; active site 2 [active]	9601	9735	6	CDD:29603
6160	Site	active site 1 [active]; active site 1 [active]	9640	9750	9	CDD:29603

6160	Site	other; dimerization interface [polypeptide binding]	3660	3764	99	CDD:176123
6160	Site	other; putative CheW interface [polypeptide binding]	8084	8185	102	CDD:206779
6160	Site	catalytic residues [active]; catalytic residues [active]	11488	11583	9	CDD:163681
6160	Site	other; dimer interface [polypeptide binding]	9658	9750	39	CDD:29603
9794	Site	putative active site [active]; putative active site [active]	6935	7504	24	CDD:163641
9794	Site	other; putative metal binding site [ion binding]	6935	7504	21	CDD:163641
9794	Site	other; dimer interface [polypeptide binding]	17066	17569	159	CDD:119394
9794	Site	other; ATP binding site [chemical binding]	14989	15453	27	CDD:73018
9794	Site	other; ATP binding site [chemical binding]	13874	14335	27	CDD:73018
9794	Site	other; ATP binding site [chemical binding]	12049	12390	30	CDD:99707
9794	Site	putative PBP binding loops; putative PBP binding loops	17186	17524	18	CDD:119394
9794	Site	other; dimer interface [polypeptide binding]	16435	16770	96	CDD:119394
9794	Site	conserved gate region; conserved gate region	17144	17449	63	CDD:119394
9794	Site	putative PBP binding loops; putative PBP binding loops	16420	16722	15	CDD:119394
9794	Site	other; substrate binding site [chemical binding]	7723	7872	33	CDD:29391
9794	Site	ABC-ATPase subunit interface; ABC-ATPase subunit interface	16498	16569	48	CDD:119394
9794	Site	ABC-ATPase subunit interface; ABC-ATPase subunit interface	17309	17380	48	CDD:119394
9794	Site	ABC transporter signature motif; ABC transporter signature motif	14159	14188	30	CDD:73018
9794	Site	ABC transporter signature motif; ABC transporter signature motif	15274	15303	30	CDD:73018
9794	Site	Walker A motif; Walker A motif	12370	12393	24	CDD:99707
9794	Site	Walker A/P-loop; Walker A/P-loop	13865	13888	24	CDD:73018
9794	Site	Walker A/P-loop; Walker A/P-loop	14980	15003	24	CDD:73018
9794	Site	H-loop/switch region; H-loop/switch region	14321	14341	21	CDD:73018
9794	Site	Walker B motif; Walker B motif	12172	12189	18	CDD:99707
9794	Site	DNA binding; DNA-binding interface [nucleotide binding]	11587	11604	15	CDD:119388
9794	Site	Walker B; Walker B	14219	14236	18	CDD:73018
9794	Site	Walker B; Walker B	15337	15354	18	CDD:73018
9794	Site	KMSKS motif; KMSKS motif	9079	9093	15	CDD:185676

9794	Site	H-loop/switch region; H-loop/switch region	15439	15453	15	CDD:73018
9794	Site	Q-loop/lid; Q-loop/lid	13994	14005	12	CDD:73018
9794	Site	D-loop; D-loop	14243	14254	12	CDD:73018
9794	Site	Q-loop/lid; Q-loop/lid	15109	15120	12	CDD:73018
9794	Site	D-loop; D-loop	15361	15372	12	CDD:73018
9794	Site	arginine finger; arginine finger	11992	11994	3	CDD:99707
14785	Site	other; substrate binding pocket [chemical binding]	4847	5356	21	CDD:176148
14785	Site	other; dimerization interface [polypeptide binding]	4910	5353	87	CDD:176148
14785	Site	active;	3541	3573	12	CDD:58567
18132	Site	other; putative substrate translocation pore	5670	6374	111	CDD:119392
18132	Site	other; molybdopterin cofactor binding site	85	471	39	CDD:30308
18132	Site	active;	8589	8882	24	CDD:29071
18132	Site	other; ATP binding site [chemical binding]	8275	8499	42	CDD:28956
18132	Site	DNA binding; DNA binding residues [nucleotide binding]	9012	9152	48	CDD:99777
18132	Site	other; dimerization interface [polypeptide binding]	7294	7431	57	CDD:100122
18132	Site	other; dimerization interface [polypeptide binding]	9045	9179	27	CDD:99777
18132	Site	G-X-G motif; G-X-G motif	8374	8418	9	CDD:28956
18132	Site	other; ligand binding site [chemical binding]	9539	9580	15	CDD:28920
18132	Site	other; non-specific DNA interactions [nucleotide binding]	9809	9850	15	CDD:28976
18132	Site	flexible hinge region; flexible hinge region	9638	9664	18	CDD:28920
18132	Site	DNA binding; DNA binding site [nucleotide binding]	9848	9868	21	CDD:28976
18132	Site	intermolecular recognition site; intermolecular recognition site	8736	8753	15	CDD:29071
18132	Site	other; sequence specific DNA binding site [nucleotide binding]	9851	9868	9	CDD:28976
18132	Site	other; dimerization interface [polypeptide binding]	8877	8885	9	CDD:29071
18132	Site	putative switch regulator; putative switch regulator	9752	9757	6	CDD:28976
18132	Site	other; putative cAMP binding site [chemical binding]	9851	9856	6	CDD:28976
18132	Site	Mg2+ binding site [ion binding]; Mg2+ binding site [ion binding]	8287	8289	3	CDD:28956
18132	Site	phosphorylation; phosphorylation site [posttranslational modification]	8727	8729	3	CDD:29071

23284	Site	other; substrate binding pocket [chemical binding]	3960	4469	21	CDD:176148
23284	Site	other; dimerization interface [polypeptide binding]	3963	4406	87	CDD:176148
23284	Site	other; tetramer interface [polypeptide binding]	1044	1403	18	CDD:176686
23284	Site	other; tetramer interface [polypeptide binding]	900	971	6	CDD:176667
23284	Site	other; NAD binding site [chemical binding]	10	18	9	CDD:143412
23284	Site	active;	954	962	6	CDD:176667
23284	Site	catalytic residues [active]; catalytic residues [active]	7	9	3	CDD:143412
23284	Site	Fe binding site [ion binding]; Fe binding site [ion binding]	960	962	3	CDD:176667
33223	Site	other; NAD(P) binding site [chemical binding]	28	705	60	CDD:176178
58390	Site	other; dimerization interface [polypeptide binding]	660	1040	69	CDD:176102
66283	Site	other; putative DNA binding site [nucleotide binding]	269	388	48	CDD:28974
66283	Site	other; putative Zn ²⁺ binding site [ion binding]	332	343	6	CDD:28974
243	SMART	DEAD-like helicases superfamily; Helicase, superfamily 1/2, ATP-binding domain	52456	53052	597	SM00487
243	SMART	Cyclic nucleotide-monophosphate binding doma; Cyclic nucleotide-binding domain	26750	27115	366	SM00100
243	SMART	Cyclic nucleotide-monophosphate binding doma; Cyclic nucleotide-binding domain	9400	9759	360	SM00100
243	SMART	helicase superfamily c-terminal domain; Helicase, C-terminal	52111	52356	246	SM00490
243	SMART	Lysin motif; Peptidoglycan-binding Lysin subgroup	21089	21223	135	SM00257
3075	SMART	Semialdehyde dehydrogenase, NAD bindin; Semialdehyde dehydrogenase, NAD-binding	7272	7619	348	SM00859
3075	SMART	HNH nucleases; HNH nuclease	11879	12049	171	SM00507
3148	SMART	ATPases associated with a variety of cellula; AAA+ ATPase domain	35218	35793	576	SM00382
3148	SMART	ATPases associated with a variety of cellula; AAA+ ATPase domain	36368	36943	576	SM00382
3148	SMART	Probable molybdopterin binding domain; Molybdopterin binding domain	39242	39682	441	SM00852
3148	SMART	Histidine kinase-like ATPases; Histidine kinase-like ATPase, ATP-binding domain	6483	6764	282	SM00387
3148	SMART	HAMP (Histidine kinases, Adenylyl cyclases,; HAMP linker domain	5529	5687	159	SM00304
9794	SMART	Semialdehyde dehydrogenase, NAD bindin; Semialdehyde dehydrogenase, NAD-binding	16847	17203	357	SM00859
9794	SMART	cheY-homologous receiver domain; Signal transduction response regulator, receiver domain	13066	13407	342	SM00448
9794	SMART	helix_turn_helix, Lux Regulon; Transcription regulator LuxR, C-terminal	12805	12978	174	SM00421
243	Superfamily	Formate dehydrogenase/DMSO reductase, domains 1-3; unintegrated	53536	55134	1599	SSF53706

243	Superfamily	Outer membrane efflux proteins (OEP); unintegrated	49386	50868	1483	SSF56954
243	Superfamily	MFS general substrate transporter; Major facilitator superfamily domain, general substrate transporter	45387	46670	1284	SSF103473
243	Superfamily	PLP-dependent transferases; Pyridoxal phosphate-dependent transferase	36813	38069	1257	SSF53383
243	Superfamily	Periplasmic binding protein-like II; unintegrated	5665	6609	945	SSF53850
243	Superfamily	Inosine monophosphate dehydrogenase (IMPDH); unintegrated	48184	49059	876	SSF51412
243	Superfamily	Xylose isomerase-like; Xylose isomerase-like, TIM barrel domain	22499	23275	777	SSF51658
243	Superfamily	Nucleotidyl transferase; unintegrated	616	1341	726	SSF52374
243	Superfamily	Adenine nucleotide alpha hydrolases-like; unintegrated	55287	55934	639	SSF52402
243	Superfamily	PRTase-like; unintegrated	11510	12133	624	SSF53271
243	Superfamily	Periplasmic binding protein-like II; unintegrated	32635	33255	621	SSF53850
243	Superfamily	Flavoproteins; unintegrated	13453	14049	597	SSF52218
243	Superfamily	GroES-like; GroES-like	29707	30231	525	SSF50129
243	Superfamily	ALDH-like; Aldehyde/histidinol dehydrogenase	7194	7664	471	SSF53720
243	Superfamily	Ribonuclease H-like; Ribonuclease H-like domain	58029	58484	456	SSF53098
243	Superfamily	AMMECR1-like; unintegrated	35780	36226	447	SSF143447
243	Superfamily	L,D-transpeptidase catalytic domain-like; unintegrated	20600	21040	441	SSF141523
243	Superfamily	cAMP-binding domain-like; Cyclic nucleotide-binding-like	9379	9807	429	SSF51206
243	Superfamily	NAD(P)-binding Rossmann-fold domains; unintegrated	29416	29817	402	SSF51735
243	Superfamily	cAMP-binding domain-like; Cyclic nucleotide-binding-like	26726	27118	393	SSF51206
243	Superfamily	YjgF-like; Endoribonuclease L-PSP/chorismate mutase-like	23164	23535	372	SSF55298
243	Superfamily	Ribonuclease H-like; Ribonuclease H-like domain	43847	44206	360	SSF53098
243	Superfamily	"Winged helix" DNA-binding domain; unintegrated	24323	24658	336	SSF46785
243	Superfamily	YggU-like; Protein of unknown function DUF167	47464	47760	297	SSF69786
243	Superfamily	"Winged helix" DNA-binding domain; unintegrated	25484	25750	267	SSF46785
243	Superfamily	"Winged helix" DNA-binding domain; unintegrated	32410	32646	237	SSF46785
243	Superfamily	"Winged helix" DNA-binding domain; unintegrated	9799	9969	171	SSF46785
243	Superfamily	LysM domain; unintegrated	21083	21229	147	SSF54106
243	Superfamily	Homeodomain-like; Homeodomain-like	4845	4943	99	SSF46689

3075	Superfamily	FAD/NAD(P)-binding domain; unintegrated	702	1955	1254	SSF51905
3075	Superfamily	P-loop containing nucleoside triphosphate hydrolases; unintegrated	3846	4571	726	SSF52540
3075	Superfamily	Xylose isomerase-like; Xylose isomerase-like, TIM barrel domain	9837	10556	720	SSF51658
3075	Superfamily	Periplasmic binding protein-like II; unintegrated	16580	17206	627	SSF53850
3075	Superfamily	Periplasmic binding protein-like II; unintegrated	13822	14448	627	SSF53850
3075	Superfamily	NAD(P)-binding Rossmann-fold domains; unintegrated	7167	7634	468	SSF51735
3075	Superfamily	MOP-like; Molybdate/tungstate binding, C-terminal	7771	8106	336	SSF50331
3075	Superfamily	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain; unintegrated	6918	7244	327	SSF55347
3075	Superfamily	Pili subunits; unintegrated	5941	6213	273	SSF54523
3075	Superfamily	"Winged helix" DNA-binding domain; unintegrated	13555	13821	267	SSF46785
3148	Superfamily	MFS general substrate transporter; Major facilitator superfamily domain, general substrate transporter	10465	11823	1359	SSF103473
3148	Superfamily	Aldolase; unintegrated	21971	22801	831	SSF51569
3148	Superfamily	FAH; Fumarylacetoacetase, C-terminal-related	20253	21032	780	SSF56529
3148	Superfamily	Formate dehydrogenase/DMSO reductase, domains 1-3; unintegrated	17268	18041	774	SSF53706
3148	Superfamily	Metalloproteases ("zincins"), catalytic domain; unintegrated	27225	27971	747	SSF55486
3148	Superfamily	HlyD-like secretion proteins; unintegrated	9380	10099	720	SSF111369
3148	Superfamily	P-loop containing nucleoside triphosphate hydrolases; unintegrated	35173	35865	693	SSF52540
3148	Superfamily	Periplasmic binding protein-like II; unintegrated	33526	34176	651	SSF53850
3148	Superfamily	Leukotriene A4 hydrolase N-terminal domain; unintegrated	27996	28640	645	SSF63737
3148	Superfamily	alpha/beta-Hydrolases; unintegrated	48599	49231	633	SSF53474
3148	Superfamily	P-loop containing nucleoside triphosphate hydrolases; unintegrated	36425	37033	609	SSF52540
3148	Superfamily	Molybdenum cofactor biosynthesis proteins; Molybdopterin binding domain	39227	39796	570	SSF53218
3148	Superfamily	Nucleic acid-binding proteins; Nucleic acid-binding, OB-fold	8267	8791	525	SSF50249
3148	Superfamily	P-loop containing nucleoside triphosphate hydrolases; unintegrated	2944	3438	495	SSF52540
3148	Superfamily	alpha/beta knot; unintegrated	44779	45177	399	SSF75217
3148	Superfamily	ATPase domain of HSP90 chaperone/DNA topoisomerase II/histidine kinase; Histidine kinase-like ATPase, ATP-binding domain	6366	6752	387	SSF55874
3148	Superfamily	"Winged helix" DNA-binding domain; unintegrated	34081	34419	339	SSF46785
3148	Superfamily	"Winged helix" DNA-binding domain; unintegrated	46403	46741	339	SSF46785

3148	Superfamily	Cyclophilin-like; Cyclophilin-like peptidyl-prolyl cis-trans isomerase domain	32227	32523	297	SSF50891
3148	Superfamily	Sporulation related repeat; Sporulation-related domain	407	619	213	SSF110997
6160	Superfamily	ABC transporter transmembrane region; ABC transporter, transmembrane domain, type 1	14783	15712	930	SSF90123
6160	Superfamily	Enolase C-terminal domain-like; unintegrated	8015	8881	867	SSF51604
6160	Superfamily	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase; unintegrated	4366	5193	828	SSF54593
6160	Superfamily	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase; unintegrated	3252	4043	792	SSF54593
6160	Superfamily	dUTPase-like; unintegrated	18252	18812	561	SSF51283
6160	Superfamily	RmlC-like cupins; RmlC-like cupin domain	2400	2957	558	SSF51182
6160	Superfamily	Glyoxalase/Bleomycin resistance protein/Dihydroxybiphenyl dioxygenase; unintegrated	12004	12552	549	SSF54593
6160	Superfamily	FAH; Fumarylacetoacetase, C-terminal-related	12854	13345	492	SSF56529
6160	Superfamily	Enolase N-terminal domain-like; unintegrated	8885	9286	402	SSF54826
6160	Superfamily	Dimeric alpha+beta barrel; Dimeric alpha-beta barrel	6317	6499	183	SSF54909
9794	Superfamily	RmlC-like cupins; RmlC-like cupin domain	9978	10808	831	SSF51182
9794	Superfamily	NAD(P)-binding Rossmann-fold domains; unintegrated	15958	16710	753	SSF51735
9794	Superfamily	P-loop containing nucleoside triphosphate hydrolases; unintegrated	5391	6125	735	SSF52540
9794	Superfamily	Periplasmic binding protein-like II; unintegrated	8815	9456	642	SSF53850
9794	Superfamily	Dihydrofolate reductase-like; Dihydrofolate reductase-like domain	1202	1681	480	SSF53597
9794	Superfamily	NAD(P)-binding Rossmann-fold domains; unintegrated	16832	17302	471	SSF51735
9794	Superfamily	Glyceraldehyde-3-phosphate dehydrogenase-like, C-terminal domain; unintegrated	17225	17689	465	SSF55347
9794	Superfamily	CheY-like; CheY-like superfamily	13030	13425	396	SSF52172
9794	Superfamily	Hedgehog/DD-peptidase; Hedgehog signalling/DD-peptidase zinc-binding domain	2466	2849	384	SSF55166
9794	Superfamily	"Winged helix" DNA-binding domain; unintegrated	8575	8910	336	SSF46785
9794	Superfamily	Fe-S cluster assembly (FSCA) domain-like; unintegrated	5148	5405	258	SSF117916
9794	Superfamily	C-terminal effector domain of the bipartite response regulators; Signal transduction response regulator, C-terminal effector	12796	13035	240	SSF46894
14785	Superfamily	Radical SAM enzymes; unintegrated	214	1056	843	SSF102114
14785	Superfamily	alpha/beta-Hydrolases; unintegrated	3139	3813	675	SSF53474
14785	Superfamily	FMN-dependent nitroreductase-like; Nitroreductase-like	1140	1697	558	SSF55469
18132	Superfamily	MFS general substrate transporter; Major facilitator superfamily domain, general substrate transporter	6754	7914	1161	SSF103473

18132	Superfamily	P-loop containing nucleoside triphosphate hydrolases; unintegrated	4370	5179	810	SSF52540
18132	Superfamily	Class I glutamine amidotransferase-like; unintegrated	3542	4324	783	SSF52317
18132	Superfamily	HAD-like; HAD-like domain	8553	9161	609	SSF56784
18132	Superfamily	SIS domain; unintegrated	5961	6398	438	SSF53697
23284	Superfamily	Multidrug efflux transporter AcrB pore domain; PN1, PN2, PC1 and PC2 subdomains; unintegrated	2573	3427	861	SSF82693
23284	Superfamily	FAH; Fumarylacetoacetase, C-terminal-related	3651	4247	597	SSF56529
33223	Superfamily	Glucocorticoid receptor-like (DNA-binding domain); unintegrated	472	570	99	SSF57716
58390	Superfamily	alpha/beta-Hydrolases; unintegrated	429	1229	801	SSF53474
243	TIGRFAMs	narG: nitrate reductase, alpha subunit; Nitrate reductase, alpha subunit	53536	55632	2097	TIGR01580
243	TIGRFAMs	outer_NodT: efflux transporter, outer membrane fac; RND efflux system, outer membrane lipoprotein, NodT	49434	50865	1432	TIGR01845
243	TIGRFAMs	bioA: adenosylmethionine-8-amino-7-oxononanoate tr; Adenosylmethionine--8-amino-7-oxononanoate aminotransferase BioA	36849	38060	1212	TIGR00508
243	TIGRFAMs	2A0108: nitrite transporter; Nitrate transporter	45465	46616	1152	TIGR00886
243	TIGRFAMs	TIGR00296: uncharacterized protein, PH0010 family; AMMECR1	35801	36226	426	TIGR00296
243	TIGRFAMs	TIGR00004: putative endoribonuclease L-PSP; YjgF-like protein	23167	23529	363	TIGR00004
3075	TIGRFAMs	salicylate_mono: salicylate 1-monooxygenase; Salicylate 1-monooxygenase	714	1955	1242	TIGR03219
3075	TIGRFAMs	ac_ald_DH_ac: acetaldehyde dehydrogenase (acetylalt; Acetaldehyde dehydrogenase	6912	7625	714	TIGR03215
3075	TIGRFAMs	IV_pilin_GFxxxE: prepilin-type N-terminal cleavage; Prokaryotic N-terminal methylation site	6151	6222	72	TIGR02532
3148	TIGRFAMs	VI_chp_6: type VI secretion protein, VC_A0110 fami; Type VI secretion system, VCA0110	22644	23828	1185	TIGR03359
3148	TIGRFAMs	RND_mfp: efflux transporter, RND family, MFP subun; RND efflux pump, membrane fusion protein	9371	10315	945	TIGR01730
3148	TIGRFAMs	KD08P_synth: 3-deoxy-8-phosphooctulonate synthase; 3-deoxy-8-phosphooctulonate synthase	21992	22765	774	TIGR01362
3148	TIGRFAMs	catechol_dmpE: 2-oxopent-4-enoate hydratase; 2-oxopent-4-enoate hydratase	20268	21032	765	TIGR03220
3148	TIGRFAMs	VI_zyme: type VI secretion system lysozyme-like pr; Type VI secretion system, lysozyme-related	25872	26324	453	TIGR03357
3148	TIGRFAMs	molyb_syn: molybdenum cofactor synthesis domain; Molybdenum cofactor synthesis	39233	39661	429	TIGR00177
6160	TIGRFAMs	MsbA_rel: ABC transporter, permease/ATP-binding pr; ABC transporter, ATP-binding/permease protein	13988	15715	1728	TIGR02204
6160	TIGRFAMs	eno: phosphopyruvate hydratase; Enolase	8018	9283	1266	TIGR01060
6160	TIGRFAMs	catechol_2_3: catechol 2,3 dioxygenase; Catechol 2,3 dioxygenase	3112	4043	932	TIGR03211
6160	TIGRFAMs	catechol_2_3: catechol 2,3 dioxygenase; Catechol 2,3 dioxygenase	4366	5277	912	TIGR03211

6160	TIGRFAMs	PPK2_P_aer: polyphosphate kinase 2; Polyphosphate kinase 2, PA0141	9676	10356	681	TIGR03707
6160	TIGRFAMs	dCTP_deam: deoxycytidine triphosphate deaminase; Deoxycytidine triphosphate deaminase	18252	18806	555	TIGR02274
9794	TIGRFAMs	ac_ald_DH_ac: acetaldehyde dehydrogenase (acetylac); Acetaldehyde dehydrogenase	16841	17758	918	TIGR03215
18132	TIGRFAMs	PyrG: CTP synthase; CTP synthase	3569	5176	1608	TIGR00337
18132	TIGRFAMs	HAD-SF-IA-v1: HAD hydrolase, family IA, variant 1; HAD-superfamily hydrolase, subfamily IA, variant 1	8832	9065	234	TIGR01549
23284	TIGRFAMs	catechol_dmpH: 4-oxalocrotonate decarboxylase; 4-oxalocrotonate decarboxylase	3651	4247	597	TIGR03218
243	TMHMM	transmembrane_regions	45474	46625	612	tmhmm
243	TMHMM	transmembrane_regions	42700	42762	63	tmhmm
243	TMHMM	transmembrane_regions	53596	53658	63	tmhmm
243	TMHMM	transmembrane_regions	57398	57460	63	tmhmm
243	TMHMM	transmembrane_regions	58053	58115	63	tmhmm
243	TMHMM	transmembrane_regions	6613	6669	57	tmhmm
243	TMHMM	transmembrane_regions	49419	49475	57	tmhmm
3075	TMHMM	transmembrane_regions	2105	3019	633	tmhmm
3075	TMHMM	transmembrane_regions	12848	12910	63	tmhmm
3075	TMHMM	transmembrane_regions	1896	1958	63	tmhmm
3075	TMHMM	transmembrane_regions	6151	6207	57	tmhmm
3148	TMHMM	transmembrane_regions	24775	25068	177	tmhmm
3148	TMHMM	transmembrane_regions	50035	50244	189	tmhmm
3148	TMHMM	transmembrane_regions	5037	5105	69	tmhmm
3148	TMHMM	transmembrane_regions	29095	29163	69	tmhmm
3148	TMHMM	transmembrane_regions	965	1027	63	tmhmm
3148	TMHMM	transmembrane_regions	37194	37256	63	tmhmm
3148	TMHMM	transmembrane_regions	38528	38590	63	tmhmm
3148	TMHMM	transmembrane_regions	45106	45168	63	tmhmm
3148	TMHMM	transmembrane_regions	1990	2046	57	tmhmm
3148	TMHMM	transmembrane_regions	9245	9301	57	tmhmm
3148	TMHMM	transmembrane_regions	18476	18532	57	tmhmm

3148	TMHMM	transmembrane_regions	51985	52041	57	tmhmm
5976	TMHMM	transmembrane_regions	93	902	372	tmhmm
6160	TMHMM	transmembrane_regions	14918	15676	327	tmhmm
6160	TMHMM	transmembrane_regions	2341	2403	63	tmhmm
6160	TMHMM	transmembrane_regions	5390	5446	57	tmhmm
6160	TMHMM	transmembrane_regions	9700	9756	57	tmhmm
6160	TMHMM	transmembrane_regions	16593	16649	57	tmhmm
9794	TMHMM	transmembrane_regions	1613	1675	63	tmhmm
9794	TMHMM	transmembrane_regions	2340	2396	57	tmhmm
18132	TMHMM	transmembrane_regions	6766	7914	744	tmhmm
18132	TMHMM	transmembrane_regions	5099	5167	69	tmhmm
23284	TMHMM	transmembrane_regions	403	459	57	tmhmm

References

- Abbai, N. S., Govender, A., Shaik, R. & Pillay, B. (2012). Pyrosequence analysis of unamplified and whole genome amplified DNA from hydrocarbon-contaminated groundwater. *Mol Biotechnol* **50**, 39–48.
- Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**, 3389–402.
- Ames, B. N., Gurney, E. G., Miller, J. A. & Bartsch, H. (1972). Carcinogens as frameshift mutagens: metabolites and derivatives of 2-acetylaminofluorene and other aromatic amine carcinogens. *Proc Natl Acad Sci USA* **69**, 3128–32.
- Andreoni, V., Cavalca, L., Rao, M. A., Nocerino, G., Bernasconi, S., Dell'Amico, E., Colombo, M. & Gianfreda, L. (2004). Bacterial communities and enzyme activities of PAHs polluted soils. *Chemosphere* **57**, 401–412.
- Baba, T., Ara, T., Hasegawa, M., Takai, Y., Okumura, Y., Baba, M., Datsenko, K. A., Tomita, M., Wanner, B. L. & Mori, H. (2006). Construction of *Escherichia coli* K-12 in-frame, single-gene knockout mutants: the Keio collection. *Mol Syst Biol* **2**, 2006–8.
- Barkay, T., Turner, R. R., Rasmussen, L. D., Kelly, C. A. & Rudd, J. W. (1998). Luminescence facilitated detection of bioavailable mercury in natural waters. *Methods Mol Biol* **102**, 231–46.
- Bell, T. H., Yergeau, E., Martineau, C., Juck, D., Whyte, L. G. & Greer, C. W. (2011). Identification of nitrogen-incorporating bacteria in petroleum-contaminated arctic soils by using [¹⁵N]DNA-based stable isotope probing and pyrosequencing. *Appl Environ Microbiol* **77**, 4163–71.
- van den Berg, B. (2005). The FadL family: unusual transporters for unusual substrates. *Curr Opin Struct Biol* **15**, 401–7.
- Blainey, P. C. (2013). The future is now: single-cell genomics of bacteria and archaea. *FEMS Microbiol Rev*.
- Boonchan, S., Britz, M. L. & Stanley, G. A. (1998). Surfactant-enhanced biodegradation of high molecular weight polycyclic aromatic hydrocarbons by

- Stenotrophomonas maltophilia*. *Biotechnol Bioeng* **59**, 482–494.
- Bosch, R., Moore, E. R., García-Valdés, E. & Pieper, D. H. (1999)**. NahW, a novel, inducible salicylate hydroxylase involved in mineralization of naphthalene by *Pseudomonas stutzeri* AN10. *J Bacteriol* **181**, 2315–22.
- Brezna, B., Khan, A. A. & Cerniglia, C. E. (2003)**. Molecular characterization of dioxygenases from polycyclic aromatic hydrocarbon-degrading *Mycobacterium* spp. *FEMS Microbiol Lett* **223**, 177–83.
- Brinkrolf, K., Brune, I. & Tauch, A. (2006)**. Transcriptional regulation of catabolic pathways for aromatic compounds in *Corynebacterium glutamicum*. *Genet Mol Res* **5**, 773–89.
- Brown, N. L., Stoyanov, J. V., Kidd, S. P. & Hobman, J. L. (2003)**. The MerR family of transcriptional regulators. *FEMS Microbiol Rev* **27**, 145–63.
- Brunet-Galmés, I., Busquets, A., Peña, A., Gomila, M., Nogales, B., García-Valdés, E., Lalucat, J., Bennasar, A. & Bosch, R. (2012)**. Complete genome sequence of the naphthalene-degrading bacterium *Pseudomonas stutzeri* AN10 (CCUG 29243). *J Bacteriol* **194**, 6642–3.
- Campbell, E. A., Westblade, L. F. & Darst, S. A. (2008)**. Regulation of bacterial RNA polymerase sigma factor activity: a structural perspective. *Curr Opin Microbiol* **11**, 121–7.
- Carl, B. & Fetzner, S. (2005)**. Transcriptional activation of quinoline degradation operons of *Pseudomonas putida* 86 by the AraC/XylS-type regulator OxoS and cross-regulation of the PqorM promoter by XylS. *Appl Environ Microbiol* **71**, 8618–26.
- Carvalhais, L. C., Dennis, P. G., Tyson, G. W. & Schenk, P. M. (2012)**. Application of metatranscriptomics to soil environments. *J Microbiol Methods* **91**, 246–51.
- Cases, I. & de Lorenzo, V. (2005)**. Promoters in the environment: transcriptional regulation in its natural context. *Nat Rev Microbiol* **3**, 105–18.
- Cébron, A., Louvel, B., Faure, P., France-Lanord, C., Chen, Y., Murrell, J. C. & Leyval, C. (2011)**. Root exudates modify bacterial diversity of phenanthrene degraders in PAH-polluted soil but not phenanthrene degradation rates. *Environ Microbiol* **13**, 722–36.

- Chakraborty, J., Ghosal, D., Dutta, A. & Dutta, T. K. (2012).** An insight into the origin and functional evolution of bacterial aromatic ring-hydroxylating oxygenases. *J Biomol Struct Dyn* **30**, 419–36.
- Chen, Y. & Murrell, J. C. (2010).** When metagenomics meets stable-isotope probing: progress and perspectives. *Trends Microbiol* **18**, 157–63.
- Chen, Y., Dumont, M. G., Neufeld, J. D., Bodrossy, L., Stralis-Pavese, N., McNamara, N. P., Ostle, N., Briones, M. J. I. & Murrell, J. C. (2008).** Revealing the uncultivated majority: combining DNA stable-isotope probing, multiple displacement amplification and metagenomic analyses of uncultivated *Methylocystis* in acidic peatlands. *Environ Microbiol* **10**, 2609–22.
- Churchill, P. F., Morgan, A. C. & Kitchens, E. (2008).** Characterization of a pyrene-degrading *Mycobacterium* sp. strain CH-2. *J Environ Sci Health B* **43**, 698–706.
- Cowles, C. E., Nichols, N. N. & Harwood, C. S. (2000).** BenR, a XylS homologue, regulates three different pathways of aromatic acid degradation in *Pseudomonas putida*. *J Bacteriol* **182**, 6339–46.
- Craig, J. W., Chang, F. Y., Kim, J. H., Obiajulu, S. C. & Brady, S. F. (2010).** Expanding small-molecule functional metagenomics through parallel screening of broad-host-range cosmid environmental DNA libraries in diverse proteobacteria. *Appl Environ Microbiol* **76**, 1633–41.
- Dagher, F., Déziel, E., Lirette, P., Paquette, G., Bisailon, J. G. & Villemur, R. (1997).** Comparative study of five polycyclic aromatic hydrocarbon degrading bacterial strains isolated from contaminated soils. *Can J Microbiol* **43**, 368–77.
- Daguené, V., McFall, E., Yumvihoze, E., Xiang, S., Amyot, M. & Poulain, A. J. (2012).** Divalent base cations hamper Hg(II) uptake. *Environ Sci Technol* **46**, 6645–53.
- Dal, S., Trautwein, G. & Gerischer, U. (2005).** Transcriptional organization of genes for protocatechuate and quinate degradation from *Acinetobacter* sp. strain ADP1. *Appl Environ Microbiol* **71**, 1025–34.
- Daniel, R. (2005).** The metagenomics of soil. *Nat Rev Microbiol* **3**, 470–8.
- Dantas, G., Sommer, M. O. A., Degnan, P. H. & Goodman, A. L. (2013).** Experimental approaches for defining functional roles of microbes in the human gut. *Annu Rev Microbiol* **67**, 459–75.

- DeBruyn, J. M., Chewing, C. S. & Sayler, G. S. (2007). Comparative quantitative prevalence of *Mycobacteria* and functionally abundant *nidA*, *nahAc*, and *nagAc* dioxygenase genes in coal tar contaminated sediments. *Environ Sci Technol* **41**, 5426–32.
- DeBruyn, J. M., Mead, T. J. & Sayler, G. S. (2012). Horizontal transfer of PAH catabolism genes in *Mycobacterium*: evidence from comparative genomics and isolated pyrene-degrading bacteria. *Environ Sci Technol* **46**, 99–106.
- Decker, K. B. & Hinton, D. M. (2013). Transcription regulation at the core: similarities among bacterial, archaeal, and eukaryotic RNA polymerases. *Annu Rev Microbiol* **67**, 113–39.
- Delgado, A. & Ramos, J. L. (1994). Genetic evidence for activation of the positive transcriptional regulator Xy1R, a member of the NtrC family of regulators, by effector binding. *J Biol Chem* **269**, 8059–8062.
- Delmont, T. O., Malandain, C., Prestat, E., Larose, C., Monier, J.-M., Simonet, P. & Vogel, T. M. (2011a). Metagenomic mining for microbiologists. *ISME J* **5**, 1837–43.
- Delmont, T. O., Robe, P., Clark, I., Simonet, P. & Vogel, T. M. (2011b). Metagenomic comparison of direct and indirect soil DNA extraction approaches. *J Microbiol Methods* **86**, 397–400.
- Delmont, T. O., Prestat, E., Keegan, K. P., Faubladiere, M., Robe, P., Clark, I. M., Pelletier, E., Hirsch, P. R., Meyer, F. & other authors. (2012a). Structure, fluctuation and magnitude of a natural grassland soil metagenome. *ISME J* **6**, 1677–87.
- Delmont, T. O., Simonet, P. & Vogel, T. M. (2012b). Describing microbial communities and performing global comparisons in the 'omic era. *ISME J* **6**, 1625–8.
- Demanèche, S., Meyer, C., Micoud, J., Louwagie, M., Willison, J. C. & Jouanneau, Y. (2004). Identification and functional analysis of two aromatic-ring-hydroxylating dioxygenases from a *Sphingomonas* strain that degrades various polycyclic aromatic hydrocarbons. *Appl Environ Microbiol* **70**, 6714–25.
- Demanèche, S., David, M. M., Navarro, E., Simonet, P. & Vogel, T. M. (2009). Evaluation of functional gene enrichment in a soil metagenomic clone library. *J Microbiol Methods* **76**, 105–7.

- Demple, B. (1996).** Redox signaling and gene control in the *Escherichia coli* *soxRS* oxidative stress regulon—a review. *Gene* **179**, 53–7.
- Denef, V. J., Mueller, R. S. & Banfield, J. F. (2010).** AMD biofilms: using model communities to study microbial evolution and ecological complexity in nature. *ISME J* **4**, 599–610.
- Dennis, J. J. & Zylstra, G. J. (2004).** Complete sequence and genetic organization of pDTG1, the 83 kilobase naphthalene degradation plasmid from *Pseudomonas putida* strain NCIB 9816-4. *J Mol Biol* **341**, 753–68.
- Díaz, E. & Prieto, M. A. (2000).** Bacterial promoters triggering biodegradation of aromatic pollutants. *Curr Opin Biotechnol* **11**, 467–75.
- Díaz, E., Ferrandez, A., Prieto, M. A. & Garcia, J. L. (2001).** Biodegradation of aromatic compounds by *Escherichia coli*. *Microbiol Mol Biol Rev* **65**, 523–69, table.
- Díaz, E. (2004).** Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility. *Int Microbiol* **7**, 173–80.
- Dunn, A. K. & Handelsman, J. (1999).** A vector for promoter trapping in *Bacillus cereus*. *Gene* **226**, 297–305.
- Dunn, A. K., Klimowicz, A. K. & Handelsman, J. (2003).** Use of a promoter trap to identify *Bacillus cereus* genes regulated by tomato seed exudate and a rhizosphere resident, *Pseudomonas aureofaciens*. *Appl Environ Microbiol* **69**, 1197–205.
- Edgar, R. C. (2004).** MUSCLE: a multiple sequence alignment method with reduced time and space complexity. *BMC Bioinformatics* **5**, 113.
- Ekkers, D. M., Cretoiu, M. S., Kielak, A. M. & Elsas, J. D. van. (2012).** The great screen anomaly - a new frontier in product discovery through functional metagenomics. *Appl Microbiol Biotechnol* **93**, 1005–20.
- Epstein, S. (2013).** The phenomenon of microbial uncultivability. *Curr Opin Microbiol*.
- Eriksson, M., Sodersten, E., Yu, Z., Dalhammar, G. & Mohn, W. W. (2003).** Degradation of polycyclic aromatic hydrocarbons at low temperature under aerobic and nitrate-reducing conditions in enrichment cultures from northern soils. *Appl Environ Microbiol* **69**, 275–284.

- Fang, H., Cai, L., Yu, Y. & Zhang, T. (2013). Metagenomic analysis reveals the prevalence of biodegradation genes for organic pollutants in activated sludge. *Bioresour Technol* **129**, 209–18.
- Fernandez, C., Ferrandez, A., Minambres, B., Diaz, E. & Garcia, J. L. (2006). Genetic characterization of the phenylacetyl-coenzyme A oxygenase from the aerobic phenylacetic acid degradation pathway of *Escherichia coli*. *Appl Environ Microbiol* **72**, 7422–7426.
- Fernández, M., Niqui-Arroyo, J. L., Conde, S., Ramos, J. L. & Duque, E. (2012). Enhanced tolerance to naphthalene and enhanced rhizoremediation performance for *Pseudomonas putida* KT2440 via the NAH7 catabolic plasmid. *Appl Environ Microbiol* **78**, 5104–10.
- Ferrandez, A., Minambres, B., Garcia, B., Olivera, E. R., Luengo, J. M., Garcia, J. L. & Diaz, E. (1998). Catabolism of phenylacetic acid in *Escherichia coli*. Characterization of a new aerobic hybrid pathway. *J Biol Chem* **273**, 25974–25986.
- Ferrandez, A., Garcia, J. L. & Diaz, E. (2000). Transcriptional regulation of the divergent paa catabolic operons for phenylacetic acid degradation in *Escherichia coli*. *J Biol Chem* **275**, 12214–12222.
- Fischer, R., Bleichrodt, F. S. & Gerischer, U. C. (2008). Aromatic degradative pathways in *Acinetobacter baylyi* underlie carbon catabolite repression. *Microbiology* **154**, 3095–103.
- Fish, J. A., Chai, B., Wang, Q., Sun, Y., Brown, C. T., Tiedje, J. M. & Cole, J. R. (2013). FunGene: the functional gene pipeline and repository. *Front Microbiol* **4**, 291.
- Fujihara, H., Yoshida, H., Matsunaga, T., Goto, M. & Furukawa, K. (2006). Cross-regulation of biphenyl- and salicylate-catabolic genes by two regulatory systems in *Pseudomonas pseudoalcaligenes* KF707. *J Bacteriol* **188**, 4690–7.
- Gabor, E. M., Alkema, W. B. L. & Janssen, D. B. (2004). Quantifying the accessibility of the metagenome by random expression cloning techniques. *Environ Microbiol* **6**, 879–86.
- Gallegos, M. T., Williams, P. A. & Ramos, J. L. (1997). Transcriptional control of the multiple catabolic pathways encoded on the TOL plasmid pWW53 of *Pseudomonas putida* MT53. *J Bacteriol* **179**, 5024–9.

- Galvão, T. C., Mencía, M. & de Lorenzo, V. (2007).** Emergence of novel functions in transcriptional regulators by regression to stem protein types. *Mol Microbiol* **65**, 907–19.
- Di Gennaro, P., Terreni, P., Masi, G., Botti, S., De Ferra, F. & Bestetti, G. (2010).** Identification and characterization of genes involved in naphthalene degradation in *Rhodococcus opacus* R7. *Appl Microbiol Biotechnol* **87**, 297–308.
- Gerischer, U. (2002).** Specific and global regulation of genes associated with the degradation of aromatic compounds in bacteria. *J Mol Microbiol Biotechnol* **4**, 111–21.
- Gillespie, D. E., Brady, S. F., Bettermann, A. D., Cianciotto, N. P., Liles, M. R., Rondon, M. R., Clardy, J., Goodman, R. M. & Handelsman, J. (2002).** Isolation of antibiotics turbomycin A and B from a metagenomic library of soil microbial DNA. *Appl Environ Microbiol* **68**, 4301–6.
- Görke, B. & Stülke, J. (2008).** Carbon catabolite repression in bacteria: many ways to make the most out of nutrients. *Nat Rev Microbiol* **6**, 613–24.
- Gottfried, A., Singhal, N., Elliot, R. & Swift, S. (2010).** The role of salicylate and biosurfactant in inducing phenanthrene degradation in batch soil slurries. *Appl Microbiol Biotechnol* **86**, 1563–71.
- Groster, A. & Edwards, E. A. (2006).** Growth of *Dehalobacter* and *Dehalococcoides* spp. during degradation of chlorinated ethanes. *Appl Environ Microbiol* **72**, 428–36.
- Habe, H. & Omori, T. (2003).** Genetics of polycyclic aromatic hydrocarbon metabolism in diverse aerobic bacteria. *Biosci Biotechnol Biochem* **67**, 225–43.
- Haddad, S., Eby, D. M. & Neidle, E. L. (2001).** Cloning and expression of the benzoate dioxygenase genes from *Rhodococcus* sp. strain 19070. *Appl Environ Microbiol* **67**, 2507–14.
- Hamann, C., Hegemann, J. & Hildebrandt, A. (1999).** Detection of polycyclic aromatic hydrocarbon degradation genes in different soil bacteria by polymerase chain reaction and DNA hybridization. *FEMS Microbiol Lett* **173**, 255–263.
- Hancock, R. E. (1984).** Alterations in outer membrane permeability. *Annu Rev Microbiol* **38**, 237–64.
- Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J. & Goodman, R. M. (1998).** Molecular biological access to the chemistry of unknown soil microbes: a

- new frontier for natural products. *Chem Biol* **5**, R245–9.
- Handelsman, J. (2004).** Metagenomics: application of genomics to uncultured microorganisms. *Microbiol Mol Biol Rev* **68**, 669–85.
- Handelsman, J. (2005).** Sorting out metagenomes. *Nat Biotechnol* **23**, 38–9.
- Hårdeman, F. & Sjöling, S. (2007).** Metagenomic approach for the isolation of a novel low-temperature-active lipase from uncultured bacteria of marine sediment. *FEMS Microbiol Ecol* **59**, 524–34.
- Haritash, A. K. & Kaushik, C. P. (2009).** Biodegradation aspects of polycyclic aromatic hydrocarbons (PAHs): a review. *J Hazard Mater* **169**, 1–15.
- Harwood, C. S. & Parales, R. E. (1996).** The beta-ketoadipate pathway and the biology of self-identity. *Annu Rev Microbiol* **50**, 553–90.
- Heinaru, E., Vedler, E., Jutkina, J., Aava, M. & Heinaru, A. (2009).** Conjugal transfer and mobilization capacity of the completely sequenced naphthalene plasmid pNAH20 from multiplasmid strain *Pseudomonas fluorescens* PC20. *FEMS Microbiol Ecol* **70**, 563–74.
- Heinrichs, D. E., Yethon, J. A. & Whitfield, C. (1998).** Molecular basis for structural diversity in the core regions of the lipopolysaccharides of *Escherichia coli* and *Salmonella enterica*. *Mol Microbiol* **30**, 221–32.
- Heitkamp, M. A., Franklin, W. & Cerniglia, C. E. (1988).** Microbial metabolism of polycyclic aromatic hydrocarbons: isolation and characterization of a pyrene-degrading bacterium. *Appl Environ Microbiol* **54**, 2549–2555.
- Helmann, J. D. (2009).** RNA polymerase: a nexus of gene regulation. *Methods* **47**, 1–5.
- Henne, A., Daniel, R., Schmitz, R. A. & Gottschalk, G. (1999).** Construction of environmental DNA libraries in *Escherichia coli* and screening for the presence of genes conferring utilization of 4-hydroxybutyrate. *Appl Environ Microbiol* **65**, 3901–7.
- Henry, C. S., Overbeek, R., Xia, F., Best, A. A., Glass, E., Gilbert, J., Larsen, P., Edwards, R., Disz, T. & other authors. (2011).** Connecting genotype to phenotype in the era of high-throughput sequencing. *Biochim Biophys Acta* **1810**, 967–77.

- Hermans, K., Nguyen, T. L. A., Roberfroid, S., Schoofs, G., Verhoeven, T., De Coster, D., Vanderleyden, J. & De Keersmaecker, S. C. J. (2011). Gene expression analysis of monospecies *Salmonella typhimurium* biofilms using differential fluorescence induction. *J Microbiol Methods* **84**, 467–78.
- Hickey, W. J., Chen, S. & Zhao, J. (2012). The phn Island: A New Genomic Island Encoding Catabolism of Polynuclear Aromatic Hydrocarbons. *Front Microbiol* **3**, 125.
- Hill, G., Mitkowski, N., Aldrich-Wolfe, L., Emele, L., Jurkonie, D., Ficke, A., Maldonado-Ramirez, S., Lynch, S. & Nelson, E. (2000). Methods for assessing the composition and diversity of soil microbial communities. *Applied soil ecology* **15**, 25–36. Elsevier.
- Hug, L. A., Beiko, R. G., Rowe, A. R., Richardson, R. E. & Edwards, E. A. (2012). Comparative metagenomics of three *Dehalococcoides*-containing enrichment cultures: the role of the non-dechlorinating community. *BMC Genomics* **13**, 327.
- Huson, D. H., Auch, A. F., Qi, J. & Schuster, S. C. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**, 377–86.
- Hynninen, A. & Virta, M. (2010). Whole-cell bioreporters for the detection of bioavailable metals. *Adv Biochem Eng Biotechnol* **118**, 31–63.
- IARC. (1983). Polynuclear aromatic compounds, Part 1, Chemical, environmental and experimental data. *IARC Monogr Eval Carcinog Risk Chem Hum* **32**, 1–453.
- Izmalkova, T. Y., Sazonova, O. I., Nagornih, M. O., Sokolov, S. L., Kosheleva, I. A. & Boronin, A. M. (2013). The organization of naphthalene degradation genes in *Pseudomonas putida* strain AK5. *Res Microbiol* **164**, 244–53.
- James, K. D. & Williams, P. A. (1998). ntn genes determining the early steps in the divergent catabolism of 4-nitrotoluene and toluene in *Pseudomonas* sp. strain TW3. *J Bacteriol* **180**, 2043–9.
- Jimenez, J. I., Minambres, B., Garcia, J. L. & Diaz, E. (2002). Genomic analysis of the aromatic catabolic pathways from *Pseudomonas putida* KT2440. *Environ Microbiol* **4**, 824–841.
- Johnsen, A. R., Wick, L. Y. & Harms, H. (2005). Principles of microbial PAH-degradation in soil. *Environ Pollut* **133**, 71–84.

- Johnsen, A. R., Schmidt, S., Hybholt, T. K., Henriksen, S., Jacobsen, C. S. & Andersen, O. (2007).** Strong impact on the polycyclic aromatic hydrocarbon (PAH)-degrading community of a PAH-polluted soil but marginal effect on PAH degradation when priming with bioremediated soil dominated by mycobacteria. *Appl Environ Microbiol* **73**, 1474–1480.
- Johnsen, A. R. & Karlson, U. (2007).** Diffuse PAH contamination of surface soils: environmental occurrence, bioavailability, and microbial degradation. *Appl Microbiol Biotechnol* **76**, 533–43.
- Jones, M. D., Crandell, D. W., Singleton, D. R. & Aitken, M. D. (2011).** Stable-isotope probing of the polycyclic aromatic hydrocarbon-degrading bacterial guild in a contaminated soil. *Environ Microbiol* **13**, 2623–32.
- Jouanneau, Y. & Meyer, C. (2006).** Purification and characterization of an arene cis-dihydrodiol dehydrogenase endowed with broad substrate specificity toward polycyclic aromatic hydrocarbon dihydrodiols. *Appl Environ Microbiol* **72**, 4726–34.
- Kahng, H.-Y. & Oh, K.-H. (2005).** Molecular detection of catabolic genes for polycyclic aromatic hydrocarbons in the reed rhizosphere of Sunchon Bay. *J Microbiol* **43**, 572–6.
- Kakirde, K. S., Parsley, L. C. & Liles, M. R. (2010).** Size Does Matter: Application-driven Approaches for Soil Metagenomics. *Soil Biol Biochem* **42**, 1911–1923.
- Kamath, R., Schnoor, J. L. & Alvarez, P. J. J. (2004).** Effect of root-derived substrates on the expression of *nah-lux* genes in *Pseudomonas fluorescens* HK44: implications for PAH biodegradation in the rhizosphere. *Environ Sci Technol* **38**, 1740–5.
- Kanaly, R. A. & Harayama, S. (2000).** Biodegradation of high-molecular-weight polycyclic aromatic hydrocarbons by bacteria. *J Bacteriol* **182**, 2059–67.
- Kasuga, K., Habe, H., Chung, J. S., Yoshida, T., Nojiri, H., Yamane, H. & Omori, T. (2001).** Isolation and characterization of the genes encoding a novel oxygenase component of angular dioxygenase from the gram-positive dibenzofuran-degrader *Terrabacter* sp. strain DBF63. *Biochem Biophys Res Commun* **283**, 195–204.
- Khan, A. A., Wang, R. F., Cao, W. W., Doerge, D. R., Wennerstrom, D. & Cerniglia, C. E. (2001).** Molecular cloning, nucleotide sequence, and expression of genes encoding a polycyclic aromatic ring dioxygenase from *Mycobacterium* sp. strain PYR-1. *Appl Environ Microbiol* **67**, 3577–85.

- Kim, B. S., Kim, S. Y., Park, J., Park, W., Hwang, K. Y., Yoon, Y. J., Oh, W. K., Kim, B. Y. & Ahn, J. S. (2007a). Sequence-based screening for self-sufficient P450 monooxygenase from a metagenome library. *J Appl Microbiol* **102**, 1392–400.
- Kim, J. & Rees, D. C. (1994). Nitrogenase and biological nitrogen fixation. *Biochemistry* **33**, 389–97.
- Kim, S. J., Kweon, O., Jones, R. C., Edmondson, R. D. & Cerniglia, C. E. (2008). Genomic analysis of polycyclic aromatic hydrocarbon degradation in *Mycobacterium vanbaalenii* PYR-1. *Biodegradation* **19**, 859–881.
- Kim, S.-J., Jones, R. C., Cha, C.-J., Kweon, O., Edmondson, R. D. & Cerniglia, C. E. (2004). Identification of proteins induced by polycyclic aromatic hydrocarbon in *Mycobacterium vanbaalenii* PYR-1 using two-dimensional polyacrylamide gel electrophoresis and de novo sequencing methods. *Proteomics* **4**, 3899–908.
- Kim, S.-J., Kweon, O., Freeman, J. P., Jones, R. C., Adjei, M. D., Jhoo, J.-W., Edmondson, R. D. & Cerniglia, C. E. (2006). Molecular cloning and expression of genes encoding a novel dioxygenase involved in low- and high-molecular-weight polycyclic aromatic hydrocarbon degradation in *Mycobacterium vanbaalenii* PYR-1. *Appl Environ Microbiol* **72**, 1045–54.
- Kim, S.-J., Kweon, O., Jones, R. C., Freeman, J. P., Edmondson, R. D. & Cerniglia, C. E. (2007b). Complete and integrated pyrene degradation pathway in *Mycobacterium vanbaalenii* PYR-1 based on systems biology. *J Bacteriol* **189**, 464–72.
- Kim, S.-J., Kweon, O. & Cerniglia, C. E. (2009). Proteomic applications to elucidate bacterial aromatic hydrocarbon metabolic pathways. *Curr Opin Microbiol* **12**, 301–9.
- Kirk, J. L., Beaudette, L. A., Hart, M., Moutoglis, P., Klironomos, J. N., Lee, H. & Trevors, J. T. (2004). Methods of studying soil microbial diversity. *J Microbiol Methods* **58**, 169–88.
- Knietsch, A., Waschowitz, T., Bowien, S., Henne, A. & Daniel, R. (2003). Construction and screening of metagenomic libraries derived from enrichment cultures: generation of a gene bank for genes conferring alcohol oxidoreductase activity on *Escherichia coli*. *Appl Environ Microbiol* **69**, 1408–16.
- Kouzuma, A., Pinyakong, O., Nojiri, H., Omori, T., Yamane, H. & Habe, H. (2006). Functional and transcriptional analyses of the initial oxygenase genes for acenaphthene degradation from *Sphingomonas* sp. strain A4. *Microbiology* **152**,

2455–67.

- Kristiansson, E., Fick, J., Janzon, A., Grabic, R., Rutgersson, C., Weijdegård, B., Söderström, H. & Larsson, D. G. J. (2011). Pyrosequencing of antibiotic-contaminated river sediments reveals high levels of resistance and gene transfer elements. *PLoS ONE* **6**, e17038.
- Krivobok, S., Kuony, S., Meyer, C., Louwagie, M., Willison, J. C. & Jouanneau, Y. (2003). Identification of pyrene-induced proteins in *Mycobacterium* sp. strain 6PY1: evidence for two ring-hydroxylating dioxygenases. *J Bacteriol* **185**, 3828–41.
- Kweon, O., Kim, S.-J., Jones, R. C., Freeman, J. P., Adjei, M. D., Edmondson, R. D. & Cerniglia, C. E. (2007). A polyomic approach to elucidate the fluoranthene-degradative pathway in *Mycobacterium vanbaalenii* PYR-1. *J Bacteriol* **189**, 4635–47.
- Kweon, O., Kim, S.-J., Freeman, J. P., Song, J., Baek, S. & Cerniglia, C. E. (2010). Substrate specificity and structural characteristics of the novel Rieske nonheme iron aromatic ring-hydroxylating oxygenases NidAB and NidA3B3 from *Mycobacterium vanbaalenii* PYR-1. *MBio* **1**.
- Labana, S., Kapur, M., Malik, D. K., Prakash, D. & Jain, R. (2007). Diversity, biodegradation and bioremediation of polycyclic aromatic hydrocarbons. In *Environmental Bioremediation Technologies*, pp. 409–443. Springer.
- Lafortune, I., Juteau, P., Déziel, E., Lépine, F., Beaudet, R. & Villemur, R. (2009). Bacterial diversity of a consortium degrading high-molecular-weight polycyclic aromatic hydrocarbons in a two-liquid phase biosystem. *Microb Ecol* **57**, 455–68.
- Lambert, I. B., Carroll, C., Laycock, N., Koziarz, J., Lawford, I., Duval, L., Turner, G., Booth, R., Douville, S. & other authors. (2001). Cellular determinants of the mutational specificity of 1-nitroso-6-nitropyrene and 1-nitroso-8-nitropyrene in the *lacI* gene of *Escherichia coli*. *Mutat Res* **484**, 19–48.
- Lämmle, K., Zipper, H., Breuer, M., Hauer, B., Buta, C., Brunner, H. & Rupp, S. (2007). Identification of novel enzymes with different hydrolytic activities by metagenome expression cloning. *J Biotechnol* **127**, 575–92.
- de Las Heras, A., Fraile, S. & de Lorenzo, V. (2012). Increasing Signal Specificity of the TOL Network of *Pseudomonas putida* mt-2 by Rewiring the Connectivity of the Master Regulator XylR. *PLoS Genet* **8**, e1002963.

- Laurie, A. D. & Lloyd-Jones, G. (1999). The *phn* genes of *Burkholderia* sp. strain RP007 constitute a divergent gene cluster for polycyclic aromatic hydrocarbon catabolism. *J Bacteriol* **181**, 531–540.
- Lee, M.-H., Lee, C.-H., Oh, T.-K., Song, J. K. & Yoon, J.-H. (2006). Isolation and characterization of a novel lipase from a metagenomic library of tidal flat sediments: evidence for a new family of bacterial lipases. *Appl Environ Microbiol* **72**, 7406–9.
- Lee, S. H., Kim, J. M., Lee, H. J. & Jeon, C. O. (2011). Screening of promoters from rhizosphere metagenomic DNA using a promoter-trap vector and flow cytometric cell sorting. *J Basic Microbiol* **51**, 52–60.
- Lee, S.-E., Seo, J.-S., Keum, Y.-S., Lee, K.-J. & Li, Q. X. (2007). Fluoranthene metabolism and associated proteins in *Mycobacterium* sp. JS14. *Proteomics* **7**, 2059–69.
- Lemieux, C. L., Lambert, I. B., Lundstedt, S., Tysklind, M. & White, P. A. (2008). Mutagenic hazards of complex polycyclic aromatic hydrocarbon mixtures in contaminated soil. *Environ Toxicol Chem* **27**, 978–90.
- Li, J. L. & Chen, B. H. (2009). Surfactant-mediated biodegradation of polycyclic aromatic hydrocarbons. *Materials* **2**, 76–94. Molecular Diversity Preservation International.
- Li, W., Shi, J., Wang, X., Han, Y., Tong, W., Ma, L., Liu, B. & Cai, B. (2004). Complete nucleotide sequence and organization of the naphthalene catabolic plasmid pND6-1 from *Pseudomonas* sp. strain ND6. *Gene* **336**, 231–40.
- Liang, Y., Gardner, D. R., Miller, C. D., Chen, D., Anderson, A. J., Weimer, B. C. & Sims, R. C. (2006). Study of biochemical pathways and enzymes involved in pyrene degradation by *Mycobacterium* sp. strain KMS. *Appl Environ Microbiol* **72**, 7821–8.
- Liu, L., Li, Y., Li, S., Hu, N., He, Y., Pong, R., Lin, D., Lu, L. & Law, M. (2012). Comparison of next-generation sequencing systems. *J Biomed Biotechnol* **2012**, 251364.
- Liu, M., Durfee, T., Cabrera, J. E., Zhao, K., Jin, D. J. & Blattner, F. R. (2005). Global transcriptional programs reveal a carbon source foraging strategy by *Escherichia coli*. *J Biol Chem* **280**, 15921–7.

- Liu, S. & Suflita, J. M. (1993).** Ecology and evolution of microbial populations for bioremediation. *Trends Biotechnol* **11**, 344–52.
- Lönneborg, R. & Brzezinski, P. (2011).** Factors that influence the response of the LysR type transcriptional regulators to aromatic compounds. *BMC Biochem* **12**, 49.
- de Lorenzo, V. (2005).** Problems with metagenomic screening. *Nat Biotechnol* **23**, 1045; author reply 1045–6.
- Lors, C., Ryngaert, A., Périé, F., Diels, L. & Damidot, D. (2010).** Evolution of bacterial community during bioremediation of PAHs in a coal tar contaminated soil. *Chemosphere* **81**, 1263–71.
- Lundstedt, S., Haglund, P. & Oberg, L. (2003).** Degradation and formation of polycyclic aromatic compounds during bioslurry treatment of an aged gasworks soil. *Environ Toxicol Chem* **22**, 1413–20.
- Lundstedt, S., White, P. A., Lemieux, C. L., Lynes, K. D., Lambert, I. B., Oberg, L., Haglund, P. & Tysklind, M. (2007).** Sources, fate, and toxic hazards of oxygenated polycyclic aromatic hydrocarbons (PAHs) at PAH-contaminated sites. *Ambio* **36**, 475–85.
- Maidak, B. L., Cole, J. R., Parker, C. T., Garrity, G. M., Larsen, N., Li, B., Lilburn, T. G., McCaughey, M. J., Olsen, G. J. & other authors. (1999).** A new version of the RDP (Ribosomal Database Project). *Nucleic Acids Res* **27**, 171–3.
- Martin, F., Malagnoux, L., Violet, F., Jakoncic, J. & Jouanneau, Y. (2012).** Diversity and catalytic potential of PAH-specific ring-hydroxylating dioxygenases from a hydrocarbon-contaminated soil. *Appl Microbiol Biotechnol*.
- Martin, R. G., Jair, K. W., Wolf, R. E. & Rosner, J. L. (1996).** Autoactivation of the *marRAB* multiple antibiotic resistance operon by the MarA transcriptional activator in *Escherichia coli*. *J Bacteriol* **178**, 2216–23.
- Martineau, C., Whyte, L. G. & Greer, C. W. (2010).** Stable isotope probing analysis of the diversity and activity of methanotrophic bacteria in soils from the Canadian high Arctic. *Appl Environ Microbiol* **76**, 5773–84.
- Mason, J. R. & Cammack, R. (1992).** The electron-transport proteins of hydroxylating bacterial dioxygenases. *Annu Rev Microbiol* **46**, 277–305.
- Mathema, V. B., Thakuri, B. C. & Sillanpää, M. (2011).** Bacterial *mer* operon-mediated detoxification of mercurial compounds: a short review. *Arch Microbiol*

193, 837–44.

McDavid, A., Finak, G., Chattopadhyay, P. K., Dominguez, M., Lamoreaux, L., Ma, S. S., Roederer, M. & Gottardo, R. (2012). Data Exploration, Quality Control and Testing in Single-Cell qPCR-Based Gene Expression Experiments. *Bioinformatics*.

McFall, S. M., Chugani, S. A. & Chakrabarty, A. M. (1998). Transcriptional activation of the catechol and chlorocatechol operons: variations on a theme. *Gene* **223**, 257–67.

McLellan, S. L., Warshawsky, D. & Shann, J. R. (2002). The effect of polycyclic aromatic hydrocarbons on the degradation of benzo[a]pyrene by *Mycobacterium* sp. strain RJGII-135. *Environ Toxicol Chem* **21**, 253–259.

van der Meer, J. R., Tropel, D. & Jaspers, M. (2004). Illuminating the detection chain of bacterial bioreporters. *Environ Microbiol* **6**, 1005–20.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R. & other authors. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics* **9**, 386.

Miller, C. D., Hall, K., Liang, Y. N., Nieman, K., Sorensen, D., Issa, B., Anderson, A. J. & Sims, R. C. (2004). Isolation and characterization of polycyclic aromatic hydrocarbon-degrading *Mycobacterium* isolates from soil. *MicrobEcol* **48**, 230–238.

Mishra, V., Lal, R. & Srinivasan. (2001). Enzymes and operons mediating xenobiotic degradation in bacteria. *Crit Rev Microbiol* **27**, 133–66.

Miyakoshi, M., Urata, M., Habe, H., Omori, T., Yamane, H. & Nojiri, H. (2006). Differentiation of carbazole catabolic operons by replacement of the regulated promoter via transposition of an insertion sequence. *J Biol Chem* **281**, 8450–7.

Molina, M. C., González, N., Bautista, L. F., Sanz, R., Simarro, R., Sánchez, I. & Sanz, J. L. (2009). Isolation and genetic identification of PAH degrading bacteria from a microbial consortium. *Biodegradation* **20**, 789–800.

Mooney, R. A., Darst, S. A. & Landick, R. (2005). Sigma and RNA polymerase: an on-again, off-again relationship? *Mol Cell* **20**, 335–45.

Moser, R. & Stahl, U. (2001). Insights into the genetic diversity of initial dioxygenases from PAH-degrading bacteria. *Appl Microbiol Biotechnol* **55**, 609–618.

- Mumtaz, M. & George, J. (1995).** Toxicological profile for polycyclic aromatic hydrocarbons. *Atlanta (GA): Agency for Toxic Substances and Disease Registry, US Department of Health and Human Services.*
- Newman, D. L. & Shapiro, J. A. (1999).** Differential *flu-lacZ* fusion regulation linked to *Escherichia coli* colony development. *Mol Microbiol* **33**, 18–32.
- Ni Chadhain, S. M., Norman, R. S., Pesce, K. V., Kukor, J. J. & Zylstra, G. J. (2006).** Microbial dioxygenase gene population shifts during polycyclic aromatic hydrocarbon biodegradation. *Appl Environ Microbiol* **72**, 4078–4087.
- Nie, M., Yin, X., Ren, C., Wang, Y., Xu, F. & Shen, Q. (2010).** Novel rhamnolipid biosurfactants produced by a polycyclic aromatic hydrocarbon-degrading bacterium *Pseudomonas aeruginosa* strain NY3. *Biotechnol Adv* **28**, 635–43.
- Nikaido, H. (2003).** Molecular basis of bacterial outer membrane permeability revisited. *Microbiol Mol Biol Rev* **67**, 593–656.
- Nishi, A., Tominaga, K. & Furukawa, K. (2000).** A 90-kilobase conjugative chromosomal element coding for biphenyl and salicylate catabolism in *Pseudomonas putida* KF715. *J Bacteriol* **182**, 1949–55.
- Nogales, J., Macchi, R., Franchi, F., Barzaghi, D., Fernandez, C., Garcia, J. L., Bertoni, G. & Diaz, E. (2007).** Characterization of the last step of the aerobic phenylacetic acid degradation pathway. *Microbiology* **153**, 357–365.
- Nojiri, H., Habe, H. & Omori, T. (2001).** Bacterial degradation of aromatic compounds via angular dioxygenation. *J Gen Appl Microbiol* **47**, 279–305.
- Nzila, A. (2013).** Update on the cometabolism of organic pollutants by bacteria. *Environ Pollut* **178**, 474–82.
- Ono, A., Miyazaki, R., Sota, M., Ohtsubo, Y., Nagata, Y. & Tsuda, M. (2007).** Isolation and characterization of naphthalene-catabolic genes and plasmids from oil-contaminated soil by using two cultivation-independent approaches. *Appl Microbiol Biotechnol* **74**, 501–10.
- Pagnout, C., Frache, G., Poupin, P., Maunit, B., Muller, J. F. & Ferard, J. F. (2007).** Isolation and characterization of a gene cluster involved in PAH degradation in *Mycobacterium* sp. strain SNP11: expression in *Mycobacterium smegmatis* mc(2)155. *Res Microbiol* **158**, 175–186.

- Parales, R. E. & Harwood, C. S. (1993).** Regulation of the *pcaIJ* genes for aromatic acid degradation in *Pseudomonas putida*. *J Bacteriol* **175**, 5829–38.
- Park, H. H., Lee, H. Y., Lim, W. K. & Shin, H. J. (2005a).** NahR: effects of replacements at Asn 169 and Arg 248 on promoter binding and inducer recognition. *Arch Biochem Biophys* **434**, 67–74.
- Park, H. H., Lim, W. K. & Shin, H. J. (2005b).** In vitro binding of purified NahR regulatory protein with promoter *Psal*. *Biochim Biophys Acta* **1725**, 247–55.
- Park, H.-J. & Kim, E.-S. (2003).** An inducible *Streptomyces* gene cluster involved in aromatic compound metabolism. *FEMS Microbiol Lett* **226**, 151–7.
- Park, W., Padmanabhan, P., Padmanabhan, S., Zylstra, G. J. & Madsen, E. L. (2002).** *nahR*, encoding a LysR-type transcriptional regulator, is highly conserved among naphthalene-degrading bacteria isolated from a coal tar waste-contaminated site and in extracted community DNA. *Microbiology* **148**, 2319–29.
- Park, W., Peña-Llopis, S., Lee, Y. & Demple, B. (2006).** Regulation of superoxide stress in *Pseudomonas putida* KT2440 is different from the SoxR paradigm in *Escherichia coli*. *Biochem Biophys Res Commun* **341**, 51–6.
- Peng, J.-J., Cai, C., Qiao, M., Li, H. & Zhu, Y.-G. (2010a).** Dynamic changes in functional gene copy numbers and microbial communities during degradation of pyrene in soils. *Environ Pollut* **158**, 2872–9.
- Peng, R.-H., Xiong, A.-S., Xue, Y., Fu, X.-Y., Gao, F., Zhao, W., Tian, Y.-S. & Yao, Q.-H. (2008).** Microbial biodegradation of polyaromatic hydrocarbons. *FEMS Microbiol Rev* **32**, 927–55.
- Peng, R.-H., Xiong, A.-S., Xue, Y., Fu, X.-Y., Gao, F., Zhao, W., Tian, Y.-S. & Yao, Q.-H. (2010b).** A profile of ring-hydroxylating oxygenases that degrade aromatic pollutants. *Rev Environ Contam Toxicol* **206**, 65–94.
- Peng, Y., Leung, H. C. M., Yiu, S. M. & Chin, F. Y. L. (2012).** IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics* **28**, 1420–8.
- Pham, V. H. T. & Kim, J. (2012).** Cultivation of unculturable soil bacteria. *Trends Biotechnol* **30**, 475–84.
- Pinyakong, O., Habe, H. & Omori, T. (2003).** The unique aromatic catabolic genes in sphingomonads degrading polycyclic aromatic hydrocarbons (PAHs). *J Gen Appl*

Microbiol **49**, 1–19.

- Pinyakong, O., Habe, H., Kouzuma, A., Nojiri, H., Yamane, H. & Omori, T. (2004).** Isolation and characterization of genes encoding polycyclic aromatic hydrocarbon dioxygenase from acenaphthene and acenaphthylene degrading *Sphingomonas* sp. strain A4. *FEMS Microbiol Lett* **238**, 297–305.
- Pothier, J. F., Wisniewski-Dyé, F., Weiss-Gayet, M., Moëgne-Loccoz, Y. & Prigent-Combaret, C. (2007).** Promoter-trap identification of wheat seed extract-induced genes in the plant-growth-promoting rhizobacterium *Azospirillum brasilense* Sp245. *Microbiology* **153**, 3608–22.
- Prieto, M. A., Galán, B., Torres, B., Ferrández, A., Fernández, C., Miñambres, B., García, J. L. & Díaz, E. (2004).** Aromatic metabolism versus carbon availability: the regulatory network that controls catabolism of less-preferred carbon sources in *Escherichia coli*. *FEMS Microbiol Rev* **28**, 503–18.
- Qin, J., Li, R., Raes, J., Arumugam, M., Burgdorf, K. S., Manichanh, C., Nielsen, T., Pons, N., Levenez, F. & other authors. (2010).** A human gut microbial gene catalogue established by metagenomic sequencing. *Nature* **464**, 59–65.
- Quevillon, E., Silventoinen, V., Pillai, S., Harte, N., Mulder, N., Apweiler, R. & Lopez, R. (2005).** InterProScan: protein domains identifier. *Nucleic Acids Res* **33**, W116–20.
- Rabin, R. S. & Stewart, V. (1992).** Either of two functionally redundant sensor proteins, NarX and NarQ, is sufficient for nitrate regulation in *Escherichia coli* K-12. *Proc Natl Acad Sci USA* **89**, 8419–23.
- Rahman, P. K. S. M. & Gakpe, E. (2008).** Production, characterisation and applications of biosurfactants-Review. *Biotechnology* **7**, 360-70.
- Ramos, J. L., Marqués, S. & Timmis, K. N. (1997).** Transcriptional control of the *Pseudomonas* TOL plasmid catabolic operons is achieved through an interplay of host factors and plasmid-encoded regulators. *Annu Rev Microbiol* **51**, 341–73.
- Rediers, H., Rainey, P. B., Vanderleyden, J. & De Mot, R. (2005).** Unraveling the secret lives of bacteria: use of in vivo expression technology and differential fluorescence induction promoter traps as tools for exploring niche-specific gene expression. *Microbiol Mol Biol Rev* **69**, 217–61.

- Reis-Filho, J. S. & others. (2009). Next-generation sequencing. *Breast Cancer Res* **11**, S12.
- Rentz, J. A., Alvarez, P. J. J. & Schnoor, J. L. (2008). Benzo[a]pyrene degradation by *Sphingomonas yanoikuyae* JAR02. *Environ Pollut* **151**, 669–77.
- Retallack, D. M., Thomas, T. C., Shao, Y., Haney, K. L., Resnick, S. M., Lee, V. D. & Squires, C. H. (2006). Identification of anthranilate and benzoate metabolic operons of *Pseudomonas fluorescens* and functional characterization of their promoter regions. *Microb Cell Fact* **5**, 1.
- Roldan, M. D., Perez-Reinado, E., Castillo, F. & Moreno-Vivian, C. (2008). Reduction of polynitroaromatic compounds: the bacterial nitroreductases. *FEMS MicrobiolRev* **32**, 474–500.
- Rondon, M. R., Goodman, R. M. & Handelsman, J. (1999). The Earth's bounty: assessing and accessing soil microbial diversity. *Trends Biotechnol* **17**, 403–9.
- Rose, C. G. (2010). *Temporal changes in the microbial community of a PAH-contaminated soil during bench-top bioremediation*. (Master's thesis). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. MR65957).
- Ross, E. M., Moate, P. J., Bath, C. R., Davidson, S. E., Sawbridge, T. I., Guthridge, K. M., Cocks, B. G. & Hayes, B. J. (2012). High throughput whole rumen metagenome profiling using untargeted massively parallel sequencing. *BMC Genet* **13**, 53.
- Ross, W., Park, S. J. & Summers, A. O. (1989). Genetic analysis of transcriptional activation and repression in the Tn21 *mer* operon. *J Bacteriol* **171**, 4009–18.
- Ruby, J. G., Bellare, P. & Derisi, J. L. (2013). PRICE: Software for the Targeted Assembly of Components of (Meta) Genomic Sequence Data. *G3* **3**, 865–80.
- Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M. & other authors. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**, e77.
- Saito, A., Iwabuchi, T. & Harayama, S. (2000). A novel phenanthrene dioxygenase from *Nocardioides* sp. Strain KP7: expression in *Escherichia coli*. *J Bacteriol* **182**, 2134–2141.

- Samanta, S. K., Singh, O. V. & Jain, R. K. (2002).** Polycyclic aromatic hydrocarbons: environmental pollution and bioremediation. *Trends Biotechnol* **20**, 243–248.
- Sambrook, J. & Russell, D. W. (2001).** *Molecular cloning: a laboratory manual*. CSHL press.
- Santos, P. M. & Sá-Correia, I. (2007).** Characterization of the unique organization and co-regulation of a gene cluster required for phenol and benzene catabolism in *Pseudomonas* sp. M1. *J Biotechnol* **131**, 371–8.
- Sato, S. I., Nam, J. W., Kasuga, K., Nojiri, H., Yamane, H. & Omori, T. (1997).** Identification and characterization of genes encoding carbazole 1,9a-dioxygenase in *Pseudomonas* sp. strain CA10. *J Bacteriol* **179**, 4850–8.
- Schamfuß, S., Neu, T. R., van der Meer, J. R., Tecon, R., Harms, H. & Wick, L. Y. (2013).** Impact of mycelia on the accessibility of fluorene to PAH-degrading bacteria. *Environ Sci Technol* **47**, 6908–15.
- Schell, M. A. (1985).** Transcriptional control of the *nah* and *sal* hydrocarbon-degradation operons by the *nahR* gene product. *Gene* **36**, 301–9.
- Schell, M. A. & Poser, E. F. (1989).** Demonstration, characterization, and mutational analysis of NahR protein binding to *nah* and *sal* promoters. *J Bacteriol* **171**, 837–46.
- Schloss, P. D. & Handelsman, J. (2005).** Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *Genome Biol* **6**, 229.
- Schmeisser, C., Steele, H. & Streit, W. R. (2007).** Metagenomics, biotechnology with non-culturable microbes. *Appl Microbiol Biotechnol* **75**, 955–62.
- Selifonova, O., Burlage, R. & Barkay, T. (1993).** Bioluminescent sensors for detection of bioavailable Hg(II) in the environment. *Appl Environ Microbiol* **59**, 3083–90.
- Shapiro, H. M. (2003).** *Practical flow cytometry*. Wiley-Liss.
- Shepherd, J. M. & Lloyd-Jones, G. (1998).** Novel carbazole degradation genes of *Sphingomonas* CB3: sequence analysis, transcription, and molecular ecology. *Biochem Biophys Res Commun* **247**, 129–35.

- Shi, P., Jia, S., Zhang, X.-X., Zhang, T., Cheng, S. & Li, A. (2013). Metagenomic insights into chlorination effects on microbial antibiotic resistance in drinking water. *Water Res* **47**, 111–20.
- Shintani, M., Urata, M., Inoue, K., Eto, K., Habe, H., Omori, T., Yamane, H. & Nojiri, H. (2007). The *Sphingomonas* plasmid pCAR3 is involved in complete mineralization of carbazole. *J Bacteriol* **189**, 2007–20.
- Sho, M., Hamel, C. & Greer, C. W. (2004). Two distinct gene clusters encode pyrene degradation in *Mycobacterium* sp. strain S65. *FEMS Microbiol Ecol* **48**, 209–20.
- Shuttleworth, K. L. & Cerniglia, C. E. (1995). Environmental aspects of PAH biodegradation. *Appl Biochem Biotechnol* **54**, 291–302.
- Siggins, A., Gunnigle, E. & Abram, F. (2012). Exploring mixed microbial community functioning: recent advances in metaproteomics. *FEMS Microbiol Ecol* **80**, 265–80.
- Silby, M. W., Winstanley, C., Godfrey, S. A. C., Levy, S. B. & Jackson, R. W. (2011). *Pseudomonas* genomes: diverse and adaptable. *FEMS Microbiol Rev* **35**, 652–80.
- Simon, C. & Daniel, R. (2011). Metagenomic analyses: past and future trends. *Appl Environ Microbiol* **77**, 1153–61.
- Simon, M. J., Osslund, T. D., Saunders, R., Ensley, B. D., Suggs, S., Harcourt, A., Suen, W. C., Cruden, D. L., Gibson, D. T. & Zylstra, G. J. (1993). Sequences of genes encoding naphthalene dioxygenase in *Pseudomonas putida* strains G7 and NCIB 9816-4. *Gene* **127**, 31–7.
- Singleton, D. R., Ramirez, L. G. & Aitken, M. D. (2009). Characterization of a polycyclic aromatic hydrocarbon degradation gene cluster in a phenanthrene-degrading *Acidovorax* strain. *Appl Environ Microbiol* **75**, 2613–20.
- Singleton, D. R., Hu, J. & Aitken, M. D. (2012). Heterologous expression of polycyclic aromatic hydrocarbon ring-hydroxylating dioxygenase genes from a novel pyrene-degrading betaproteobacterium. *Appl Environ Microbiol* **78**, 3552–9.
- Smith, K. E. C., Thullner, M., Wick, L. Y. & Harms, H. (2009). Sorption to humic acids enhances polycyclic aromatic hydrocarbon biodegradation. *Environ Sci Technol* **43**, 7205–11.

- de Souza, N. (2013).** Genetics: Single-cell genetics. *Nature Methods* **10**, 820–820. Nature Publishing Group.
- Staley, J. T. & Konopka, A. (1985).** Measurement of in situ activities of nonphotosynthetic microorganisms in aquatic and terrestrial habitats. *Annu Rev Microbiol* **39**, 321–46.
- Stenglein, M. D., Sanders, C., Kistler, A. L., Ruby, J. G., Franco, J. Y., Reavill, D. R., Dunker, F. & Derisi, J. L. (2012).** Identification, characterization, and in vitro culture of highly divergent arenaviruses from boa constrictors and annulated tree boas: candidate etiological agents for snake inclusion body disease. *MBio* **3**, e00180–12.
- Stepanauskas, R. (2012).** Single cell genomics: an individual look at microbes. *Curr Opin Microbiol* **15**, 613–20.
- Stewart, V. & Parales, J. (1988).** Identification and expression of genes *narL* and *narX* of the *nar* (nitrate reductase) locus in *Escherichia coli* K-12. *J Bacteriol* **170**, 1589–97.
- Stingley, R. L., Khan, A. A. & Cerniglia, C. E. (2004a).** Molecular characterization of a phenanthrene degradation pathway in *Mycobacterium vanbaalenii* PYR-1. *Biochem Biophys Res Commun* **322**, 133–46.
- Stingley, R. L., Brezna, B., Khan, A. A. & Cerniglia, C. E. (2004b).** Novel organization of genes in a phthalate degradation operon of *Mycobacterium vanbaalenii* PYR-1. *Microbiology* **150**, 3749–61.
- Story, S. P., Parker, S. H., Hayasaka, S. S., Riley, M. B. & Kline, E. L. (2001).** Convergent and divergent points in catabolic pathways involved in utilization of fluoranthene, naphthalene, anthracene, and phenanthrene by *Sphingomonas paucimobilis* var. EPA505. *J Ind Microbiol Biotechnol* **26**, 369–82.
- Suenaga, H. (2012).** Targeted metagenomics: a high-resolution metagenomics approach for specific gene clusters in complex microbial communities. *Environ Microbiol* **14**, 13–22.
- Suenaga, H., Koyama, Y., Miyakoshi, M., Miyazaki, R., Yano, H., Sota, M., Ohtsubo, Y., Tsuda, M. & Miyazaki, K. (2009).** Novel organization of aromatic degradation pathway genes in a microbial community as revealed by metagenomic analysis. *ISME J* **3**, 1335–48.

- Takizawa, N., Kaida, N., Torigoe, S., Moritani, T., Sawada, T., Satoh, S. & Kiyohara, H. (1994).** Identification and characterization of genes encoding polycyclic aromatic hydrocarbon dioxygenase and polycyclic aromatic hydrocarbon dihydrodiol dehydrogenase in *Pseudomonas putida* OUS82. *J Bacteriol* **176**, 2444–9.
- Taupp, M., Mewis, K. & Hallam, S. J. (2011).** The art and design of functional metagenomic screens. *Curr Opin Biotechnol* **22**, 465–72.
- Tecon, R., Wells, M. & van der Meer, J. R. (2006).** A new green fluorescent protein-based bacterial biosensor for analysing phenanthrene fluxes. *Environ Microbiol* **8**, 697–708.
- Tittabutr, P., Cho, I. K. & Li, Q. X. (2011).** Phn and Nag-like dioxygenases metabolize polycyclic aromatic hydrocarbons in *Burkholderia* sp. C3. *Biodegradation* **22**, 1119–33.
- Torsvik, V., Daae, F. L., Sandaa, R. A. & Ovreås, L. (1998).** Novel techniques for analysing microbial diversity in natural and perturbed environments. *J Biotechnol* **64**, 53–62.
- Tropel, D. & van der Meer, J. R. (2004).** Bacterial transcriptional regulators for degradation pathways of aromatic compounds. *Microbiol Mol Biol Rev* **68**, 474–500.
- Turgeon, N., Laflamme, C., Ho, J. & Duchaine, C. (2006).** Elaboration of an electroporation protocol for *Bacillus cereus* ATCC 14579. *J Microbiol Methods* **67**, 543–8.
- Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S. & Banfield, J. F. (2004).** Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**, 37–43.
- Uchiyama, T. & Miyazaki, K. (2010).** Substrate-induced gene expression screening: a method for high-throughput screening of metagenome libraries. *Methods Mol Biol* **668**, 153–68.
- Uchiyama, T. & Watanabe, K. (2007).** The SIGEX scheme: high throughput screening of environmental metagenomes for the isolation of novel catabolic genes. *Biotechnol Genet Eng Rev* **24**, 107–16.

- Uchiyama, T. & Watanabe, K. (2008).** Substrate-induced gene expression (SIGEX) screening of metagenome libraries. *Nat Protoc* **3**, 1202–12.
- Uchiyama, T., Abe, T., Ikemura, T. & Watanabe, K. (2005).** Substrate-induced gene-expression screening of environmental metagenome libraries for isolation of catabolic genes. *Nat Biotechnol* **23**, 88–93.
- Ullrich, S. M., Tanton, T. W. & Abdrashitova, S. A. (2001).** Mercury in the aquatic environment: a review of factors affecting methylation. *Critical Reviews in Environmental Science and Technology* **31**, 241–293.
- Urata, M., Miyakoshi, M., Kai, S., Maeda, K., Habe, H., Omori, T., Yamane, H. & Nojiri, H. (2004).** Transcriptional regulation of the *ant* operon, encoding two-component anthranilate 1,2-dioxygenase, on the carbazole-degradative plasmid pCAR1 of *Pseudomonas resinovorans* strain CA10. *J Bacteriol* **186**, 6815–23.
- Urata, M., Uchimura, H., Noguchi, H., Sakaguchi, T., Takemura, T., Eto, K., Habe, H., Omori, T., Yamane, H. & Nojiri, H. (2006).** Plasmid pCAR3 contains multiple gene sets involved in the conversion of carbazole to anthranilate. *Appl Environ Microbiol* **72**, 3198–205.
- Uroz, S., Ioannidis, P., Lengelle, J., Cébron, A., Morin, E., Buée, M. & Martin, F. (2013).** Functional Assays and Metagenomic Analyses Reveals Differences between the Microbial Communities Inhabiting the Soil Horizons of a Norway Spruce Plantation. *PLoS ONE* **8**, e55929.
- Uyttebroek, M., Spoden, A., Ortega-Calvo, J. J., Wouters, K., Wattiau, P., Bastiaens, L. & Springael, D. (2007).** Differential responses of eubacterial, *Mycobacterium*, and *Sphingomonas* communities in polycyclic aromatic hydrocarbon (PAH)-contaminated soil to artificially induced changes in PAH profile. *J Environ Qual* **36**, 1403–1411.
- Vagner, V., Dervyn, E. & Ehrlich, S. D. (1998).** A vector for systematic gene inactivation in *Bacillus subtilis*. *Microbiology* **144** (Pt 11), 3097–104.
- Vaillancourt, F. H., Bolin, J. T. & Eltis, L. D. (2006).** The ins and outs of ring-cleaving dioxygenases. *Crit Rev Biochem Mol Biol* **41**, 241–267.
- Valentini, M. & Lapouge, K. (2012).** Catabolite repression in *Pseudomonas aeruginosa* PAO1 regulates the uptake of C(4)-dicarboxylates depending on succinate concentration. *Environ Microbiol*.

- Vartoukian, S. R., Palmer, R. M. & Wade, W. G. (2010). Strategies for culture of “unculturable” bacteria. *FEMS Microbiol Lett* **309**, 1–7.
- Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E. & other authors. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science* **304**, 66–74.
- Vinas, M., Sabate, J., Espuny, M. J. & Solanas, A. M. (2005). Bacterial community dynamics and polycyclic aromatic hydrocarbon degradation during bioremediation of heavily creosote-contaminated soil. *Appl Environ Microbiol* **71**, 7008–7018.
- Vogel, T. M., Simonet, P., Jansson, J. K., Hirsch, P. R., Tiedje, J. M., Van Elsas, J. D., Bailey, M. J., Nalin, R., Philippot, L. & others. (2009). TerraGenome: a consortium for the sequencing of a soil metagenome. *Nature Reviews Microbiology* **7**, 252.
- Wackett, L. P. (2012). Annotation for environmental metagenomes. *Environmental Microbiology* **14**, 3066–3067. Wiley Online Library.
- Wang, X. & Quinn, P. J. (2010). Lipopolysaccharide: Biosynthetic pathway and structure modification. *Prog Lipid Res* **49**, 97–107.
- Watanabe, T., Fujihara, H. & Furukawa, K. (2003). Characterization of the second LysR-type regulator in the biphenyl-catabolic gene cluster of *Pseudomonas pseudoalcaligenes* KF707. *J Bacteriol* **185**, 3575–82.
- Wexler, M. & Johnston, A. W. (2010). Wide host-range cloning for functional metagenomics. *Methods Mol Biol* **668**, 77–96.
- Whynot, C. (2009). *The efficacy of different bioremediation strategies in removing mutagenic hazard from contaminated soil.* (Master’s thesis). Retrieved from ProQuest Dissertations and Theses. (Accession Order No. MR47539).
- Wilson, K. (1997). Preparation of Genomic DNA from Bacteria. *Current Protocols in Molecular Biology* **2**, 1–2.
- Winsley, T., van Dorst, J. M., Brown, M. V. & Ferrari, B. C. (2012). Capturing greater 16S rRNA gene sequence diversity within the domain Bacteria. *Appl Environ Microbiol* **78**, 5938–41.
- Wolfe, M. F., Schwarzbach, S. & Sulaiman, R. A. (1998). Effects of mercury on wildlife: a comprehensive review. *Environmental Toxicology and Chemistry* **17**,

146–160. Wiley Online Library.

- Wommack, K. E., Bhavsar, J. & Ravel, J. (2008).** Metagenomics: read length matters. *Appl Environ Microbiol* **74**, 1453–63.
- Wösten, M. M. (1998).** Eubacterial sigma-factors. *FEMS Microbiol Rev* **22**, 127–50.
- Xia, Y., Ju, F., Fang, H. H. P. & Zhang, T. (2013).** Mining of novel thermo-stable cellulolytic genes from a thermophilic cellulose-degrading consortium by metagenomics. *PLoS ONE* **8**, e53779.
- Yamaguchi, A., Tamang, D. G. & Saier, M. H. (2007).** Mercury transport in bacteria. *Water, Air, & Soil Pollution* **182**, 219–234. Springer.
- Yergeau, E., Arbour, M., Brousseau, R., Juck, D., Lawrence, J. R., Masson, L., Whyte, L. G. & Greer, C. W. (2009).** Microarray and real-time PCR analyses of the responses of high-arctic soil bacteria to hydrocarbon pollution and bioremediation treatments. *Appl Environ Microbiol* **75**, 6258–67.
- Yergeau, E., Sanschagrín, S., Beaumier, D. & Greer, C. W. (2012a).** Metagenomic analysis of the bioremediation of diesel-contaminated Canadian high arctic soils. *PLoS ONE* **7**, e30058.
- Yergeau, E., Lawrence, J. R., Sanschagrín, S., Waiser, M. J., Korber, D. R. & Greer, C. W. (2012b).** Next-generation sequencing of microbial communities in the Athabasca River and its tributaries in relation to oil sands mining activities. *Appl Environ Microbiol* **78**, 7626–37.
- Yu, K. & Zhang, T. (2012).** Metagenomic and metatranscriptomic analysis of microbial community structure and gene expression of activated sludge. *PLoS ONE* **7**, e38183.
- Yun, J. & Ryu, S. (2005).** Screening for novel enzymes from metagenome and SIGEX, as a way to improve it. *Microb Cell Fact* **4**, 8.
- Zhang, C. & Anderson, A. J. (2013).** Utilization of pyrene and benzoate in *Mycobacterium* isolate KMS is regulated differentially by catabolic repression. *J Basic Microbiol* **53**, 81–92.
- Zhou, F. & Xu, Y. (2010).** cBar: a computer program to distinguish plasmid-derived from chromosome-derived sequence fragments in metagenomics data. *Bioinformatics* **26**, 2051–2.

Zhu, W., Lomsadze, A. & Borodovsky, M. (2010). *Ab initio* gene identification in metagenomic sequences. *Nucleic Acids Res* **38**, e132.

Zylstra, G. J., Wang, X. P., Kim, E. & Didolkar, V. A. (1994). Cloning and analysis of the genes for polycyclic aromatic hydrocarbon degradation. *Ann N Y Acad Sci* **721**, 386–98.