Metabolic analysis of the soil microbe *Dechloromonas aromatica* str. RCB: indications of a surprisingly complex life-style and cryptic anaerobic pathways for aromatic degradation

Kennan Kellaris Salinero [1§], Keith Keller[2], William S. Feil[1], Helene Feil[1], Stephan Trong[3], Genevieve Di Bartolo[3], and Alla Lapidus[3]

[1]University of California, Berkeley, CA, USA

[2]Virtual Institute of Microbial Stress and Survival, Berkeley, CA, 94710, USA

[3]DOE Joint Genome Institute, Walnut Creek, CA 94598, USA

[§]Corresponding author

Email address:

KKS: kellarkv@nature.berkeley.edu

Background

Initial interest in *Dechloromonas aromatica* strain RCB arose from its ability to anaerobically degrade benzene. It is also able to reduce perchlorate and oxidize chlorobenzoate, toluene, and xylene, creating interest in using this organism for bioremediation. Little physiological data has been published for this microbe. It is considered to be a free-living organism.

Results

The *a priori* prediction that the *D. aromatica* genome would contain previously characterized "central" enzymes involved in anaerobic aromatic degradation proved to be false, suggesting the presence of novel anaerobic aromatic degradation pathways in this species. These missing pathways include the benzyl succinyl synthase (*bss*ABC) genes (responsible for formate addition to toluene) and the central benzoylCoA pathway for monoaromatics. In depth analyses using existing TIGRfam, COG, and InterPro models, and the creation of *de novo* HMM models, indicate a highly complex lifestyle with a large number of environmental sensors and signaling pathways, including a relatively large number of GGDEF domain signal receptors and multiple quorum sensors. A number of proteins indicate interactions with an as yet unknown host, as indicated by the presence of predicted cell host remodeling enzymes, effector enzymes, hemolysin-like proteins, adhesins, NO reductase, and both type III and type VI secretory complexes. Evidence of biofilm formation including a proposed exopolysaccharide complex with the somewhat rare exosortase (epsH), is also present. Annotation described in this paper also reveals evidence for several metabolic pathways that have yet to be observed experimentally, including a sulphur oxidation (soxFCDYZAXB) gene cluster, Calvin cycle enzymes,

and nitrogen fixation (including RubisCo, ribulose-phosphate 3-epimerase, and nif gene families, respectively).

Conclusions

Analysis of the *D. aromatica* genome indicates there is much to be learned regarding the metabolic capabilities, and life-style, for this microbial species. Examples of recent gene duplication events in signaling as well as dioxygenase clusters are present, indicating selective gene family expansion as a relatively recent event in *D. aromatica*'s evolutionary history. Gene families that constitute metabolic cycles presumed to create *D. aromatica*'s environmental 'foot-print' indicate a high level of diversification between its predicted capabilities and those of its close relatives, *A. aromaticum* str EbN1 and *Azoarcus* BH72.

Background

D. aromatica strain RCB is a gram negative Betaproteobacterium found in soil environments [1-6]. Other members of the Betaproteobacteria class are found in environmental samples (such as soil and sludge) or are pathogens (such as *Ralstonia solanacearum* in plants and *Neisseria meningitidis* in humans) and in general the genus *Dechloromonas* has been found to be ubiquitous in the environment.

A facultative anaerobe, *D. aromatica* was initially isolated from Potomac River sludge contaminated with BTEX compounds (benzene, toluene, ethylbenzene and xylene) based on its ability to anaerobically degrade chlorobenzoate [3]. This microbe is capable of aromatic hydrocarbon degradation and perchlorate reduction, and can oxidize Fe(II) and H_2S [6]. Although several members of the Rhodocyclales group of Betaproteobacteria are of interest to the scientific community due to their ability to anaerobically degrade derivatives of benzene, *D. aromatica* was the first

pure culture capable of anaerobic degradation of the stable underivitized benzene molecule to be isolated [3]. This, along with its ability to reduce perchlorate (a teratogenic contaminant introduced into the environment by man) and inquiry into its use in biocells [1] has led to interest in using this organism for bioremediation and energy production. Since the isolation of *D. aromatica*, other species of *Azoarcus* have been found to possess the ability to anaerobically degrade benzene, but have not been genomically sequenced [7].

The pathway for anaerobic benzene degradation has been partially deduced [5], but the enzymes responsible for this process have yet to be identified, and remain elusive even after the intensive annotation efforts described here-in. Conversely, central anaerobic pathways for aromatic compounds described in various other species were not found to be present in this genome [8].

Methods

Sequencing

Three libraries (3kb, 8kb and 30kb) were generated by controlled shearing (Hydroshear, Genomic Solutions, Ann Arbor, MI) of spooled genomic DNA isolated from *D. aromatica* strain RCB and inserted into pUC18, pCUGIblu21, and pcc1Fos vectors, respectively. Clonal DNA was amplified using rolling circular amplification (http://www.jgi.doe.gov/) and sequenced on ABI 3700 capillary DNA sequencers (Applied Biosystems, Foster City, CA) using BigDye technology (Perkin Elmer Corporation, Waltham, MA). Paired end-reads [9] were used to aid in assembly, and proved particularly useful in areas of repeats.

The Phrap algorithm [10, 11] was used for initial assembly. Finishing and manual curation was conducted on CONSED v14 software [12], supplemented with a suite of finishing analysis tools provided by the Joint Genome Institute. *In silico*

cross-over errors were corrected by manual creation of fake reads to guide the assembly by forcing the consensus to follow the correct path.

Gaps were closed through a combination of primer walks on the gap-spanning clones from the 3 and 8kb libraries (identified by paired-end analysis in the CONSED software) as well as sequencing of mapped, unique PCR products from freshly prepared genomic DNA.

The final step required to create a finished single chromosomal sequence was to establish the number of tandem repeats for a 672 base DNA sequence of unknown length. This was done by creating the full tandem repeat insert from unique upstream and downstream primers using long-range PCR. We then determined the size of product (amplified DNA) between the unique sequences.

Protein sequence predictions/orfs

Annotation done at Oak Ridge National Laboratory consisted of gene calls using CRITICA [13], glimmer [14], and Generation (http://compbio.ornl.gov/generation/index.shtml). Annotation at the Virtual Institute for Microbial Stress and Survival (http://www.microbesonline.org) used bidirectional best hits as well as recruitment to TIGRfam hidden Markov models (HMMs), as described in Alm et al. [15]. Briefly, protein coding predictions derived from NCBI, or identified using CRITICA, with supplemental input from Glimmer, were analyzed for domain identities using the models deposited in the InterPro, UniProt, PRODOM, Pfam, PRINTS, SMART, PIR SuperFamily, SUPERFAMILY, and TIGRfam databases [15]. Orthologs were identified using bidirectional unique best hits with greater than 75% coverage. RPS-BLAST against the NCBI COGs (Clusters of

Orthologous Genes) in the CDD database were used to assign proteins to COG models when the best hit E-value was $<1e^{-5}$ and coverage was >60%.

Manual curation

Each and every predicted protein in the VIMSS database [15] was assessed to compare insights obtained from recruitment to models from the various databases (TIGRfams, COGs, EC and InterPro). Assignments that offered the most definitive functional assignment were captured in an excel spreadsheet with data entries for all proteins predicted in the VIMSS database. Extensive manual curation of the predicted protein set was carried out using a combination of tools including the VIMSS analysis tools, HMMs, and phylogenomic analysis, as described below. Changes in gene functional predictions and naming were captured in the excel spreadsheet, and predictions with strong phylogenetic evidence of function posted using the interactive VIMSS web-based annotation interface.

Phylogenomic analysis: Flower Power , SCI PHY and HMM scoring

Phylogenomic profiling was done to derive functional assignments based on near-neighbors or to confirm the absence of a given protein in the *D. aromatica* genome. Four basic approaches were used (see Fig. 1). In the first, we employed the Flower Power and SCI-PHY (Subfamily Classification in Phylogenomics) utilities from the UCBerkeley web server (http://phylogenomics.berkeley.edu [16, 17]) to profile enzymes specifically from the near-neighbor *A. aromaticum* EbN1in creating Hidden Markov Models (HMMs) that could be scored against the *D. aromatica* protein set. This approach allowed us to determine, with extremely high confidence, whether a protein is truly not present in the *D. aromatica* genome, whether or not

automatic annotation had identified it correctly. The same protocols were employed as described in the following section, except for the selection of seed sequences. Phylogenetic trees were assessed by comparing included species with experimentally published orthologs, and in some cases, HMMs were scored against the full set of *D. aromatica* proteins (VIMSS annotation set) in order to compare relative probability scores for individual proteins in the genome [17].

The second use of HMM modeling and phylogenomic tree-building was used to create models for proteins of interest, with known function, to determine whether orthologs are present in the *D. aromatica* genome. Protein models were created based on candidates having enzymatic function related to metabolic function for *D. aromatica* (eg benzylsuccinate synthetase), or for proteins that were suggested by clusters of *D. aromatica* predictions displaying pfam, COG, EC or TIGRfam annotations indicating the possibility of pathways established in other organisms (eg the *SOX* cluster). For aromatic degradation enzymes, the BRENDA database (http://www.brenda.uni-koeln.de/) [18, 19] was used to select protein sequences for enzymes of known function based on experimental evidence. A characterized protein was used to seed a Flower Power HMM-based recruitment of all related sequences from the Genbank non-redundant protein set, requiring an identity to existing sequences ranging from 0.15 to 0.19 (0.18 most typically produced a robust recruitment set while retaining specificity), and requiring global alignment. Species represented in the recruited set were compared to all species characterized as having that enzyme function, as captured in BRENDA, to assess overall coverage of known enzyme candidates. Often, more than one enzyme was found to be responsible for the same enzymatic process. Therefore, if deemed appropriate, a second Flower Power sequence recruitment was conducted using a seed sequence not captured in the first

set of homologous proteins, to create a second HMM model and phylogenetic tree profile for the isozymic set. If the HMM model was to be scored against the full set of *D. aromatica* proteins, alignments were edited using the Belvu alignment editing tool [20] using the methods of Brown et al. [21]. The resulting modified alignment was used to create HMMs using the w0.5 build and hmmscore utilities of UCSC as employed by K. Sjölander [22-26]. HMMs were then scored against the complete set of predicted proteins from *D. aromatica*.

The third approach was to carry out *de novo* generation of HMMs for protein sequences of interest that were not adequately described by either TIGRfam or COGs models, or for which no models were available. Proteins or enzymes potentially involved in the metabolic pathways or cell processes of particular interest from the *D. aromatica* VIMSS protein set were used as seed sequences for Flower Power recruitment of phylogenetically related proteins in the Genbank non-redundant data set. In a few instances, Flower Power alignments were used as input for the SCI PHY utility, and then uploaded into PhyloFacts [27, 28]. Subfamilies generated using the SCI PHY minimum-encoding-cost criteria were viewed in phylogenetic trees, and the functionality of the protein of interest was inferred based on clade (sub-family) membership, ideally from the experimentally supported functionality of other proteins within the same clade [25, 26, 29]. During the course of this study, TIGRfams 7.0 was released, which contained several models that replicated ones generated during this study. In all cases where HMM modelling was used, their annotation predictions agreed with ours.

In the fourth approach using HMM models, internal clustering of all *D. aromatica* proteins (using the VIMSS protein set) was employed to create a set of paralogous proteins within the *D. aromatica* genome. Clustering of proteins within

the *D. aromatica* genome was carried out using an internal suite of computational tools as employed by K. Sjölander, UCBerkeley [16]. This analysis resulted in the identification of full-length paralogs within the *D. aromatica* genome. Comparison of sets of paralogs allowed identification of putative enzymes forming sequential steps in a given catabolic pathway, many of which display physical proximity along the chromosome (eg the mhp families of proteins). It also allowed identification of smaller subunits in multi-enzyme complexes, which were often missed by the high-throughput annotation pipelines employed.

Gene Family Expansion

The clustered set of predicted proteins (putative paralogs) was also used as a candidate set to search for protein paralogs that are candidates for recent gene duplication events. After an initial assessment of the VIMSS gene information/homolog data, candidate proteins were used as seed sequences for Flower Power and internal tree-viewing tools or SCI-PHY analyses. These two approaches employed the neighbor-joining trees using the Scoredist correction setting in the Belvu alignment editor, or the SCI-PHY utility and tree viewer. In either case resulting phylogenomic tree builds were assessed for contiguous protein alignments with two or more proteins from *D.* aromatica. If the most similar homolog (% amino acid identity between aligned sequences) was also from the *D. aromatica* genome, the protein set is considered to be a candidate for a gene duplication event, either in the *D. aromatica* genome or in a predecessor species.

Resequencing to verify absence of plasmid structure

After finishing the *D. aromatica* genome, analysis of the annotated gene set revealed the notable absence of several anaerobic aromatic degradation pathways that were expected to be present, due to their presence in *A. aromaticum* EbN1 (an evolutionary near-neighbor, as determined by 16sRNA phylogeny). Because many catabolic pathways are encoded on plasmid DNA, we felt it was important to preclude this possibility. We re-isolated DNA from a clonal preparation of *D. aromatica* that experimentally supported anaerobic benzene degradation, using three different plasmid purification protocols, each based on different physical parameters. All three generated a single band of DNA. The protocol that generated the highest yield of DNA was used to create a complete, new library of 2kb inserts, and the library was submitted to sequence analysis using the protocols previously cited.

Results

Overview of Gene and Protein Features

The finished sequence for *D. aromatica* reveals a single circular, closed chromosome of 4,501,104 nucleotides created from 130,636 screened reads, with an average G+C content of 60% and an extremely high level of sequence coverage (average depth of 24 reads/base; see Table 1). Specific probing for plasmids confirmed no plasmid structure was present in the clonal species sequenced, which supports anaerobic benzene degradation. It is noted however that the presence of two *tra* clusters (putative conjugal transfer genes; VIMSS582582-582597 and VIMSS582865-582880), as well as plasmid partitioning proteins, indicates this microbial species is likely to be transformationally competent and thus likely to be able to support plasmid DNA structures.

The Virtual Institute for Microbial Stress and Survival (VIMSS, http://www.microbesonline.org) and the Joint Genome Institute (http://genome.jgi-psf.org/finished_microbes/decar/decar.home.html) report 4170 and 4204 protein coding genes, respectively (Table 2). Cross-database comparisons to assure the highest probability of capturing candidate orfs for analysis were made. The majority of proteins are shared between data sets. Variations in N-termini start sites were noted, both between JGI and VIMSS datasets and between initial and later annotation runs (approximately 200 N-termini differences between four runs of orf predictions were noted for the initial two annotation runs, Joint Genome Institutes, done at Oak Ridge National Laboratories – ORNL, and VIMSS).

The most definitive functional classification, TIGRfams, currently covers 33% of predicted proteins, leaving 2802 genes with no TIGRfam classification (see Table 2). Many proteins in the non-covered set were investigated further using K. Sjölander's HMM building protocols (many of which are available at http://phylogenomics.berkeley.edu), to supplement TIGRfams. The Clusters of Orthologous Genes (COG) assignments were used for classification in the families of signaling proteins, but specific function predictions for these proteins also required further analyses. The metabolic and signaling pathways are discussed below, and the identity of orthologs within these pathways are based on analysis of phylogenomic profiles of clusters obtained by HMM analysis, with comparison to proteins having experimentally defined function.

Anaerobic aromatic degradation – absence of known enzymes indicates novel pathways

Given that one of the primary metabolic capabilities of interest for this microbe is anaerobic degradation of benzene, we first sought to identify key enzymes

for anaerobic degradation of monoaromatic compounds in the *D. aromatica* genome Fumarate addition to toluene via benzylsuccinate synthase is recognized as the common mechanism for anaerobic degradation by a phylogenomically diverse population of microbes [30-32] and has been called “the paradigm of anaerobic hydrocarbon oxidation”[33]. Benzoyl CoA is likewise considered a central intermediate in anaerobic degradation, and is further catabolized via benzoyl CoA reductase (BcrAB) [33]. In *A. aromaticum* EbN1, ten major catabolic pathways have been found for anaerobic aromatic degradation, and nine of the ten converge on benzoyl-CoA [34].

To explore this difference between the two genomes further, all characterized anaerobic aromatic degradation pathways from near-neighbor *A. aromaticum* EbN1 [35] were defined by HMMs to establish presence or absence of proteins in both the *D. aromatica* and *Azoarcus* BH72 genomes (see Fig1A for methodology). A key catalytic enzyme or subunit for each enzymatic step was used as a seed sequence to recruit proteins from a non-redundant set of Genbank proteins for phylogenetic analysis. Benzylsuccinyl synthase, present in *A. aromaticum* EbN1 [35, 36] as well as *Thauera aromatica* [8], and *Geobacter metallireducens* [37], is surprisingly absent from the genomes of both *D. aromatica* and *Azoarcus* BH72 (see Table 3). *D. aromatica* does encode a protein in the pyruvate formate lyase family, but further analysis shows that it is more closely related to the *E. coli* homolog of this protein (which is not involved in aromatic catabolism) than to BssA.

For all pathways except the ubiquitous phenylacetic acid catabolic cluster, which is involved in the aerobic degradation of phenylalanine, and the PpcAB phenylphosphate carboxylase enzymes involved in benzoate degradation, all key anaerobic aromatic degradation proteins present in *A. aromaticum* EbN1 are missing

from the *D. aromatica* genome (Table 3), and the majority are also not present in *Azoarcus* BH72. The lack of overlap for genes encoding anaerobic aromatic enzymes between these two species was completely unexpected, as both *A. aromaticum* EbN1 and *D. aromatica* are metabolically diverse degraders of aromatic compounds. In general *Azoarcus* BH72 appears to share many families of proteins with *D. aromatica* that are not present in *A. aromaticum* EbN1 (eg signaling proteins, noted below).

The extremely high divergence of encoded protein families in this functional grouping differs from the general population of central metabolic and housekeeping genes: *Azoarcus* BH72, *Azoarcus aromaticum* EbN1 and *D. aromatica* are evolutionarily near-neighbors within currently sequenced genomes, as defined both by the high level of protein homology within house-keeping genes (defined by the COG J family of proteins), and 16sRNA sequence. *Azoarcus* BH72 and *A. aromaticum* EbN1 display the highest percent homology between housekeeping proteins within this triad, with 138 of the 156 COG J proteins in *A. aromaticum* EbN1 displaying highest similar to their BH72 counterparts. On average these two genomes display 83.5% amino acid identity across shared COG J proteins. *D. aromatica* is an outlier in the triad, with slightly higher similarity to *Azoarcus* BH72 than *A. aromaticum* EbN1 (43 of *D. aromatica*'s 169 COG J proteins are most homolous to *A. aromaticum* EbN1 orthologs with an average 71% identity, and 67 are most homologous to *Azoarcus* BH72 with an average 72% identity).

Aerobic aromatic degradation

D. aromatica encodes several aerobic pathways for aromatic degradation, including six groups of oxygenase clusters that each share a high degree of sequence similarity to the phenylpropionate and phenol degradation (Hpp, Mhp) pathways in

Comamonas species[38, 39]. The *mhp* genes of *E. coli* and *Comamonas* are involved in catechol and protocatechuate pathways for aromatic degradation via hydroxylation, oxidation, and subsequent ring cleavage of the dioxygen species. Only one of these clusters encodes an *mhpA*-like gene; it begins with VIMSS584143 MhpC, and is composed of orthologs of MhpABCDEF&R, and is in the same overall order and orientation as the *Comamonas* cluster as well as the *E. coli mhp* gene families [40]. These pathways are also phylogenomically related to the biphenyl/polychlorinated biphenyl (Bhp) degradation pathways in *Pseudomonad* species [40]. For *Comamonas testosteroni*, this pathway is thought to be associated with lignin degradation [39]. Hydroxyphenyl propionate (HPP), an alkanoic acid of phenol, is the substrate for Mhp, and is also produced by animals in the digestive breakdown of polyphenols found in seed components [41]. Each gene cluster appears to represent a multi-component pathway, and is made up of five or more of various combinations of dioxygenase, hydroxylase, aldolase, dehydrogenase, hydratase, decarboxylase and thioesterase enzymes.

Interestingly, the single predicted MhpA protein in *D. aromatica* (VIMSS584155), which is predicted to support an initial hydroxylation of a substituted phenol substrate, shares 64.4% identity to *Rhodococcus* OhpB 3-(2-hydroxyphenyl) propionate monooxygenase (GI:8926385) vs. 26.4% for *Comamonas testosteroni* (GI:5689247), yet the remainder of the *ohp* genes in the *Rodococcus ohp* clade do not share synteny with the *D. aromatica mhp* gene cluster.

Other aromatic oxygenases

Two chromosomally adjacent monooxygenase clusters, syntenic to genes found in *Burkholderia* and *Ralstonia* spp, indicate that *D. aromatica* might have broad

substrate hydroxylases that support the degradation of toluene, vinyl chlorides, and TCE.

One monooxygenase gene cluster, composed of VIMSS581514 to 581519, is orthologous to the *tbuA1UBVA2C*/*tmoAECDBF/touABCDEF/phlKLMNOP* and *tbc2ABCDEF* gene families (from *P. stutzeri, R. pickettii*, and *Burkholderia* JS150; see Table 4). This gene cluster includes a transport protein that is orthologous to TbuX/TodX/XylN (VIMSS581520). Specificity for the initial monooxygenase is not known, but phylogenetic analysis places VIMSS581514 monooxygenase with near-neighbors TbhA [42], reported as a toluene and aliphatic carbohydrate monooxygenase (76.5% sequence identity), and BmoA [43], a benzene monooxygenase of low regiospecificity (79.6% sequence identity). The region is also highly syntenic with, and homologous to, the *tmo*AECDBF (AY552601) gene cluster responsible for *P. mendocina*'s ability to utilize toluene as a sole carbon and energy source [44].

Just downstream on the chromosome is a *phc/dmp/phh/phe/aph*-like cluster of genes, composed of the genes VIMSS812947 and VIMSS 581535 to 581540. Overall, chromosomal organization is somewhat different for *D. aromatica* as compared to *Ralstonia* and *Burkholderia*. *D. aromatica* has a seventeen gene insert that encodes members of the *mhp*-like family of aromatic oxygenases between the tandem oxygenase clusters (see Fig. 4), with an inversion of the second region compared to *R. eutropha* and *Burkholderia*. Clade analysis indicates a broad substrate phenol degradative pathway in this cluster, with high sequence identity to the TOM gene cluster of *Bradyrhizobium*, which has the ability to oxidize dichloroethylene, vinyl chlorides, and TCE [45, 46]. The VIMSS581522 response regulator gene that occurs between the two identified monooxygenase gene clusters shares 50.3% identity to the

Thaurea aromatica tutB gene and 48.2% to the *Pseudomonas* sp. Y2 styrene response regulator (identified by phylogenetic analysis). VIMSS581522 is likely to be involved in the chemotactic response in conjunction with VIMSS581521 (histidine kinase) and VIMSS581523 (methyl accepting chemotaxis protein), which would confer the ability to display a chemotactic response to aromatic compounds.

Overall, several mono- and di-oxygenases were found in the genome, indicating *D. aromatica* has diverse abilities in the aerobic oxidation of heterocyclic compounds.

There are several gene clusters indicative of benzoate transport and catabolism. All recognized pathways are aerobic. The benzoate dioxygenase cluster BenABCDR is encoded in VIMSS582483-582487, and is very similar to (and clades with) the xylene degradation (*xyl*XYZ) cluster of *Pseudomonas*, as well as gene clusters in *Azotobacter vinelandii*, *R. eutropha*, *Burkholderia*, and *Corynebacterium glutamicum*.

There is also an *hcaA* oxygenase gene cluster, embedded in one of the *mhp* clusters (see Fig. 3). Specificity of the large subunit of the dioxygenase (VIMSS582049) appears to be most likely for a bicyclic aromatic compound, as it shows highest identity to dibenzothiophene and naphthalene dioxygenases.

Dechloromonas aromatica's sensitivity to the environment

Cell Signaling

D. aromatica has a large number of genes involved in signaling pathways, with 383 predicted signaling proteins categorized in COG T (signal transduction mechanisms) and a total of 395 proteins (nearly 10% of the genome) either recruited to COG T or possessing annotated signal transduction domains. Signaling appears to be an area that has undergone recent gene expansion, as twelve recent gene

duplication events in this functional group are predicted by phylogenetic analysis, as described in a later section.

Complex lifestyles are implicated in large genomes with diverse signaling capability, and in general genomes with a very large number of annotated open reading frames (orfs) have high numbers of predicted signal transducing proteins, as shown in Fig. 5, though some species, such as *Rhodococcus* RHA1 and *Psychroflexus torques* are notable exceptions to this trend. However, assessment of COG T population size relative to other genomes with a similar number of predicted orfs (Fig. 5) indicates that *D. aromatica* is one of a handful of species that have a large relative number of signaling proteins vs similarly sized genomes. Other organisms displaying this characteristic include *Magnetospirillum magnetotacticum* MS-1, *Stigmatella aurantiaca, Myxococcus Xanthus* DK1622, *Magnetospirillum Magneticum* AMB-1, *Ocenospririllum sp.* MED92, *Hahella chejuensis* KCTC2396 and *Desulfuromonas acetoxidans*. Within the Betaproteobacteria, *Chromobacterium violaceum* and *Thiobacillus denitrificans* have a relatively large number of signaling cascade genes, but still have far fewer than found in *D. aromatica*, with 308 predicted COG T proteins (7% of the genome) and 158 COG T proteins (5.6% of the genome), respectively predicted in these two microbial species, vs the 383 noted in *D. aromatica*.

Histidine kinase encoding proteins are particularly well-represented, with 91 predicted histidine kinase proteins, including a large number of nitrate/nitrogen responsive elements. Furthermore, the presence of 47 putative histidine kinases predicted to contain two transmembrane (TM) domains, likely to encode membrane-bound sensors (see Fig. 2), suggests that *D. aromatica* is likely to be highly sensitive to environmental signals. Nearly half (48%) of the predicted histidine kinases are

contiguous to a putative response regulator on the chromosomal DNA, indicating they likely constitute functionally expressed kinase/response regulator pairs. This is atypically high for contiguous placement on the chromosome vs most other genomes assessed, but is still less than the number reported for *E. coli*, where 26 or the 29 histidine kinases are organized contiguously with response regulators [47].

A relatively high level of diguanylate cyclase (GGDEF domain [48-50]) signaling capability is implied in *D. aromatica* by the presence of 57 proteins encoding a GGDEF domain (Interpro IPR000160) or a GGDEF response regulator (COG1639) [48]. *E. coli*, for comparison, encodes 19. This gene family also appears to have undergone recent expansion in this microbe's evolutionary history. Microbes having a large number of proteins or even a diverse array of COG T elements do not *a priori* encode a large number of GGDEF elements, as *Stigmatella aurantiaca*, *Myxococcus*, *Xanthus* DK1622 and *Burkholderia pseudomallei* 668, by contrast, have very large genomes with extensive COG T populations, yet each have 20 or fewer proteins identified as having GGDEFdomains (see Table 5), and *Prochloroccus* spp. appear to have none. Conversely, *Oceanospirillum* has a relatively small genome, yet has 112 proteins identified as likely GGDEF domain/IPR000160 proteins. GGDEF/EAL domain response regulators have been implicated in root colonization in *Pseudomonas putida* (Matilla et al. 2007); in *E. coli* the GGDEF domain-containing YddV protein upregulates the transcription of a number of cell wall modification enzymes [50], and in point of fact, *D. aromatica*'s VIMSS581804, a GGDEF domain containing homolog of the YddV *E. coli* protein, occurs upstream of a cluster of sixteen cell wall division proteins (VIMSS581805-581820).

Cellular interactions with community – quorum sensing

Quorum sensing uses specific membrane bound receptors to detect autoinducers released into the environment. It is involved in both intra- and inter-species density detection [51, 52]. Cell density has been shown to regulate a number of cellular responses, including bioluminescence, swarming, expression of virulence factors, secretion, and motility (as reviewed in Withers et al. 2001[53]).

D. aromatica encodes six histidine kinase receptor proteins that are similar to the quorum sensing protein QseC of *E. coli* (VIMSS580745, 582451, 582897, 583274, 3337577 (formerly 583538), and 583893), five of which co-occur on the chromosome with homologs of the CheY like QseB regulator, and two of which appear to be the product of a recent duplication event (VIMSS583893 & 3337577). Of the six QseC homologs, phylogenetic analysis indicates VIMSS582451 is most similar to QseC from *E. coli*, where the QseBC complex regulates motility via the FlhCD master flagellar regulators. *D. aromatica* encodes homologs of both FlhC and FlhD (VIMSS582640 and 582641).

N-acyl-homoserine lactone is the autoinducer typical for gram negative bacteria [54]. However, *Ralstonia* Betaproteobacteria have been reported to display a diversity of candidate cell density signaling compounds other than AHL [55], and *D. aromatica* lacks any recognizable AHL synthesis genes. *Ralstonia eutropha* and *R. solanacearum* likewise encode several proteins in the *qse*C gene family (seven each in *R. eutropha* strains JMP and H16, five in *R. solanacearum* strain GMI1000 and six in strain UW551). Identification of the signaling agents in *D. aromatica*, whether there is shared chemistry with the *Ralstonia* species, and the utility of having a diverse array of putative quorum sensing proteins, remains to be determined.

Cellular interactions with the environment – stress

Carbon Storage

Poly-hydroxyalkanoates (PHAs) store carbon energy, are synthesized from the catabolism of lipids, and constitute up to almost 90% of the dry weight of the Betaproteobacteria species *Comamonas testosteroni* [56]. These lipid-like carbon/energy storage polymers are found in granular inclusions. PhaR candidate VIMSS583509 (as identified by phylogenetic clustering) is likely to be the regulatory protein and is found near other proteins associated with PHA granule biosynthesis and utilization in *D. aromatica* (VIMSS583511-513).

Phasins are relatively small proteins (180-200 aas) that have been shown to associate with PHA inclusions [57]. There are six copies of phasin-type proteins, with indications of recent gene duplication for three of the phasin-type proteins (VIMSS581881, 582264, and 3337571 (formerly 583582). There are also three homologs of the active subunit poly-B-hydroxybutyrate polymerase (PhaC orthologs) and two pha reductase candidates present in a direct repeat, which is also found in *Legionella pneumophila*. Interestingly, one PhaC-like protein, VIMSS583511, is 70% identical to NodG of *Azospirillum brasilense*, a nodulation protein [58]. No PhaA-like ketothiolase ortholog is present. The presence of an amplified gene pool for carbon storage granules in *D. aromatica* may confer the ability to survive under low nutrient conditions, and poly-3-hydroxybutyrate accumulation has recently been observed in *A. aromaticum* EbN1 cultures displaying reduced growth [59].

Phosphate

Inorganic polyphosphate storage appears likely, as both polyphosphate kinase (Ppk, VIMSS582444) and exopolyphosphatase (Ppx, VIMSS583870) are present. These genes are similar to those encoded in *Pseudomonas aeruginosa*, in that they are in disparate regions of the chromosome [60]. Polyphosphate has been implicated in

stress response due to low nutrients in the environment [61], and also in DNA uptake [62].

Phosphate transport appears to be encoded in a large cluster of genes (VIMSS581746-581752), and the response to phosphate starvation by the PhoH homolog VIMSS583854.

Cellular interactions with community/environment – secretion

TypeI secretion

Fifteen transport clusters include a 'TolC-like' outer membrane component, and recent gene family expansion is noted within several families of ABC transporters for this genome. Identification of paralogs within the genome reveals ABC transporters as the largest populated family of genes, similar to *E. coli* [63] (and to most microbes in general). TolC was originally identified in *E. coli* as the channel that exports hemolysin [64], and hemolysin-like proteins are present. Two groups of ABC transporters occur as a cluster of five transport genes; these five-component transporters have been implicated in the uptake of external macromolecules [65].

The presence of putative lytic factors, lipases, proteases, antimicrobials, invasins, hemolysins, RTXs and colicins near potential type I transport systems indicate that these might be effector molecules used by *D. aromatica* for interactions with host cells (eg. for cell wall remodeling). Iron acquisition is likely to be supported by a putative FeoAB protein cluster (VIMSS583997, 583998), as well as several siderophore-like receptors and a putative FhuE protein (outer membrane receptor for ferric iron uptake; VIMSS583312). Other effector-type proteins, likely to be involved in cell/host interactions (and which in some species have a role in pathogenicity [66]), are present in this genome. Adhesins, haemagglutinins, and oxidative stress neutralizers are relatively abundant in *D. aromatica*. A number of

transporters occur near the six putative soluble lytic murein transglycosylases, indicating potential cell wall remodeling capabilities, possibly for host colonization. Homologs of these transporters were shown to support invasin-type functions in other microbes [66]. Interaction with a host is further implicated by: VIMSS581582, encoding a potential cell wall-associated hydrolase, VIMSS581622, encoding a predicted ATPase, and VIMSS581623, encoding a putative membrane-bound lytic transglycosylase.

Eleven tandem copies of a 672 nucleotide insert comprise a region of the chromosome that challenged the correct assembly of the genome, and finishing this region was the final step for the sequencing phase of this project (see Methods). Unexpectedly, analysis of this region revealed a potential open reading frame encoding a very large protein that has been variously predicted at 4854, 2519 or 2491 amino acids in size during sequential automated protein prediction analyses (VIMSS3337779/ formerly 582095). This putative protein, even in its smallest configuration, contains a hemolysin-type calcium-binding region, a cadherin-like domain, and several RTX domains, which have been associated with adhesion and virulence. Intragenic tandem repeats have been observed in other genomes and have been shown to be functionally relevant in fungi, where they encode cell wall proteins [67]. Internal repeats of up to 100 residues with multiple copies have also been found in proteins from *Vibrio, Colwellia, Bradyrhizobium,* and *Shewanella* spp. (termed "VCBS" proteins as defined by TIGRfam1965).

Other potential effector proteins include: three hemolysin-like proteins adjacent to type I transporters, eight proteins with a predicted hemolysin-related function, including VIMSS583067, a hemolysin activation/secretion protein, VIMSS580979, hemolysin A, VIMSS583372, phospholipase/hemolysin,

VIMSS581868, a homolog of hemolysin III, predicted by TIGRfam1065 to have cytolytic capability, VIMSS582079, a transport/hemolysin, and VIMSS581408, a general hemolysin. Five predicted proteins have possible LysM/invasin domains, including: VIMSS580547, 581221, 581781, 582766, and 583769. One gene, VIMSS583068, encodes a putative 2079 amino acid filamentous haemagglutinin, as well as a hasA-like domain, making it a candidate for hasA-like function (hasA is a hemophore that captures heme for iron aquisition [68]).

Type II secretion

Besides the constitutive Sec and Tat pathways, *D. aromatica* has several candidates for dedicated export secretons of unknown function.

Type II candidates include a cluster of nine genes, with 3-4 putative orthologs of PulDEFG interspersed with a lytic transglycosylase and a hemolysin (VIMSS582071-582085). This region is syntenic to an eight gene cluster in *Thiobacillus denitrificans*; several phylogenomically related Betaproteobacteria (*R. solanacearum, R. eutropha, Chromobacterium violaceum,* and *A. aromaticum EbN1)* encode a similar gene cluster of 10-11 type II secretion proteins.

Another candidate for type II secretion occurs from VIMSS581889 to VIMSS581897. It includes *pul*DEFG type subunits and an *exe*A ATPase like protein. It is bracketed by signaling components histidine kinase, adenylate cyclase, and a protein bearing similarity to the nitrogen response regulator *gln*G (VIMSS581898), which has been shown to be involved in NH_3 assimilation in other species [69].

In addition, there is a nine-gene cluster that encodes several proteins related to toluene resistance (VIMSS581899 to 581906).

A pilus-like gene cluster (which can also be classified as type IV secretion) occurs in VIMSS584278-584553, which encode a putative hydrolase, lytic

transglycosylase, cation transporter, pilin peptidase, pilin ATPase and PulF-type protein. This assembly resembles other pilin assemblies associated with attachment to a substrate, such as the PilA/chitin regulated pilus that is responsible for chitin/host colonization in *Vibrio cholerae* [70].

Another large pilus-like cluster (VIMSS584160-584173) occurs in close proximity to the *mhp*CEFDBAR oxygenase genes (see eg VIMSS584157, *mhp*R).

Type III secretion

D. aromatica has been shown to be chemotactic under various circumstances. The flagellar proteins (FliAEFGHIJKLMNOPQR, FlaABCDEFGHIJK) are followed by an additional cluster of 15 chemotaxis/signal transduction genes (VIMSS580462-580476), and homologs of FlhC and D regulatory elements required for the expression of flagellar proteins [71], identified by phylogenetic clustering, are also present. Since *D. aromatica* has a flagellum and displays chemotactic behavior, it is likely that the flagellar gene cluster is solely related to locomotion, though type III secretion systems can also encode dedicated protein translocation machineries that deliver bacterial pathogenicity proteins directly to the cytosol of eukaryotic host cells [72].

Type IV secretion

There are two copies of a twenty-one gene cluster that includes ten putative conjugal transfer (Tra) sex-pilus type genes in the *D. aromatica* genome (VIMSS582582-582601 and VIMSS582864-582884), indicating a typeIV secretion structure that is related to non-pathogenic cell-cell interactions [73].

Type VI secretion

A large cluster of transport proteins that is related to the virulence associated genetic locus HIS-1 of *Pseudomonas aeruginosa* and the VAS genes of *V. cholerae* [74, 75] includes homologs of hcp1, IcmF and clpV (as VIMSS583005, 582995 and 583009, respectively, in *D. aromatica*; see Table 6). This IcmF-associated (IAHP) cluster has been associated with mediation of host interactions, via export of effector proteins that lack signal sequences [75], and most bacteria that contain IcmF clusters are pathogenic agents that associate with eukaryotic cell hosts [76]. Further evidence for typeVI secretion is found in the presence of three proteins containing a Vgr secretion motif modeled by TIGRfam3361, which is found only in genomes having type VI secretory apparatus.

Biofilm formation

There is a rather large cluster of exopolysaccharide export (eps) associated genes, including a proposed exosortase (epsH, VIMSS582792). Presence of the eps family proteins (VIMSS582786, VIMSS582790-582801) indicates capsular exopolysaccharide production, associated with either host cell interactions (including root colonization[77]) or biofilm production in soil sediments [78]. *D. aromatica* is one of a small number of species (19 out of 280 genomes assessed by Haft et al [78]) that also encodes the PEP-CTERM export system. The PEP-CTERM signal, present in sixteen proteins in this genome, are proposed to be exported via a potential exportase, represented in this genome by epsH (VIMSS582792). The presence of proteins encoding this putative exportase is seen only in genomes also encoding the eps genes.

Metabolic Cycles

Nitrogen

D. aromatica closely reflects several metabolic pathways of *R. capsulatus*, which is present in the rhizosphere, and its assimilatory nitrate/nitrite reductase cluster is highly similar to the *R. capsulatus* cluster [79]. Encoded nitrate response elements also support a possible plant association for this microbe, as nitrate can act as a terminal electron acceptor in the oxygen-limited rhizosphere. Alternatively, nitrous oxide (NO) reduction can indicate the ability to respond to anti-microbial NO production by a host (used by the host to mitigate infection [80]).

Nitrate is imported into the cytosol by NasDEFT in *Bacillus subtilis* [81]. A homologous set of genes are encoded by the cluster VIMSS580377-580380, and a homolog of *nar*K is immediately downstream at VIMSS580384, and is likely involved in nitrite extrusion. Upstream, a putative *nas*A/*nir*BDC cluster (assimilatory nitrate and nitrite reduction) is encoded near the *nar*XL-like nitrate response element. VIMSS580393 encodes a nitrate reductase that is homologous to the NasA cytosolic nitrate reductase of *Klebsiella pneumoniae* [81]. Community studies have correlated the presence of NasA-encoding bacteria with the ability to use nitrate as the sole source of nitrogen [82]. The large and small subunits of nitrite reductase (*nir*B, VIMSS580390 and *nir*D-like ferredoxin gene, VIMSS580391), are immediately adjacent to a transporter with a putative nitrite transport function (NirC-like protein, VIMSS580389). The NirB orf is also highly homologous to *both* NasB (nitrite reductase) and NasC (NADH reductase which passes electrons to NasA) of *Klebsiella pneumoniae*. HMMs created from alignments seeded by the NasB and NasC genes scored at $3.2e^{-193}$ and $4.0e^{-159}$, respectively, to the VIMSS580390 NirB protein. *D. aromatica* is similar to *Methylococcus capsulatus, Ralstonia solanacearum, Polaromonas,* and *Rhodoferax ferrireducens* for *nas*A, *nir*B and *nir*D gene clusters.

However, the presence of the putative transporter *nir*C (VIMSS580389) shares unique similarity to the *E. coli* and *Salmonella nir*BCD clusters.

Putative periplasmic, dissimilatory nitrate and nitrite reduction, which are candidates for denitrification capability [83], are encoded by the *nir*D1 and *nap*DABCD genes (VIMSS 3337807/581796-581799). A probable cytochrome c' is encoded by VIMSS582015. Although most denitrifiers are free living, plant-associated denitrifiers do exist [84]. There is no dissimilatory nitrate reductive complex *nar*GHIJ, but rather, NarG and NarH-like proteins are found in the evolutionarily-related perchlorate reductase alpha and beta subunits [4]. These proteins are present in the *pcr*ABCD*cld* cluster, VIMSS582649-582652 and VIMSS584327, as previously reported for *Dechloromonas* species [85].

Ammonia incorporation appears to be metabolically feasible via a putative glu-ammonia ligase (VIMSS583081), which incorporates free ammonia into the cell via ligation to a glutamic acid. It is encoded near an ABC transport complex. In addition, a putative carbon-nitrogen ligase (VIMS583083) would form glutamine from the above product, using glutamic acid as an amido-N-donor, and thus as a nitrogen source. An ammonium transporter and cognate regulator are likely encoded in the Amt and GlnK-like proteins VIMSS581101 and 581102.

Urea catabolism as a further source of nitrogen is suggested by two different urea degradation enzyme clusters. The first co-occurs with a urea ABC-transport system, just upstream of a putative nickel-dependent urea amidohydrolase (urease) enzyme cluster (VIMSS583666, 583671-583674, and VIMSS583677-583683). The second pathway is suggested by a cluster of urea carboxylase/allophanate hydrolase enzymes (VIMSS581083-581085) [86].

Nitric oxide (NO) reductase

The chromosomal region around *D. aromatica*'s two *nos*Z homologs is notably different from near-neighbors *A. aromaticum* EbN1 and *Ralstonia solanacearum* which encode a nosRZDFYL cluster. Instead, two identical *nos*Z reductase-like genes (annotated as *nos*Z1 and *nos*Z2, VIMSS583543 and VIMSS583547) are adjacent to two cytochrome c553s, a ferredoxin, metal transport accessory proteins, and are uniquely embedded within a histidine kinase/response regulator cluster. This indicates the NO response might be involved in cell signaling and as a possible general detoxification mechanism for nitric oxide.

Nitrogen Fixation

Nitrogen fixation capability in *D. aromatica* is indicated by a complex of *nif*-like genes (see Table 7), that include putative nitrogenase alpha (NifD, VIMSS583693) and beta (NifK, VIMSS583694) subunits of the molybdenum-iron protein, an ATP-binding iron-sulfur protein (NifH, VIMSS583692), and the regulatory protein NifL (VIMSS583623). The nitrogen-fixation *nif* genes present in the free-living soil microbe *Azotobacter vinelandii* share significant sequence homology and synteny to *D. aromatica* for this protein family. *D. aromatica* further encodes a complex that is likely to transport electrons to the nitrogenase, by using a six subunit *rnf*ABCDGE-like cluster (VIMSS583616-583619, 583621 and 583622) that is phylogenomically related to the *Rhodobacter capsulatus* complex used for nitrogen fixation [87]. There is a second *rnf*-like NADH oxidoreductase complex composed of VIMSS583911-583916, of unknown involvement. *A. aromaticum* EbN1 and *Azoarcus* BH72 each encode two *rnf*-like clusters as well.

Hydrogenases associated with nitrogen fixation

Uptake hydrogenase works with the nitrogen fixation cycle in root nodule symbionts to increase efficiency of nitrogen fixation via oxidation of the co-produced hydrogen (H_2) [88]. *D. aromatica* encodes a cluster of 29 predicted orfs (Hydrogenase-1 cluster, VIMSS581358-581387; Table 8) that includes a hydrogenase cluster syntenic to the *hox*KGZMLOQR(T)V genes found in *Azotobacter vinelandii*, which reversibly oxidize H_2 in that organism [89]. This cluster is followed by a second hydrogenase (Hydrogenase-2 cluster, VIMSS581373-581379). The hydrogenase assembly proteins, *hyp*ABF and CDE are also present (VIMSS581368-581370 and 581380-581381, and VIMSS3337851 (formerly 581382) as well as proteins related to the hydrogen uptake (*hup*) genes of various rhizobial microbes [88]. The second region, with the *hyp* and *hyd*-like clusters, lacks overall synteny to any one genome. It does, however, display regions of genes that share synteny with other rhizobial microbes, with *Rhodoferax ferrireducens* producing the highest percent identity across the cluster, both in terms of synteny and protein identity.

Interestingly, VIMSS581384 encodes a homolog of the HoxA hydrogenase transcriptional regulator, which has been shown to be expressed only during symbiosis in some species [90].

Regulation is indicated by homologs of NtrX (VIMSS581123) and NtrY (VIMSS581124); the NtrXY pathway comprises a two-component signaling system involved in the regulation of nitrogen fixation in *Azorhizobium caulinodans* ORS571 [91].

Embedded in the hydrogenase cluster are two gene families involved in urea metabolism (Table 8). This includes the urea transport proteins (UrtABCDE) and urea hydrolase enzyme family (Ure protein family).

Carbon Fixation, the Calvin-Benson-Bassham cycle

The genes indicative of carbon fixation, using the Calvin cycle, are present in the *D. aromatica* genome. This includes Ribulose 1,5-bisphosphate carboxylase (RuBisCo, VIMSS581681), phosphoribulokinase (cbbP/PrkB, VIMSS581690), and a fructose bisphosphate (VIMSS581693) of the Calvin cycle subtype. The RuBisCo *cbbM* gene is of the fairly rare type II form. This sub-type is shared by a only a few microbial species, and *D. aromatica's* displays a surprisingly high 77% amino acid identity to the *cbbM* gene found in the deep-sea tube worm *Riftia pachyptila* symbiont [92], though all species encoding this form displays fairly high amino acid conservation. Further putative cbb proteins are encoded by VIMSS581680 & 581688, candidates for cbbR (regulator for the cbb operon) & cbbY (found downstream of RubisCo in *R. sphaeroides* [93]), respectively.

There is a potential glycolate salvage pathway indicated by the presence of phosphoglycolate phosphatase (Gph, VIMSS583850). Phosphoglycolate results from the oxidase activity of RuBisCo in the Calvin cycle, when concentrations of carbon dioxide are low relative to oxygen. In *Ralstonia (Alcaligenes) eutropha* and *Rhodobacter sphaeroides*, the PGP gene (*cbbZ*) is located on an operon along with other Calvin cycle enzymes, including RuBisCo. This gene, however, is removed from the other *cbb* genes on the chromosome in *D. aromatica*.

The *cco*NOQP gene cluster codes for a cbb-type cytochrome oxidase that functions as the terminal electron donor to O_2 in the aerobic respiration of *Rhodobacter capsulatus* [94]. Note that this cluster is present in a large number of Betaproteobacteria.

Reverse TCA

2-oxoglutarate:ferredoxin oxidoreductase (also known as 2-oxoglutarate synthase, or *kor*) is a key enzyme in the reductive TCA pathway, fixing carbon through CO_2 incorporation into succinyl-CoA. This pathway was suggested for *D. aromatica* [95]. The *D. aromatica* candidate for this key enzyme in the reverse TCA pathway is not a robust fit to the 2-oxoglutarate synthase model, but appears to be an indole pyruvate oxidoreductase, which has been shown instead to oxidize aryl pyruvates generated by the transamination of aromatic amino acids. This enzyme forms aryl acetyl-CoA derivatives in peptide fermentation [96] or peptide catabolism [97]. HMM analysis produced a score of 3 e^{-17} for VIMSS580731 against an alignment of known 2-oxoglutarate synthases, vs. a score of 8 e^{-230} to a model of indole pyruvate oxidoreductases. This indicates this gene is more likely to be involved in peptide catabolism rather than CO_2 incorporation using the reverse TCA cycle. Comparison of the gene cluster alignment to *A. aromaticum* EbN1 corroborates this conclusion, as the gene cluster in *D. aromatica* is syntenic to the indolepyruvate ferredoxin oxidoreductase isotype of *A. aromaticum* EbN1, which carries multiple paralogs of the cognate 2-oxoglutarate synthase (*kor*) genes (EC1.2.7.3).

Wood-Ljungdahl

The Wood-Ljungdahl pathway also appears to be absent, due to the lack of the enzyme carbon monoxide dehydrogenase (EC1.2.99.2; COG1151). Although a CoxL carbon monoxide dehydrogenase candidate (VIMSS583262) is present, it clusters with putative aldehyde dehydrogenases within the *B. japonicum* genome. It does not cluster with either form I or form II of CO dehydrogenase that have been characterized in other Betaproteobacteria [98].

Sulfur

Sulfate and thiosulfate transport appear to be encoded in the gene cluster composed of an OmpA type protein (VIMSS581631) followed by orthologs of a sulfate/thiosulfate specific binding protein Sbp (VIMSS581632), a CysU or T sulfate/thiosulfate transport system permease T protein (VIMSS581633), a CysW ABC-type sulfate transport system permease component (VIMSS581634), and a CysA ATP-binding component of sulfate permease (VIMSS581635).

In addition, candidates for the transcriptional regulator of sulfur assimilation from sulfate are present and include: CysB, CysH, and CysI (VIMSS582362, 582360 and 582362, respectively).

The cytoplasmic sulfur reductase SorAB [99] is not present in *D. aromatica* nor *A. aromaticum* EbN1, although it is found in several other Betaproteobacteria, including *R. metallidurans, R. eutropha, R. solanacearum, C. violaceum,* and *B. japonicum*.

A probable sulfur oxidation enzyme cluster is present and contains homologs of SoxFRCDYZAXB [100], with a putative SoxCD sulfur dehydrogenase, SoxF sulfide dehydrogenase, and SoxB sulfate thiohydrolase, which is predicted to support thiosulfate oxidation to sulfate (see Table 9). A syntenic *sox* gene cluster is also found in *Anaeromyxobacter dehalogens* (although it lacks *sox*FR) and *Ralstonia eutropha*, but not in *A. aromaticum* EbN1. Thiosulfate oxidation, however, has not been supported under the laboratory conditions tested thus far, and experimental support for this physiological capability awaits further investigation.

Gene Family Expansion

To determine candidates for recent gene duplication events, extensive phylogenomic profile analyses were conducted for all sets of paralogs in the genome. Flower Power recruitment and clustering against the non-redundant Genbank protein set was done, and the resulting alignments were analyzed using the tree-building SCI PHY or Belvu based neighbor-joining utilities. The alignment of two or more *D. aromatica* protein sequences in a clade such that they displayed higher % identity to each other than to orthologs present in other species (as one would expect for vertical inheritance patterns of single genes) was interpreted as an indication of a possible recent duplication event, either in the *D. aromatica* genome itself or in a progenitor species. Results of this analysis are shown in Table 10.

Potential gene family expansion is indicated in several functional groups, including the following: signaling proteins (including cAMP signaling, histidine kinases, and others), as well as Mhp-like aromatic oxidation complexes, Nos proteins, hemolysins and transport proteins.

Most duplications indicate that a single gene, rather than sets of genes, were replicated. An exception is the Tra/Type IV transport cluster (VIMSS582581-582601 and VIMSS582864-582884) noted previously. In the protein sets for the histidine kinase/response regulator, duplication of histidine kinase appears to occur without duplication of the adjacent response regulator. The paralogs created by recent duplication events are typically found well-removed from one another on the chromosome, although some tandem repeats of single genes were noted. However, the highest percent identity was not found between pairs of genes in tandem repeats.

Discussion

One of the more striking findings is the absence of key enzymes for monoaromatic degradation in the absence of oxygen. It has generally been assumed

that two of these, those for benzoyl-CoA reduction and fumarate addition to toluene, occur as central intermediaries for all anaerobic aromatic degraders [31, 33]. These enzymes have been identified in other soil microbes, such as *Aromatoleum aromaticum* EbN1 and *Geobacter metallireducens*, that carry out catabolic cycles that are similar to *D. aromatica's*, and *A. aromaticum* EbN1 is a near-neighbor of *D. aromatica* in phylogenomic tree profiling.

Populated KEGG maps in the IMG and VIMSS databases, based on BLAST analyses, indicate the presence of some of these enzymes in *D. aromatica*, yet more careful analysis shows the candidate enzymes to be members of a general family, rather than true orthologs of the enzyme in question. The most reliable prediction-of-function approaches for genomically sequenced protein orfs are obtained using the more computationally intensive HMM modeling and scoring utilities. This allows the protein in question to be assessed by phylogenetic alignment to protein families or sub-families with experimentally known function, providing much more accurate predictions [101, 102]. Both TIGRfams and COGs families, which also employ HMM models, were highly informative in deducing protein family membership for this genome, particularly for highly conserved metabolic pathways (TIGRfams) and signaling families of proteins (COGs).

However, the majority of catabolic enzymes of interest for *D. aromatica* are not covered by TIGRfams or COGs families. For this reason Flower Power clustering, SCI PHY subfamily clade analysis, and HMM scoring were used to ascertain the presence or absence of proteins of interest. HMM models for benzoyl CoA reductase and benzylsuccinylsynthase, previously denoted as "central" to anaerobic catabolism of aromatics, give clear evidence that these enzymes are not present in this genome. Moreover, the set of recruited proteins for both benzoyl-CoA

reductase and benzyl-succinyl synthase indicate they are not as universally present as has been suggested.

Comparative genomics have previously established that large amounts of DNA present in one species can be absent even from a different strain within the same species [103]. In addition, the underestimation of the diversity of aromatic catabolic pathways (both aerobic and anaerobic) has been noted previously [104], and a high level of enzymatic diversity has been seen for pathways that have the same starting and end products, including anaerobic benzoate oxidation [105]. *D. aromatica* itself appears to be a diverse source of protein function through gene family expansion of signaling proteins, dehydrogenases, quorum sensors, phasins, and *mhp* genes.

The lack of anaerobic pathways seen in other microbes for the degradation of aromatic organic compounds, together with the observation that anaerobic degradation of benzene occurs at relatively sluggish reaction rates, indicates that the pathways incumbent in *D. aromatica* for aromatic degradation, occurring under anaerobic conditions, might serve in a detoxification role.Another intriguing possibility is that oxidation is dependent on intracellularly produced oxygen, which is likely to be a rate-limiting step. *Alicycliphilus denitrificans* strain BC couples benzene degradation under anoxic conditions with chlorate reduction, utilizing the oxygen produced by chlorite dismutase in conjunction with a monooxygenase and subsequent catechol degradation for benzene catabolism [106]. A similar mechanism may account for anaerobic benzene oxidation coupled to perchlorate and chlorate reduction in *D. aromatica*. However, anaerobic benzene degradation coupled with nitrate reduction is also utilized by this organism, and remains enigmatic [5].

Relatively little is known about the overall metabolic capabilities of this organism, and the majority of these predictions have not yet been observed *in vivo*. In

point of fact, both carbon fixation and nitrogen fixation gene families, as well as the *sox* gene cluster, were found during annotation of the predicted protein set. Except for nitrogen fixation, which is generally supported in *Dechloromonas* spp (personal communication, JD Coates), growth conditions supporting these metabolic capabilities have not yet been found.

The large number of signaling proteins present in the genome indicate a potentially complex lifestyle for this species.

The presence of several *qseC/B* gene pairs indicates the possibility of specific responses that are dependant on different sensing strategies. In other species, expression of ABC exporters is regulated by quorum sensing systems [68]; gene family expansion is indicated in the ABC export gene pool as well as the qseC/B sensors in *D. aromatica*.

Several gene families are present that indicate interactions with a eukaryotic host species, including response elements that potentially neutralize host defense molecules, in particular nitric oxide and other nitrogenous species. *Dechloromonas* spp. have been found in anoxic rice root communities, and increase relative to other microbial species in the root community in the presence of added nitrate [107]. *D. aromatica* can respire on nitrate, but the typical denitrification pathway is not present in this organism (for comparison, see Philippot et al. 2001 [108]), as the enzyme complexes responsible for the first step, nitrate reduction, and the last step, nitric oxide reductase, do not fit canonical models. The first step, typically carried out by membrane bound nitrate reductase (NarDKGHJI), is instead represented by a perchlorate reductase that is evolutionarily related to the Nar proteins, but which utilizes perchlorate, rather than nitrate, as the electron acceptor. *D. aromatica*'s *nosRZDFYL* operon lacks the nosRFYL genes, and displays other notable differences

with most nitrate reducing microbes. Rather than one *nosZ* gene (the reductase for nitrous oxide), this gene has been duplicated, resulting in two identical *nosZ* genes in the genome, along with *nosD*, and a *napGH*–like pair that potentially couples quinone oxidation to cytochrome q reduction. These genes are embedded in a cluster of signaling histidine kinase/response regulator genes. The Epsilonproteobacteria *Wolinella succinogenes*, with its "unprecedented" *nos* gene cluster [109], is quite similar to *D. aromatica* for *nos* genes (that is, both have two *nosZ* genes, a *nosD* gene and a *napGH* pair in the same order and orientation), but the *W. succinogenes* genome lacks the embedded signaling protein cluster. Further, the NirD1 and NapDABC (VIMSS 3337807/581796-581799) proteins, along with the cytochrome c' protein (VIMSS582015), which has been shown to bind nitric oxide (NO) prior to its reduction [110], are all present, and potentially act in detoxification roles. In general, it appears that nitrogen species are important in a number of signaling capacities for this organism. It has been shown that formation of anaerobic biofilms of *P. aeruginosa* (which cause chronic lung infections in cystic fibrosis) require NO reductase when quorum has been reached [111], so a role in signaling and complex cell behavior is possible.

Surprisingly, *W. succinogenes*, shares other genome features with *D. aromatica.* It encodes only 2042 orfs, yet has a large number of signaling proteins, histidine kinases, and GGDEF proteins relative to its genome size. It also encodes *nif* genes, several genes similar to virulence factors, and similarity in the nitrous oxide enzyme cluster noted above. *W. succinogenes* is evolutionarily related to two pathogenic species (*Helicobacter pylori* and *Campylobacter jejuni*), and displays eukaryotic host interactions, yet is not known to be pathogenic [112]. *Vibrio fischeri* likewise shares several pathogenic congeners with *Vibrio* spp that are pathogenic, yet

its cohort of identified pathogen-like effector molecules are involved in symbiosis with squid bioluminescent membranes [113], and the distinction between effector molecules causing a pathogenic interaction and a symbiotic one is unclear.

The type IV pili systems might be involved in biofilm development, as interactions with biofilm surfaces are affected by force-generating motility structures, including type IV pili and flagella [114]. In *E. coli*, biofilm formation is dependent on the presence of conjugative plasmids [115, 116], and biofilms have been proposed to be ecological niches in which conjugation is highly favored [114]. Although plasmids were not found in the sequenced genome, *D. aromatica* displays many predicted proteins that indicate the uptake of external DNA is possible and perhaps even likely. Further, quorum sensing is a deciding input for biofilm formation, and the presence of an exopolysaccharide synthetic cluster lends further support for biofilm formation. It is intriguing that derivatives of nitrous oxide, which is an evident substrate for *D. aromatica*, are a key signal for biofilm formation vs cell dispersion in the microbe *P. aeruginosa* [117].

The presence of the *cbb*M gene suggests the ability to carry out the energetically costly fixation of CO_2, though such functionality has yet to be observed, and carbon dioxide fixation capability has been found in only a few members of the microbial community. The cbbM subtypes is also not typical: Within *Rhodobacteria*, in a genotype analysis of purple nonsulfur bacteria in aquatic sediments, *Rhodoferax fermentans, Rhodospirillum fulvum* and *Rhodospirillum rubrum* were found to be the only isolates of purple nonsulfur bacteria that carry the *cbb*M type II isoform of RuBisCo (the subtype found in *D. aromatica* [118]).

In *Leptospira interrogans*, which use mammalian reservoir hosts, the presence of 79 genes that encode the two-component sensor histidine kinase-response regulator

proteins has been proposed to be due to a complex array of environmental signals, both inside and outside the host [119], and in the cyanobacterium *Nostoc punctiforme,* the high number of signal transductants is attributed to a complex lifestyle with broad symbiotic competence [120]. Cyanobacteria in general appear to encode large numbers of signaling proteins, as seen in Fig. 5. In contrast, *Magnetospirillum magneticum* is a free-living microbe that does not have an observed host relationship, yet which has a high number of signaling molecules[121]. In that species, signaling capability is suggested to support high sensitivity to different environments, and to be instrumental for energy taxis. The underlying dynamics are further complicated by the finding that the Betaproteobacterium *Ralstonia solanacearum* is a rhizosphere symbiont, yet has far fewer signal transducing proteins encoded in its genome than *D. aromatica*. Never-the-less, the capability for diverse response to various environmental signals appear most likely to be associated with a symbiotic lifestyle for *D. aromatica,* given the many other proteins implicated in interaction with a host species.

Conclusions

Dechloromonas aromatica strain RCB appears to support a highly complex lifestyle which likely involves both biofilm formation and interaction with a eukaryotic host. It lacks predicted enzyme families for anaerobic aromatic catabolism, though it supports degradation of several aromatic species in the absence of oxygen. The enzymes responsible for this metabolic function remain to be identified and characterized. It also encodes proteins suggestive of the ability to fix nitrogen and CO_2, as well as thiosulfate oxidation. Converse to aromatic degradation, these enzymatic functionalities have yet to be experimentally demonstrated. In short, this genome was full of surprises.

The utility of TIGRfams and COGs families in these analyses cannot be overstated. New releases of TIGRfams during the course of this analysis provided new insights and identified new functionality (malonate degradation cluster, PEP-Cterm transport and the epsH putative translocon, and urea degradation all were identified in the TIGRfam 7.0 additions). The HMM model building and assessment utilized as the major annotation approach for this study was employed to cover those protein families of interest that are not currently covered by TIGRfams. We utilized K. Sjölander's modelling and analysis tools, which are highly similar to those used to produce TIGRfams models. Overall, the extensive use of HMMs during this analysis allowed high confidence in predicted protein function, as well as ascertaining that a given gene is actually not present. This proved to be of enormous importance given the large number of characterized pathways for aromatic degradation that are absent in this genome.

…

Authors' contributions

AL coordinated and oversaw the assembly of the genome. WSF, HF and GDB did the initial assembly of the genome. KKS conducted genome assembly, sequence finishing and gap closure activities, and created the final assembly. ST provided internal Joint Genome Institute assembly and analysis tools, and support in their use. KK was involved in the semi-automated genome annotation, and provided support for the VIMSS dataset and data lists from that set. KKS conducted all manual annotation work described here-in.

Acknowledgements

KKS sincerely thanks Tanja Woyke for her very helpful suggestions and direction for creation of the tables and figures, Patrick Chain for helpful suggestions on the manuscript, Dan Kirshner for technical help on computational work, Nandini Krishnamurthy for building an internally clustered data-set of *D. aromatica* proteins as well as help with computational tools, Ching Shang for ideas regarding

biochemical pathways, and Frank W Larimer for a cogent and extremely helpful critique of the manuscript. Special thanks go to JD Coates and LA Achenbach for providing the initial funding under which this project was started. The genome finishing portion of this study was performed under the auspices of the US Department of Energy's Office of Science, Biological and Environmental Research Program, and by the University of California, Lawrence Livermore National Laboratory under Contract No. W-7405-Eng-48, Lawrence Berkeley National Laboratory under Contract No. DE-AC02-05CH11231 and Los Alamos National Laboratory under Contract No. W-7405-ENG-36. The majority of the annotation was done as an independent project with considerable support from Adam Arkin, Katherine Huang, Morgan Price, Eric Alm, and Kimmen Sjölander – sufficient gratitude cannot be expressed for their generous contributions of help and time.

References

1. Thrash JC, Van Trump JI, Weber KA, Miller E, Achenbach LA, Coates JD: **Electrochemical stimulation of microbial perchlorate reduction**. *Environ Sci Technol* 2007, **41**(5):1740-1746.
2. Coates JD, Chakraborty R, McInerney MJ: **Anaerobic benzene biodegradation--a new era**. *Res Microbiol* 2002, **153**(10):621-628.
3. Coates JD, Chakraborty R, Lack JG, O'Connor SM, Cole KA, Bender KS, Achenbach LA: **Anaerobic benzene oxidation coupled to nitrate reduction in pure culture by two strains of Dechloromonas**. *Nature* 2001, **411**(6841):1039-1043.
4. Bender KS, Shang C, Chakraborty R, Belchik SM, Coates JD, Achenbach LA: **Identification, characterization, and classification of genes encoding perchlorate reductase**. *J Bacteriol* 2005, **187**(15):5090-5096.
5. Chakraborty R, Coates JD: **Hydroxylation and carboxylation--two crucial steps of anaerobic benzene degradation by Dechloromonas strain RCB**. *Appl Environ Microbiol* 2005, **71**(9):5427-5432.
6. Chakraborty R, O'Connor SM, Chan E, Coates JD: **Anaerobic degradation of benzene, toluene, ethylbenzene, and xylene compounds by Dechloromonas strain RCB**. *Appl Environ Microbiol* 2005, **71**(12):8649-8655.
7. Kasai Y, Takahata Y, Manefield M, Watanabe K: **RNA-based stable isotope probing and isolation of anaerobic benzene-degrading bacteria from gasoline-contaminated groundwater**. *Appl Environ Microbiol* 2006, **72**(5):3586-3592.

8. Beller HR, Spormann AM: **Analysis of the novel benzylsuccinate synthase reaction for anaerobic toluene activation based on structural studies of the product**. *J Bacteriol* 1998, **180**(20):5454-5457.
9. Edwards A, Voss H, Rice P, Civitello A, Stegemann J, Schwager C, Zimmermann J, Erfle H, Caskey CT, Ansorge W: **Automated DNA sequencing of the human HPRT locus**. *Genomics* 1990, **6**(4):593-608.
10. Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment**. *Genome Res* 1998, **8**(3):175-185.
11. Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities**. *Genome Res* 1998, **8**(3):186-194.
12. Gordon D, Abajian C, Green P: **Consed: a graphical tool for sequence finishing**. *Genome Res* 1998, **8**(3):195-202.
13. Badger JH, Olsen GJ: **CRITICA: coding region identification tool invoking comparative analysis**. *Mol Biol Evol* 1999, **16**(4):512-524.
14. Delcher AL, Harmon D, Kasif S, White O, Salzberg SL: **Improved microbial gene identification with GLIMMER**. *Nucleic Acids Res* 1999, **27**(23):4636-4641.
15. Alm EJ, Huang KH, Price MN, Koche RP, Keller K, Dubchak IL, Arkin AP: **The MicrobesOnline Web site for comparative genomics**. *Genome Res* 2005, **15**(7):1015-1022.
16. Krishnamurthy N, Brown D, Sjolander K: **FlowerPower: clustering proteins into domain architecture classes for phylogenomic inference of protein function**. *BMC Evol Biol* 2007, **7 Suppl 1**:S12.
17. Brown DP, Krishnamurthy N, Sjolander K: **Automated protein subfamily identification and classification**. *PLoS Comput Biol* 2007, **3**(8):e160.
18. Schomburg I, Chang A, Schomburg D: **BRENDA, enzyme data and metabolic information**. *Nucleic Acids Res* 2002, **30**(1):47-49.
19. Barthelmes J, Ebeling C, Chang A, Schomburg I, Schomburg D: **BRENDA, AMENDA and FRENDA: the enzyme information system in 2007**. *Nucleic Acids Res* 2007, **35**(Database issue):D511-514.
20. Sonnhammer EL, Hollich V: **Scoredist: a simple and robust protein sequence distance estimator**. *BMC Bioinformatics* 2005, **6**:108.
21. Brown D, Krishnamurthy N, Dale JM, Christopher W, Sjolander K: **Subfamily hmms in functional genomics**. *Pac Symp Biocomput* 2005:322-333.
22. Hughey R, Krogh A: **Hidden Markov models for sequence analysis: extension and analysis of the basic method**. *Comput Appl Biosci* 1996, **12**(2):95-107.
23. Karplus K, Karchin R, Barrett C, Tu S, Cline M, Diekhans M, Grate L, Casper J, Hughey R: **What is the value added by human intervention in protein structure prediction?** *Proteins* 2001, **Suppl 5**:86-91.
24. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D: **Hidden Markov models in computational biology. Applications to protein modeling**. *J Mol Biol* 1994, **235**(5):1501-1531.
25. Brown D, Sjolander K: **Functional classification using phylogenomic inference**. *PLoS Comput Biol* 2006, **2**(6):e77.
26. Sjolander K: **Phylogenomic inference of protein molecular function: advances and challenges**. *Bioinformatics* 2004, **20**(2):170-179.

27. Glanville JG, Kirshner D, Krishnamurthy N, Sjolander K: **Berkeley Phylogenomics Group web servers: resources for structural phylogenomic analysis**. *Nucleic Acids Res* 2007, **35**(Web Server issue):W27-32.
28. Krishnamurthy N, Brown DP, Kirshner D, Sjolander K: **PhyloFacts: an online structural phylogenomic encyclopedia for protein functional and structural classification**. *Genome Biol* 2006, **7**(9):R83.
29. Zmasek CM, Eddy SR: **RIO: analyzing proteomes by automated phylogenomics using resampled inference of orthologs**. *BMC Bioinformatics* 2002, **3**:14.
30. Chakraborty R, Coates JD: **Anaerobic degradation of monoaromatic hydrocarbons**. *Appl Microbiol Biotechnol* 2004, **64**(4):437-446.
31. Egland PG, Pelletier DA, Dispensa M, Gibson J, Harwood CS: **A cluster of bacterial genes for anaerobic benzene ring biodegradation**. *Proc Natl Acad Sci U S A* 1997, **94**(12):6484-6489.
32. Heider J, Fuchs G: **Anaerobic metabolism of aromatic compounds**. *Eur J Biochem* 1997, **243**(3):577-596.
33. Boll M, Fuchs G, Heider J: **Anaerobic oxidation of aromatic compounds and hydrocarbons**. *Curr Opin Chem Biol* 2002, **6**(5):604-611.
34. Rabus R, Kube M, Heider J, Beck A, Heitmann K, Widdel F, Reinhardt R: **The genome sequence of an anaerobic aromatic-degrading denitrifying bacterium, strain EbN1**. *Arch Microbiol* 2005, **183**(1):27-36.
35. Kuhner S, Wohlbrand L, Fritz I, Wruck W, Hultschig C, Hufnagel P, Kube M, Reinhardt R, Rabus R: **Substrate-dependent regulation of anaerobic degradation pathways for toluene and ethylbenzene in a denitrifying bacterium, strain EbN1**. *J Bacteriol* 2005, **187**(4):1493-1503.
36. Kube M, Heider J, Amann J, Hufnagel P, Kuhner S, Beck A, Reinhardt R, Rabus R: **Genes involved in the anaerobic degradation of toluene in a denitrifying bacterium, strain EbN1**. *Arch Microbiol* 2004, **181**(3):182-194.
37. Kane SR, Beller HR, Legler TC, Anderson RT: **Biochemical and genetic evidence of benzylsuccinate synthase in toluene-degrading, ferric iron-reducing Geobacter metallireducens**. *Biodegradation* 2002, **13**(2):149-154.
38. Arai H, Ohishi T, Chang MY, Kudo T: **Arrangement and regulation of the genes for meta-pathway enzymes required for degradation of phenol in Comamonas testosteroni TA441**. *Microbiology* 2000, **146 (Pt 7)**:1707-1715.
39. Arai H, Yamamoto T, Ohishi T, Shimizu T, Nakata T, Kudo T: **Genetic organization and characteristics of the 3-(3-hydroxyphenyl)propionic acid degradation pathway of Comamonas testosteroni TA441**. *Microbiology* 1999, **145 (Pt 10)**:2813-2820.
40. Hofer B, Eltis LD, Dowling DN, Timmis KN: **Genetic analysis of a Pseudomonas locus encoding a pathway for biphenyl/polychlorinated biphenyl degradation**. *Gene* 1993, **130**(1):47-55.
41. Ward NC, Croft KD, Puddey IB, Hodgson JM: **Supplementation with grape seed polyphenols results in increased urinary excretion of 3-hydroxyphenylpropionic Acid, an important metabolite of proanthocyanidins in humans**. *J Agric Food Chem* 2004, **52**(17):5545-5549.
42. Ma YaH, D.S.: **The catechol 2,3-dioxygenase gene and toluene monooxygenase genes from Burkholderia sp. AA1, an isolate capable of degrading aliphatic hydrocarbons and toluene**. *J Ind Microbiol Biotechnol* 2000, **25**:127-131.

43. Kitayama A, Kawakami,Y. and Nagamune,T.: **Gene organization and low regiospecificity in aromatic-ring hydroxylati on of a benzene monooxygenase of Pseudomonas aeruginosa JI104**. *J Ferment Bioeng* 1996, **82**:421-425.
44. Tao Y, Fishman A, Bentley WE, Wood TK: **Oxidation of benzene to phenol, catechol, and 1,2,3-trihydroxybenzene by toluene 4-monooxygenase of Pseudomonas mendocina KR1 and toluene 3-monooxygenase of Ralstonia pickettii PKO1**. *Appl Environ Microbiol* 2004, **70**(7):3814-3820.
45. Zhang H, Luo,H. and Kamagata,Y.: **Characterization of the Phenol Hydroxylase from Burkholderia kururiensis KP23 Involved in Trichloroethylene Degradation by Gene Cloning and Disruption**. *Microbes Environ* 2003, **18**:167-173.
46. Canada KA, Iwashita S, Shim H, Wood TK: **Directed evolution of toluene ortho-monooxygenase for enhanced 1-naphthol synthesis and chlorinated ethene degradation**. *J Bacteriol* 2002, **184**(2):344-349.
47. Mizuno T, Kaneko T, Tabata S: **Compilation of all genes encoding bacterial two-component signal transducers in the genome of the cyanobacterium, Synechocystis sp. strain PCC 6803**. *DNA Res* 1996, **3**(6):407-414.
48. Galperin MY: **Structural classification of bacterial response regulators: diversity of output domains and domain combinations**. *J Bacteriol* 2006, **188**(12):4169-4182.
49. Ryjenkov DA, Tarutina M, Moskvin OV, Gomelsky M: **Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: insights into biochemistry of the GGDEF protein domain**. *J Bacteriol* 2005, **187**(5):1792-1798.
50. Mendez-Ortiz MM, Hyodo M, Hayakawa Y, Membrillo-Hernandez J: **Genome-wide transcriptional profile of Escherichia coli in response to high levels of the second messenger 3',5'-cyclic diguanylic acid**. *J Biol Chem* 2006, **281**(12):8090-8099.
51. Federle MJ, Bassler BL: **Interspecies communication in bacteria**. *J Clin Invest* 2003, **112**(9):1291-1299.
52. Fuqua C, Parsek MR, Greenberg EP: **Regulation of gene expression by cell-to-cell communication: acyl-homoserine lactone quorum sensing**. *Annu Rev Genet* 2001, **35**:439-468.
53. Withers H, Swift S, Williams P: **Quorum sensing as an integral component of gene regulatory networks in Gram-negative bacteria**. *Curr Opin Microbiol* 2001, **4**(2):186-193.
54. Huang JJ, Han JI, Zhang LH, Leadbetter JR: **Utilization of acyl-homoserine lactone quorum signals for growth by a soil pseudomonad and Pseudomonas aeruginosa PAO1**. *Appl Environ Microbiol* 2003, **69**(10):5941-5949.
55. Cha C, Gao P, Chen YC, Shaw PD, Farrand SK: **Production of acyl-homoserine lactone quorum-sensing signals by gram-negative plant-associated bacteria**. *Mol Plant Microbe Interact* 1998, **11**(11):1119-1129.
56. Thakor N, Trivedi U, Patel KC: **Biosynthesis of medium chain length poly(3-hydroxyalkanoates) (mcl-PHAs) by Comamonas testosteroni during cultivation on vegetable oils**. *Bioresour Technol* 2005, **96**(17):1843-1850.
57. Potter M, Madkour MH, Mayer F, Steinbuchel A: **Regulation of phasin expression and polyhydroxyalkanoate (PHA) granule formation in Ralstonia eutropha H16**. *Microbiology* 2002, **148**(Pt 8):2413-2426.

58. Delledonne M, Porcari R, Fogher C: **Nucleotide sequence of the nodG gene of Azospirillum brasilense**. *Nucleic Acids Res* 1990, **18**(21):6435.
59. Trautwein K, Kuhner S, Wohlbrand L, Halder T, Kuchta K, Steinbuchel A, Rabus R: **Solvent stress response of the denitrifying bacterium "Aromatoleum aromaticum" strain EbN1**. *Appl Environ Microbiol* 2008, **74**(8):2267-2274.
60. Zago A, Chugani S, Chakrabarty AM: **Cloning and characterization of polyphosphate kinase and exopolyphosphatase genes from Pseudomonas aeruginosa 8830**. *Appl Environ Microbiol* 1999, **65**(5):2065-2071.
61. Rao NN, Kornberg A: **Inorganic polyphosphate supports resistance and survival of stationary-phase Escherichia coli**. *J Bacteriol* 1996, **178**(5):1394-1400.
62. Reusch RN, Sadoff HL: **Putative structure and functions of a poly-beta-hydroxybutyrate/calcium polyphosphate channel in bacterial plasma membranes**. *Proc Natl Acad Sci U S A* 1988, **85**(12):4176-4180.
63. Linton KJ, Higgins CF: **The Escherichia coli ATP-binding cassette (ABC) proteins**. *Mol Microbiol* 1998, **28**(1):5-13.
64. Wandersman C, Delepelaire P: **TolC, an Escherichia coli outer membrane protein required for hemolysin secretion**. *Proc Natl Acad Sci U S A* 1990, **87**(12):4776-4780.
65. Richarme G, Caldas TD: **Chaperone properties of the bacterial periplasmic substrate-binding proteins**. *J Biol Chem* 1997, **272**(25):15607-15612.
66. Binet R, Letoffe S, Ghigo JM, Delepelaire P, Wandersman C: **Protein secretion by Gram-negative bacterial ABC exporters--a review**. *Gene* 1997, **192**(1):7-11.
67. Verstrepen KJ, Jansen A, Lewitter F, Fink GR: **Intragenic tandem repeats generate functional variability**. *Nat Genet* 2005, **37**(9):986-990.
68. Omori K, Idei A: **Gram-negative bacterial ATP-binding cassette protein exporter family and diverse secretory proteins**. *J Biosci Bioeng* 2003, **95**(1):1-12.
69. Pahel G, Tyler B: **A new glnA-linked regulatory gene for glutamine synthetase in Escherichia coli**. *Proc Natl Acad Sci U S A* 1979, **76**(9):4544-4548.
70. Meibom KL, Li XB, Nielsen AT, Wu CY, Roseman S, Schoolnik GK: **The Vibrio cholerae chitin utilization program**. *Proc Natl Acad Sci U S A* 2004, **101**(8):2524-2529.
71. Tominaga A, Lan R, Reeves PR: **Evolutionary changes of the flhDC flagellar master operon in Shigella strains**. *J Bacteriol* 2005, **187**(12):4295-4302.
72. Hueck CJ: **Type III protein secretion systems in bacterial pathogens of animals and plants**. *Microbiol Mol Biol Rev* 1998, **62**(2):379-433.
73. Marsh JW, Taylor RK: **Genetic and transcriptional analyses of the Vibrio cholerae mannose-sensitive hemagglutinin type 4 pilus gene locus**. *J Bacteriol* 1999, **181**(4):1110-1117.
74. Mougous JD, Cuff ME, Raunser S, Shen A, Zhou M, Gifford CA, Goodman AL, Joachimiak G, Ordonez CL, Lory S *et al*: **A virulence locus of Pseudomonas aeruginosa encodes a protein secretion apparatus**. *Science* 2006, **312**(5779):1526-1530.
75. Pukatzki S, Ma AT, Sturtevant D, Krastins B, Sarracino D, Nelson WC, Heidelberg JF, Mekalanos JJ: **Identification of a conserved bacterial protein**

secretion system in Vibrio cholerae using the Dictyostelium host model system. *Proc Natl Acad Sci U S A* 2006, **103**(5):1528-1533.
76. Parsons DA, Heffron F: **sciS, an icmF homolog in Salmonella enterica serovar Typhimurium, limits intracellular replication and decreases virulence**. *Infect Immun* 2005, **73**(7):4338-4345.
77. Santaella C, Schue M, Berge O, Heulin T, Achouak W: **The exopolysaccharide of Rhizobium sp. YAS34 is not necessary for biofilm formation on Arabidopsis thaliana and Brassica napus roots but contributes to root colonization**. *Environ Microbiol* 2008, **10**(8):2150-2163.
78. Haft DH, Paulsen IT, Ward N, Selengut JD: **Exopolysaccharide-associated protein sorting in environmental organisms: the PEP-CTERM/EpsH system. Application of a novel phylogenetic profiling heuristic**. *BMC Biol* 2006, **4**:29.
79. Cabello P, Pino C, Olmo-Mira MF, Castillo F, Roldan MD, Moreno-Vivian C: **Hydroxylamine assimilation by Rhodobacter capsulatus E1F1. requirement of the hcp gene (hybrid cluster protein) located in the nitrate assimilation nas gene region for hydroxylamine reduction**. *J Biol Chem* 2004, **279**(44):45485-45494.
80. Anjum MF, Stevanin TM, Read RC, Moir JW: **Nitric oxide metabolism in Neisseria meningitidis**. *J Bacteriol* 2002, **184**(11):2987-2993.
81. Ogawa K, Akagawa E, Yamane K, Sun ZW, LaCelle M, Zuber P, Nakano MM: **The nasB operon and nasA gene are required for nitrate/nitrite assimilation in Bacillus subtilis**. *J Bacteriol* 1995, **177**(5):1409-1413.
82. Allen AE, Booth MG, Frischer ME, Verity PG, Zehr JP, Zani S: **Diversity and detection of nitrate assimilation genes in marine bacteria**. *Appl Environ Microbiol* 2001, **67**(11):5343-5348.
83. Siddiqui RA, Warnecke-Eberz U, Hengsberger A, Schneider B, Kostka S, Friedrich B: **Structure and function of a periplasmic nitrate reductase in Alcaligenes eutrophus H16**. *J Bacteriol* 1993, **175**(18):5867-5876.
84. Baek SH, Shapleigh JP: **Expression of nitrite and nitric oxide reductases in free-living and plant-associated Agrobacterium tumefaciens C58 cells**. *Appl Environ Microbiol* 2005, **71**(8):4427-4436.
85. Waller AS, Cox EE, Edwards EA: **Perchlorate-reducing microorganisms isolated from contaminated sites**. *Environ Microbiol* 2004, **6**(5):517-527.
86. Kanamori T, Kanou N, Atomi H, Imanaka T: **Enzymatic characterization of a prokaryotic urea carboxylase**. *J Bacteriol* 2004, **186**(9):2532-2539.
87. Schmehl M, Jahn A, Meyer zu Vilsendorf A, Hennecke S, Masepohl B, Schuppler M, Marxer M, Oelze J, Klipp W: **Identification of a new class of nitrogen fixation genes in Rhodobacter capsulatus: a putative membrane complex involved in electron transport to nitrogenase**. *Mol Gen Genet* 1993, **241**(5-6):602-615.
88. Baginsky C, Brito B, Imperial J, Palacios JM, Ruiz-Argueso T: **Diversity and evolution of hydrogenase systems in rhizobia**. *Appl Environ Microbiol* 2002, **68**(10):4915-4924.
89. Menon AL, Mortenson LE, Robson RL: **Nucleotide sequences and genetic analysis of hydrogen oxidation (hox) genes in Azotobacter vinelandii**. *J Bacteriol* 1992, **174**(14):4549-4557.
90. Durmowicz MC, Maier RJ: **Roles of HoxX and HoxA in biosynthesis of hydrogenase in Bradyrhizobium japonicum**. *J Bacteriol* 1997, **179**(11):3676-3682.

91. Pawlowski K, Klosse U, de Bruijn FJ: **Characterization of a novel Azorhizobium caulinodans ORS571 two-component regulatory system, NtrY/NtrX, involved in nitrogen fixation and metabolism**. *Mol Gen Genet* 1991, **231**(1):124-138.
92. Robinson JJ, Stein JL, Cavanaugh CM: **Cloning and sequencing of a form II ribulose-1,5-biphosphate carboxylase/oxygenase from the bacterial symbiont of the hydrothermal vent tubeworm Riftia pachyptila**. *J Bacteriol* 1998, **180**(6):1596-1599.
93. Gibson JL, Tabita FR: **Analysis of the cbbXYZ operon in Rhodobacter sphaeroides**. *J Bacteriol* 1997, **179**(3):663-669.
94. Thony-Meyer L, Beck C, Preisig O, Hennecke H: **The ccoNOQP gene cluster codes for a cb-type cytochrome oxidase that functions in aerobic respiration of Rhodobacter capsulatus**. *Mol Microbiol* 1994, **14**(4):705-716.
95. Raymond J: **The Evolution of Biological Carbon adn Nitrogen Cycling- a Genomic Perspective**. In: *Molecular Geomicrobiology.* Edited by Banfield JF, Cervini-Silva, Javiera and Nealson, Kenneth M., vol. 59. Chantilly, Virginia: The Minerological Society of America; 2005: 211-231.
96. Mai X, Adams MW: **Indolepyruvate ferredoxin oxidoreductase from the hyperthermophilic archaeon Pyrococcus furiosus. A new enzyme involved in peptide fermentation**. *J Biol Chem* 1994, **269**(24):16726-16732.
97. Porat I, Waters BW, Teng Q, Whitman WB: **Two biosynthetic pathways for aromatic amino acids in the archaeon Methanococcus maripaludis**. *J Bacteriol* 2004, **186**(15):4940-4950.
98. King GM: **Molecular and culture-based analyses of aerobic carbon monoxide oxidizer diversity**. *Appl Environ Microbiol* 2003, **69**(12):7257-7265.
99. Kappler U, Friedrich CG, Truper HG, Dahl C: **Evidence for two pathways of thiosulfate oxidation in Starkeya novella (formerly Thiobacillus novellus)**. *Arch Microbiol* 2001, **175**(2):102-111.
100. Friedrich CG, Bardischewsky F, Rother D, Quentmeier A, Fischer J: **Prokaryotic sulfur oxidation**. *Curr Opin Microbiol* 2005, **8**(3):253-259.
101. Storm CE, Sonnhammer EL: **Comprehensive analysis of orthologous protein domains using the HOPS database**. *Genome Res* 2003, **13**(10):2353-2362.
102. Eisen JA: **Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis**. *Genome Res* 1998, **8**(3):163-167.
103. Sibley MH, Raleigh EA: **Cassette-like variation of restriction enzyme genes in Escherichia coli C and relatives**. *Nucleic Acids Res* 2004, **32**(2):522-534.
104. Laurie AD, Lloyd-Jones G: **The phn genes of Burkholderia sp. strain RP007 constitute a divergent gene cluster for polycyclic aromatic hydrocarbon catabolism**. *J Bacteriol* 1999, **181**(2):531-540.
105. Pelletier DA, Harwood CS: **2-Hydroxycyclohexanecarboxyl coenzyme A dehydrogenase, an enzyme characteristic of the anaerobic benzoate degradation pathway used by Rhodopseudomonas palustris**. *J Bacteriol* 2000, **182**(10):2753-2760.
106. Weelink SA, Tan NC, ten Broeke H, van den Kieboom C, van Doesburg W, Langenhoff AA, Gerritse J, Junca H, Stams AJ: **Isolation and characterization of Alicycliphilus denitrificans strain BC, which grows on**

benzene with chlorate as the electron acceptor. *Appl Environ Microbiol* 2008, **74**(21):6672-6681.
107. Scheid D SS, Conrad R: **Identification of rice root associated nitrate, sulfate and ferric iron reducing bacteria during root decomposition**. *FEMS Microbiology Ecology* 2004, **50**:101-110.
108. Philippot L, Mirleau P, Mazurier S, Siblot S, Hartmann A, Lemanceau P, Germon JC: **Characterization and transcriptional analysis of Pseudomonas fluorescens denitrifying clusters containing the nar, nir, nor and nos genes**. *Biochim Biophys Acta* 2001, **1517**(3):436-440.
109. Simon J, Einsle O, Kroneck PM, Zumft WG: **The unprecedented nos gene cluster of Wolinella succinogenes encodes a novel respiratory electron transfer pathway to cytochrome c nitrous oxide reductase**. *FEBS Lett* 2004, **569**(1-3):7-12.
110. Turner SM, Moir JW, Griffiths L, Overton TW, Smith H, Cole JA: **Mutational and biochemical analysis of cytochrome c', a nitric oxide-binding lipoprotein important for adaptation of Neisseria gonorrhoeae to oxygen-limited growth**. *Biochem J* 2005, **388**(Pt 2):545-553.
111. Yoon SS, Hennigan RF, Hilliard GM, Ochsner UA, Parvatiyar K, Kamani MC, Allen HL, DeKievit TR, Gardner PR, Schwab U *et al*: **Pseudomonas aeruginosa anaerobic respiration in biofilms: relationships to cystic fibrosis pathogenesis**. *Dev Cell* 2002, **3**(4):593-603.
112. Baar C, Eppinger M, Raddatz G, Simon J, Lanz C, Klimmek O, Nandakumar R, Gross R, Rosinus A, Keller H *et al*: **Complete genome sequence and analysis of Wolinella succinogenes**. *Proc Natl Acad Sci U S A* 2003, **100**(20):11690-11695.
113. Ruby EG, Urbanowski M, Campbell J, Dunn A, Faini M, Gunsalus R, Lostroh P, Lupp C, McCann J, Millikan D *et al*: **Complete genome sequence of Vibrio fischeri: a symbiotic bacterium with pathogenic congeners**. *Proc Natl Acad Sci U S A* 2005, **102**(8):3004-3009.
114. Watnick P, Kolter R: **Biofilm, city of microbes**. *J Bacteriol* 2000, **182**(10):2675-2679.
115. Ghigo JM: **Natural conjugative plasmids induce bacterial biofilm development**. *Nature* 2001, **412**(6845):442-445.
116. Reisner A, Haagensen JA, Schembri MA, Zechner EL, Molin S: **Development and maturation of Escherichia coli K-12 biofilms**. *Mol Microbiol* 2003, **48**(4):933-946.
117. Barraud N, Hassett DJ, Hwang SH, Rice SA, Kjelleberg S, Webb JS: **Involvement of nitric oxide in biofilm dispersal of Pseudomonas aeruginosa**. *J Bacteriol* 2006, **188**(21):7344-7353.
118. Oda Y, Wanders W, Huisman LA, Meijer WG, Gottschal JC, Forney LJ: **Genotypic and phenotypic diversity within species of purple nonsulfur bacteria isolated from aquatic sediments**. *Appl Environ Microbiol* 2002, **68**(7):3467-3477.
119. Nascimento AL, Ko AI, Martins EA, Monteiro-Vitorello CB, Ho PL, Haake DA, Verjovski-Almeida S, Hartskeerl RA, Marques MV, Oliveira MC *et al*: **Comparative genomics of two Leptospira interrogans serovars reveals novel insights into physiology and pathogenesis**. *J Bacteriol* 2004, **186**(7):2164-2172.

120. Meeks JC, Elhai J, Thiel T, Potts M, Larimer F, Lamerdin J, Predki P, Atlas R: **An overview of the genome of Nostoc punctiforme, a multicellular, symbiotic cyanobacterium**. *Photosynth Res* 2001, **70**(1):85-106.
121. Matsunaga T, Okamura Y, Fukuda Y, Wahyudi AT, Murase Y, Takeyama H: **Complete Genome Sequence of the Facultative Anaerobic Magnetotactic Bacterium Magnetospirillum sp. strain AMB-1**. *DNA Res* 2005, **12**(3):157-166.
122. Kahng HY, Malinverni JC, Majko MM, Kukor JJ: **Genetic and functional analysis of the tbc operons for catabolism of alkyl- and chloroaromatic compounds in Burkholderia sp. strain JS150**. *Appl Environ Microbiol* 2001, **67**(10):4805-4816.
123. Pikus JD, Studts JM, Achim C, Kauffmann KE, Munck E, Steffan RJ, McClay K, Fox BG: **Recombinant toluene-4-monooxygenase: catalytic and Mossbauer studies of the purified diiron and rieske components of a four-protein complex**. *Biochemistry* 1996, **35**(28):9106-9119.
124. Bingle LE, Bailey CM, Pallen MJ: **Type VI secretion: a beginner's guide**. *Curr Opin Microbiol* 2008, **11**(1):3-8.

Figures

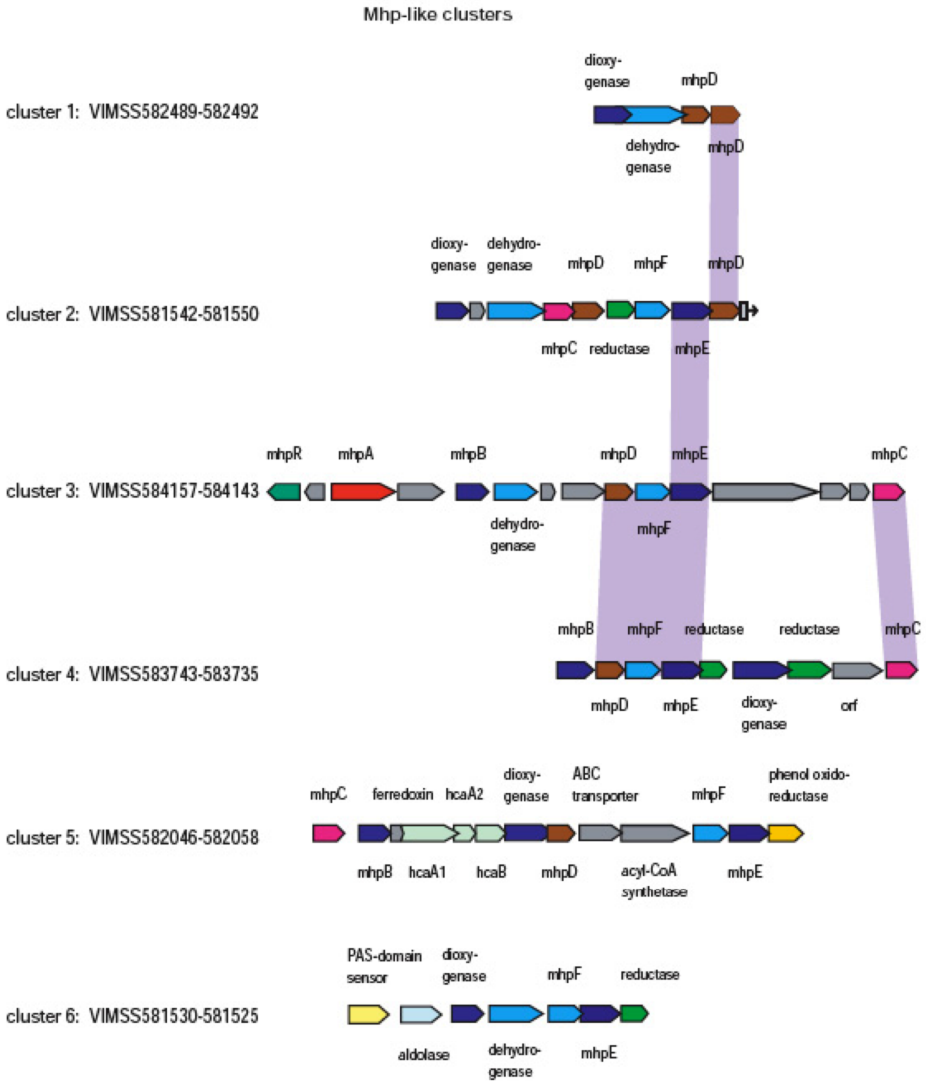

Figure 1
Aerobic degradation of aromatic compounds: multiple Mhp-like dioxygenase clusters. Each of the six mhp-like gene clusters in the D. aromatica genome is depicted. Recent gene duplications between individual proteins are shown by a purple connector between duplicates. Naming convention was chosen for simplicity and consistency, and names all proteins paralogous to a given Mhp protein with the Mhp name (MhpABCDEF or R), but does not imply enzymatic specificity for the substrates listed here-in, though the general enzymatic reaction is highly likely to be conserved. Mhp: meta cleavage of hpp, (hydroxyphenyl)propionate. MhpA, 3HPP hydroxylase; MhpB, DHPP 1,2-dioxygenase; MhpC, 2-hydroxy-6-ketonona-2,4-diene-dioate hydrolase; MhpD, 2, deto-4-pentenoate hydratase; MhpE, 4-hydroxy-2-ketovalerate aldolas; MhpF, acetaldehyde dehydrogenase.

Figure 1. Annotation pipeline using HMM models.

Hidden Markov Models were used in various steps to aid in annotation of proteins predictions from the *D. aromatica* genome. Models were also made for proteins of interest to determine whether an ortholog of that protein exists in the *D. aromatica* genome. A. Determining the presence or absence of enzymes of known function in the *D. aromatica* genome. Enzymes of known function that constitute the anaerobic

aromatic catabolic pathways from *Aromatoleum aromaticum* EbN1 were used to seed HMM models that were then either assessed for recruitment of *D. aromatica* proteins to the model, or scored against the *D. aromatica* protein set. B. Determining the presence or absence of enzymes of known function specifically from the the BRENDA database, using the same protocol as (A). C. Proteins of interest from the *D. aromatica* genome were used as seed sequences in HMM recruitment schema to determine putative function. Functional predictions were based on evolutionarily-related proteins recruited to the same clade in phylogenomic tree profiles. D. *D. aromatica* proteins were internally clustered to determine paralogous proteins within the genome, for further analysis.

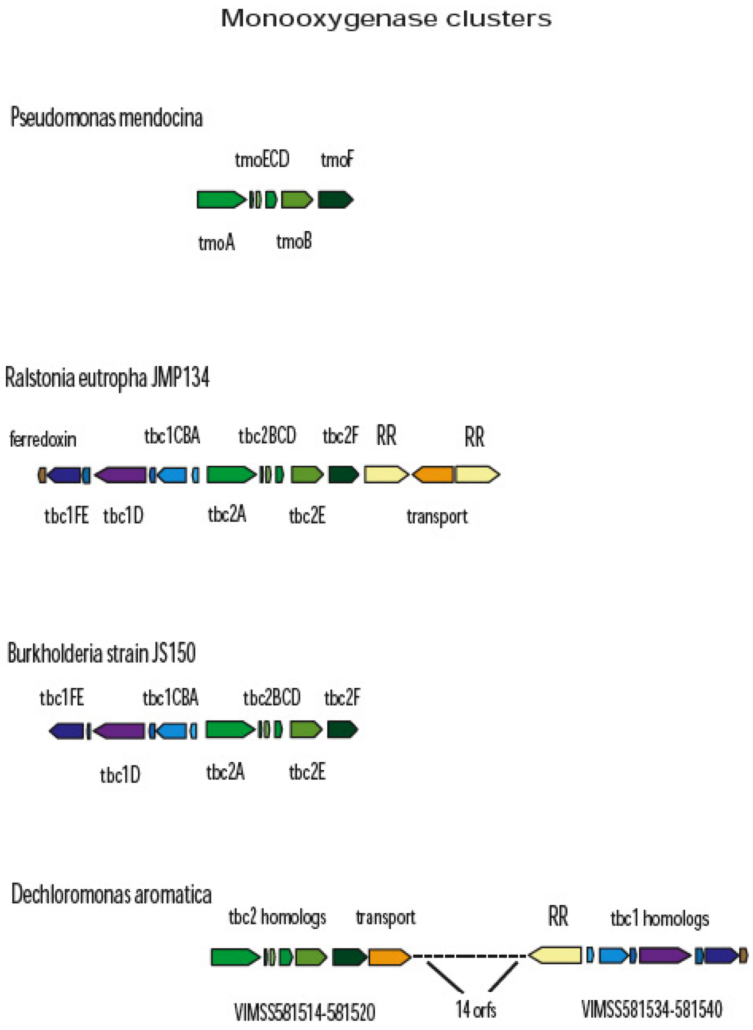

Figure 2
Catabolic oxygenases of aromatic compounds: Synteny between D. aromatica, P. mendocina, Burkholderia and R. eutropha. Orthologous gene clusters for P. mendocina, R. eutropha JMP134, Burkholderia JS150 and D. aromatica are shown. D. aromatica possesses two oxygenase gene clusters that are syntenic to the tbc1 and 2 catabolic gene clusters of Burkholderia JS150, but with an inversion and insertion in the chromosome. Also shown are the tmo (toluene mono-oxygenase) toluene degradative cluster of P. mendocina and the tbc1 & tbc2-like (tcb: toluene, chlorobenzene, and benzene utilization) gene cluster of R. eutropha (VIMSS 896207–896222, Burkholderia protein names were used for consistency). The first seven orfs (encoding a tbc1-like cluster) of R. eutropha JMP134 are orthologous to the PoxABCDEFG (phenol hydroxylase) and P0123456 genes of Ralstonia sp E2 and R. eutropha H16, respectively. Orthologs can be identified as having the same size and color scheme.

Figure 2. Overview of predicted metabolic cycles and signaling proteins in *D. aromatica*

Various metabolic cycles, secretory apparatus and signaling cascades predicted in the annotation process are depicted. TM: transmembrane. Gene names are discussed in the relevant sections of this paper.

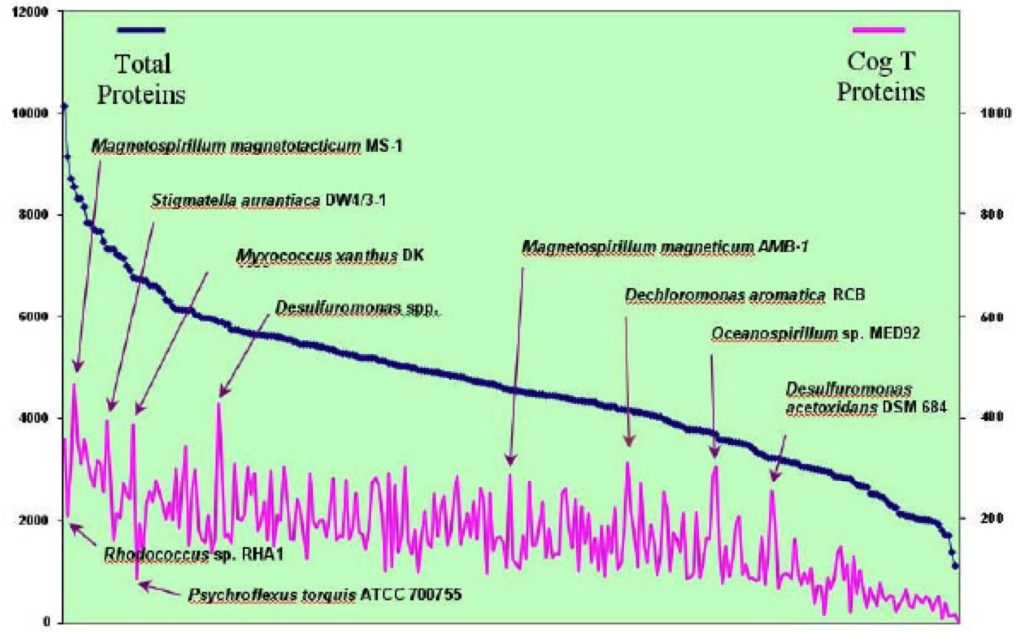

Figure 3
Number of predicted signaling proteins versus total protein count. Microbial genomes, displaying total number of predicted open reading frames (orfs, left axis) and total number of predicted signaling proteins (defined as COG T, right axis). Microbes displaying a high number of signaling orfs relative to total predicted proteins are labelled (above COG T line), as well as two large-sized genomes having a relatively low number of annotated COG T proteins (labelled below COG T line).

Figure 3. Aerobic degradation of aromatic compounds: multiple mhp-like dioxygenase clusters

Each of the six *mhp*-like gene clusters in the *D. aromatica* genome is depicted. Recent gene duplications between individual proteins are shown by a purple connector between duplicates. Naming convention was chosen for simplicity and consistency, and names all proteins paralogous to a given Mhp protein with the Mhp name (MhpABCDEF or R), but does not imply enzymatic specificity for that protein family.

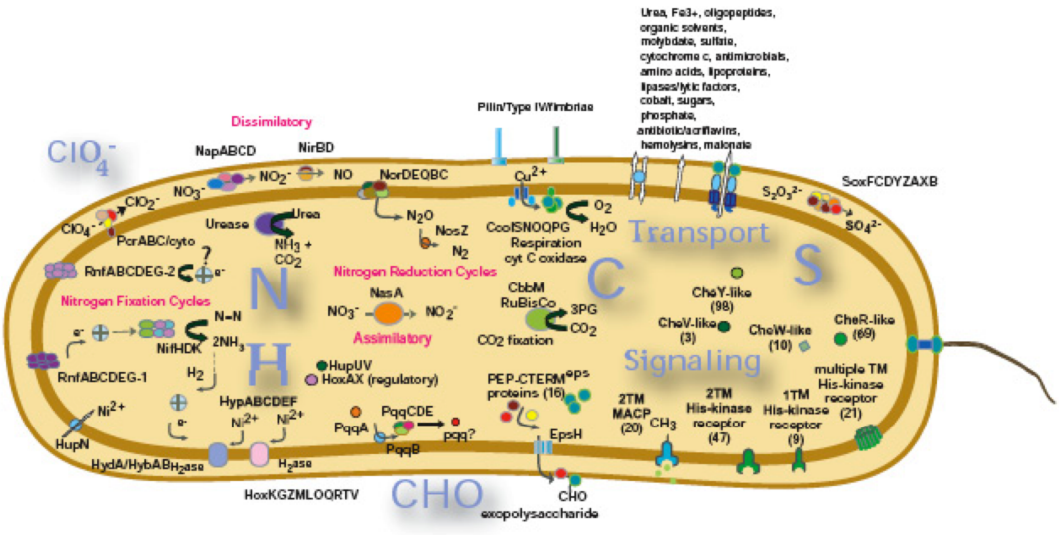

Figure 4
Overview of predicted metabolic cycles, membrane transporters and signaling proteins in D. aromatica. Various metabolic cycles, secretory apparatus and signaling cascades predicted in the annotation process are depicted. TM: transmembrane. Gene names are discussed in the relevant sections of this paper. Areas of the cell depicting Nitrogen, Hydrogen, Carbon and Sulfur cycles are indicated by "N," "H," "C," and "S."

Figure 4. Catabolic oxygenases of aromatic compounds: synteny between *D. aromatica*, *Burkholderia* JMP104 and *R. eutropha*

Orthologous gene clusters for *P. mendocina*, *R. eutropha* JMP104, *Burkholderia* str JS150 and *D. aromatica* are shown. *D. aromatica* possesses two oxygenase gene clusters that are syntenic to the *tbc*1 and 2 catabolic gene clusters of *Burkholderia* str JS150 [122], but with an inversion and insertion in the chromosome. Also shown are the *tmo* toluene degradative cluster of *P. mendocina* [123] and the *tbc*1 & *tbc*2-like gene cluster of *R. eutropha* (from the VIMSS database). The first seven orfs (encoding a *tbc*1-like cluster) of *R. eutropha* JMP104 are orthologous to the PoxABCDEFG and P0123456 genes of *Ralstonia* sp E2 and *R. eutropha* H16, respectively.

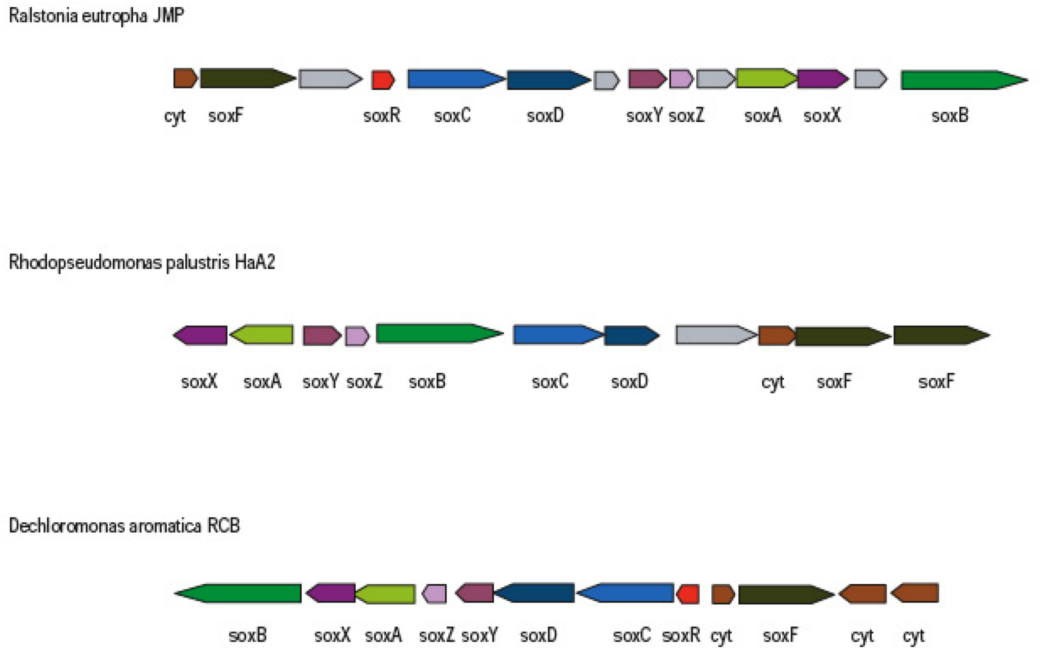

Figure 5
Sulfur oxidation (thiosulfate to sulfate) candidates in R. eutropha, R. palustris, and D. aromatica. Proposed model for this periplasmic complex is as follows: SoxXA, oxidatively links thiosulfate to SoxY; SoxB, potential sulfate thiohydrolase, interacts with SoxYZ (hydrolyzes sulfate from SoxY to regenerate); SoxCD, a sulfur dehydrogenase; oxidizes persulfide on SoxY to cysteine-S-sulfate and potentially yields 6 electrons per sulphate; SoxC, sulfite oxidase/dehydrogenase with homology to nitrate reductase, induced by thiosulfate; SoxDE, both c-type cytochromes with two heme-type binding sites; and SoxF, a FAD flavoprotein with sulfide dehydrogenase activity. Cyt, cytochrome.

Figure 5. Number of predicted signaling proteins versus total protein count. Microbial genomes, displaying total number of predicted open reading frames (orfs, left axis) and total number of predicted signaling proteins (defined as COG T, right axis). Microbes displaying a high number of signaling orfs relative to total predicted proteins are labelled (above COG T line), as well as two large-sized genomes having a relatively low number of annotated COG T proteins (labelled below COG T line).

Figure 6: Sulfur oxidation (thiosulfate to sulfate) candidates in *R. eutropha*, *R. palustris*, and *D. aromatica*.

Proposed model (see Curr Opinion Microbiol 8:253-259) for this periplasmic complex is as follows: SoxXA, oxidatively links thiosulfate to SoxY; SoxB, potential sulfate thiohydrolase, interacts wtih SoxYZ (hydrolyzes sulfate from SoxY to regenerate); SoxCD, oxidizes persulfide on SoxY to cysteine-S-sulfate, a sulfur dehydrogenase, yields 6 electrons per sulfate; SoxC, sulfite oxidase/dehydrogenase with homology to nitrate reductase, induced by thiosulfate; SoxDE, both c-type cytochromes with two

heme-type binding sites; and SoxF, FAD flavoprotein with sulfide dehydrogenase activity. Cyt, cytochrome.

…

Tables

Genome Statistics	Number
Genome length (bp)	4,501,104
Plasmids	none
Depth of Coverage	24
Reads:	
40kb	4399
8kb*	64680
3kb*	61325
PCR	196
Total used in final assembly	130600
Total in all contigs	132499
*includes primer-walking reads for finishing	
5s rRNA	4
16s rRNA	4
23s rRNA	4
Total number of predicted orfs:	
VIMSS	4170
JGI	4204

Table 1 - *Dechloromonas aromatica* RCB genome.
Statistics for finishing the *D. aromatica* genome, and resulting annotation, are shown.

VIMSS	Number	Percent
Total Number of Predicted Proteins (orfs)	4170	100%
TIGRFAMS:		
Categorized by TIGRFAM (total: domain, hypothetical, equivalog)	1368	33%
Categorized by TIGRFAM equivalog	723	18%
Number of predicted orfs without any type of TIGRFAM coverage	2802	67%
Number of predicted orfs with no annotation	539	13%
Pseudogenes	41	1%

Table 2 - Statistics for open reading frame predictions.

Statistics for predicted coding sequences (orfs) in the VIMSS database.

Anaerobic aromatic pathways in *Aromatoleum aromaticum* str. EbN1	*A. aromaticum* EbN1 - representative protein used for HMM model	*Azoarcus* BH72 ortholog	*D. aromatica* RCB ortholog
1) phenylalanine			
Pat	VIMSS813888 : pat (COG1448; EC 2.6.1.57)	-	-
Pdc	VIMSS817385 : pdc (COG3961)	-	-
Pdh	VIMSS816687 : pdh (COG1012)	-	-
IorAB	VIMSS813644 : iorA (COG4321)	-	+
2) phenylacetate			
PadBCD	VIMSS816693 : padB	-	-
PadEFGHI	VIMSS816700 : padI	-	-
PadJ	VIMSS816701 : padJ	-	-
3) benzyl alcohol/benzaldehyde			
Adh	VIMSS815388 : adh (COG1062)	-	-
Ald	VIMSS816847 : ald (COG1012; EC1.2.1.28)	+	-
4) p-cresol			
PchCF	VIMSS813733 : pchC (EC: 1.17.99.1)	-	-
PchA	VIMSS815385 : pchC	-	-
	VIMSS813734 : pchF (Vanillyl-alcohol oxidase (EC 1.1.3.38))	-	-
	VIMSS815387 : pchF	-	-
	VIMSS815384 : pchA (COG1012)		
5) phenol			
PpsABC	VIMSS816923 : ppsA phenylphosphate synthase	-	-
PpcABCD	VIMSS815367 : ppcA	-	-
6) 4-hydroxybenzoate			
PcaK	VIMSS816471 : pcaK (COG2271)	-	-
HbcL	VIMSS816681 : hbcL1 4-hydroxybenzoate CoA ligase		
HcrCBA	VIMSS815644 : hcrB		
	VIMSS815645 : hcrA	-	-
7) toluene			
BssDCABEFGH	VIMSS814633 : bssA	-	-
BbsABCDEFGH(IJ)	VIMSS814644 : bbsH	-	-
	VIMSS814645 : bbsG	-	-
	VIMSS814647 : bbsF	-	-
	VIMSS814649 : bbsD	-	-
	VIMSS814651 : bbsB	-	-
8) ethylbenzene			
EbdABC	VIMSS814907 : ebdA	-	+ (pcrA)
Ped	VIMSS814906 : ebdB	-	+ (pcrB)
	VIMSS814905 : ebdC	-	+ (pcrC)
	VIMSS814904 : ebdD	-	+ (pcrD)
	VIMSS814903 : ped	-	-
9) benzoate			
BenK	VIMSS816652 : benK		
BclA	VIMSS815152 : bclA	+	-
BcrCBAD	VIMSS813961 : bcrB	-	-
Dch Had Oah	VIMSS813959 : bcrA	-	-
10) 3-Hydroxybenzoate			
HbcL	VIMSS813951 : hbcL 3-hydroxybenzoate CoA ligase	-	-
BcrADB'C'			

Table 3. Anaerobic aromatic degradation enzymes in near-neighbor *Aromatoleum aromaticum* EbN1.

Anaerobic aromatic degradation enzymes in near-neighbor *Aromatoleum aromaticum* EbN1 are largely absent from *Azoarcus* BH72 and *D. aromatica* RCB. Protein profiles (HMMs) were used to detect the presence or absence of anaerobic enzymes involved in degradation of aromatic compounds, as described in [34, 36]. The ethylbenzene dehydrogenase molybdenum-containing enzyme complex is described by TIGRfams 3479, 3478 and 3482, which define a type II DMSO reductase family of enzymes that includes *D. aromatica*'s perchlorate reductase PcrABD subunits [4].

VIMSS id	Orthologs	Putative function	Size, aas	Number of paralogs in genom
581514	TbuA1/TmoA/TouA/PhlK/Tbc2A	methane/phenol/toluene hydroxylase	501	1
581515	TbuU/TmoE/TouB/PhlL/Tbc2B	toluene-4-monooxygenase	88	1
581516	TbuB/TmoC/TouC/PhlM/Tbc2C	ferredoxin subunit of ring-hydroxylating dioxygenase	111	1
581517	TbuV/TmoD/TouD/PhlN/Tbc2C	monooxygenase	146	1
581518	TbuA2/TmoB/TouE/PhlO/Tbc2E	hydroxylase	328	2
581519	TbuC/TmoF/TouF/PhlP/Tbc2F	flavodoxin reductase	338	6
581520	TbuX/TodX/XylN	membrane protein; transport	464	1
581521	histidine kinase	signal transduction	963	1
581522	NarL	cheY like protein	208	30
581523	methyl-accepting chemotaxis protein	chemotaxis sensory transducer, membrane bound	532	1
581524	4-oxalocrotonate tautomerase	tautomerase	144	1
581525	oxidoreductase	oxidoreductase/dehydrogenase	254	13
581526	MhpE	aldolase	354	5
581527	MhpF	EC1.2.1.10 Acetaldehyde dehydrogenase (acetylating)	305	5
581528	aldehyde dehydrogenase	NAD+-dependent dehydrogenase (potential EC1.2.1.60	489	5
581529	ring-cleaving extradiol dioxygenase	extradiol ring-cleavage dioxygenase	311	3
581530	aldolase	aldolase	266	1
581531	S box domain	signal transduction	143	1
584293	orf		63	1
581532	orf		80	1
584294	EAL domain containing protein	diguanylate phosphodiesterase; signaling	65	1
581533	transcriptional regulator	LysR-type	300	23
581534	response regulator, tbuT family	activator of aromatic catabolism	558	1
812947	PhcK/DmpK/PhhK/PheA1/Tcb1A/AphK	monooxygenase	89	
581535	PhcL/DmpL/PhhL/PheA2/Tcb1B/AphL	hydroxylase	329	2
581536	PhcM/DmpM/PhhM/PheA3/Tcb1C/AphM	monooxygenase	89	1
581537	PhcN/DmpN/PhhN/PheA4/Tcb1D	aromatic hydroxylase	517	1
581538	PhcO/DmpO/PhhO/PheA5/Tcb1E/AphO	aromatic hydroxylase	118	1
581539	PhcP/Dmp/PhhP/PheA6/Tcb1F/AphQ	hydroxylase reductase	353	6
581540	ferredoxin	2Fe-2S ferredoxin, iron-sulfur binding site	112	1
581541	transcriptional regulator	IPR000524: Bacterial regulatory protein GntR, HTH	235	5
581542	ring-cleaving extradiol dioxygenase	IPR011588: Glyoxalase/extradiol ring-cleavage	308	3
581543	orf		142	3
581544	aldehyde dehydrogenase	EC1.2.1.8 Betaine-aldehyde dehydrogenase	484	5
581545	MhpC	2-hydroxy-6-ketonona-2,4-dienedioic acid hydrolase	274	6
581546	MhpD	2-keto-4-pentenoate hydratase	296	8
581547	oxidoreductase	EC1.1.1.100 3-oxoacyl-[acyl-carrier-protein] reductase	264	13
581548	MhpF	EC1.2.1.10 Acetaldehyde dehydrogenase (acetylating)	304	5

Table 4. Aromatic degradation in *D. aromatica*: Mono- and Di-oxygenases.
The large cluster of aromatic degradation enzymes in the *D. aromatica* genome shown includes two mono-oxygenase clusters in a linear array on the *D. aromatica* chromosome, with 17 predicted genes intergenically inserted, which encode *mhp* 'cluster 6' (Fig. 3) and several predicted signaling proteins. The second monooxygenase cluster is followed by *mhp* 'cluster 2'. (Fig. 3).

Genome	IPR000160, GGDEF domain
Prochlorococcus	0
Rhodobacter sphaeroides 2.4.1	16
Escherichia coli K12	19
Nostoc punctiforme	22
Ralstonia solanacearum	23
Aromatoleum aromaticum (formerly Azoarcus) EbN1	26
Ralstonia eutropha JMP134	28
Bradyrhizobium japonicum	35
Pseudomonas fluorescens Pf-5	40
Chromobacterium violaceum ATCC 12472	43
Magnetospirillum magneticum AMB-1	46
Azoarcus BH72	51
Shewanella oneidensis MR-1	54
Alteromonadales bacterium TW-7	55
Dechloromonas aromatica RCB	57
Hahella chejuensis KCTC 2396	65
Shewanella sediminis HAW-EB3	71
Desulfuromonas spp.	109

Table 5. Number of annotated diguanylate cyclase domains (IPR000160) in various genomes.

Total number of proteins annotated with a GDDEF domain as quantified in the VIMSS database.

TIGRfam	VIMSS id	Ortholog	TIGRfam description	# Amino Acids
TIGR03348	582995	IcmF-like	IcmF homologs associated with type VI secretion systems	1270
TIGR03373	582996	Type VI secretion (minor 4)	Found exclusively in type VI secretion-associated gene clusters	327
-	582997	Outer membrane protein (OMP)	OmpA/MotB-like protein	261
TIGR03363	582998	Type VI secretion (chp* 8)	One of two related families in type VI secretion systems that contain an ImpA-related N-terminal domain	365
TIGR01646 & TIGR03361	582999	VgrG-like	Rhs element Vgr protein	911
TIGR03349	583000	Type VI secretion	Type VI secretion systems, IcmF family	257
TIGR03353	583001	Type VI secretion (chp 4)	Associated with type VI secretion loci	447
TIGR03352	583002	IcmF-associated homologous protein (IAHP)	Associated with type VI secretion loci	189
TIGR03558	583003	Type VI secretion (chp 5)	Evp-like (*Edwardella* virulence protein)	170
TIGR03555	583004	Type VI secretion (chp 2)	Evp-like (*Edwardella* virulence protein)	494
TIGR03344	583005	Hcp Hemolysin-coregulated protein	Hemolysin coregulated protein, an exported, homohexameric ring-forming virulence protein from *Pseudomonas aeruginosa*	178
TIGR03357	583006	Type VI secretion associated, lysozyme-like	Similar to acidic lysozyme activity for some phage-encoded members; representing a different subgroup where all members are associated with bacterial type VI secretion system	159
TIGR03359	583007	Type VI secretion (chp 6)	Associated with type VI secretion in a number of pathogenic bacteria. Mutation is associated with impaired virulence, such as impaired infection of plants.	610
TIGR03347	583008	IAHP-related protein	IAHP-related loci, type VI secretion system	354
TIGR03345	583009	ClpV1	Related to chaperone ClpB, but ATPase function only	899
TIGR01646 & TIGR03361	583011	Type VI secretion Rhs element VgrP	Found in Rhs classes G and E	933

Table 6. Type VI secretion cluster.

The type VI secretion locus is shown, with brief TIGRfam family description. Chp: conserved hypothetical protein. Type VI secretion is a fairly recently described secreton occurring in gram negative bacteria. It was initially associated with pathogenicity, but has since been found to be involved in host interactions, and not restricted to pathogens (see eg Bingle et al 2008 [124]).

VIMSS id	Ortholog	Size, aa
583652	FldA, flavodoxin typical for nitrogen fixation	186
583653	hypothetical protein	86
583654	NafY-1, nitrogenase accessory factor Y	247
583655	NifB, nitrogenase cofactor biosynthesis protein	500
583656	4Fe-4S ferredoxin	92
583657	Nitrogenase-associated protein	159
583658	flavodoxin	423
583659	ferredoxin, nitric oxide synthase	95
583660	2Fe-2S ferredoxin	120
583661	NifQ	190
583662	DraG	326
583663	Histidine kinase	1131
583664	Che-Y like receiver	308
583666	UrtA urea transport	420
583667/ 3337562	Cyanate lyase	147
583668	S-box sensor, similar to oxygen sensor arcB	794
583669	ABC transporter	398
3337561	Protein of unknown function involved in nitrogen fixation	72
583671	UrtB urea transport	525
583672	UrtC urea transport	371
583673	UrtD urea transport	278
583674	UrtE urea transport	230
583677	UreH urease accessory protein	288
583678	Urea amidohydrolase gamma	100
583679	Urea amidohydrolase beta	101
583680	Urea amidohydrolase alpha/ UreC urease accessory protein	569
583681	UreE urease accessory protein	175
583682	UreF urease accessory protein	228
583683	UreG urease accessory protein	201
583685	nitroreductase	558
583686	ferredoxin, subunit of nitrite reductase	122
583691	DraT	328
583692	NifH nitrogenase iron protein EC1.18.6.1	296
583693	NifD nitrogenase molybdenum-iron protein alpha chain EC1.18.6.1	490
583694	NifK nitrogenase molybdenum-iron protein beta chain EC1.18.6.1	522
583695/ 3337559	NifT	80
3337558	ferredoxin	63
583696	NafY-2 nitrogenase accessory factor Y	243
583710	NifW nitrogen fixation protein	137
3337555	NifZ	151
583711/ 3337554	NifM	271

Table 7. Putative nitrogen fixation gene cluster in *D. aromatica.*

The annotated nitrogen fixation homologs (Nif proteins) are embedded with a cluster of urea transport and degradation genes (Ure, Urea amidohydrolase and Urt transport families).

VIMSS id	Orthologs	Putative function	Size, aas
581358	HoxK/HyaA /HupS	Hydrogenase-1 small subunit	363
581359	HoxG/HyaB/HupL	Hydrogenase-1, nickel-dependent, large subunit	598
581360	HoxZ/HyaC/HupC	Ni/Fe-hydrogenase 1 b-type cytochrome subunit	235
581361	HoxM/HyaD/HupD	Hydrogenase expression/formation protein	204
581362	HoxL/HypC/HupF	Hydrogenase assembly chaperone	100
581363	HoxO/HyaE/HupG	Hydrogenase-1 expression	152
581364	HoxQ/HyaF/HupH	Nickel incorporation into hydrogenase-1 proteins	287
581365	HoxR/HupI	Rubredoxin-type Fe(Cys)4 protein	66
581366	HupJ/(similar to HoxT)	Hydrogenase accessory protein	166
581367	HoxV/HupV	Membrane-bound hydrogenase accessory protein	308
581368	HypA	Hydrogenase nickel insertion protein	113
581369	HypB	Hydrogenase accessory factor Ni(2+)-binding GTPase	352
581370	HypF	Hydrogenase maturation protein	763
581371	ABC protein	Periplasmic component, ABC transporter	260
581372	GGDEF domain	Signal transduction, GGDEF	523
581373	Hyb0	Hydrogenase-2 small subunit	394
581374	HybA	Fe-S-cluster-containing hydrogenase component	351
581375	HybB	Cytochrome Ni/Fe component of hydrogenase-2	386
581376	HybC/HynA	Hydrogenase-2 large subunit	570
581377	HybD/HynC	Ni,Fe-hydrogenase maturation factor	159
581378	HupF/HypC	Hydrogenase assembly chaperone	96
581379	HybE/HupJ	Hydrogenase accessory protein	183
581380	HypC	Hydrogenase maturation protein	81
581381	HypD	Hydrogenase maturation protein	374
581382/ 3337851	HypE	Hydrogenase maturation protein	330
581383	HoxX/HypX	Formation of active hydrogenase	558
581384	HoxA	Response regulator with CheY domain (signal transduction)	495
581385	HoxB/HupU	Regulatory [NiFe] Hydrogenase small subunit (sensor)	333
581386	HoxC/HupV	Regulatory [NiFe] Hydrogenase large subunit (sensor)	472
581397	HupT	Histidine kinase with PAS domain sensor	448
581398	HoxN/HupN/NixA	Nickel transporter	269

Table 8. Hydrogenase clusters associated with nitrogen fixation.
Hydrogenase-1 is composed of *hox/hup* genes, hydrogenase-2 of the *hyb* genes, and the nickel insertion/maturation complex, *hyp*, is present in two clusters (*hyp*ABF and *hyp*CDE).

VIMSS id	Ortholog	Putative function	Size, aas
582151	cytochrome c553	EC1.9.3.1 Cytochrome-c oxidase	200
582152	cytochrome c553	EC1.9.3.1 Cytochrome-c oxidase	205
582153	SoxF	FAD-dependent sulfide dehydrogenase	425
582154	cytochrome c553	EC1.9.3.1 Cytochrome-c oxidase	101
582155	transcriptional regulator	DNA binding regulatory protein (ArsR family)	106
582156	SoxC	EC1.8.3.1 Sulfite oxidase	448
582157	SoxD	cytochrome c551/c552	348
582158	SoxY	thiosulfate acceptor	155
582159	SoxZ	thiosulfate acceptor subunit	103
582160	SoxA	cytochrome c heme-binding	271
582161	SoxX	cytochrome c, monohaem	215
582162	SoxB	thiohydrolase, regenerates soxY	573

Table 9. Putative sulfur oxidation (sox) cluster.

Potential sulfur oxidation cluster (*sox* genes) as annotated in the *D. aromatica* genome.

Protein family function	Number of duplicates	Numbers of triplicates
Transport	13	2
Signal transduction	12	
Mhp family	6	3
Cytochromes	3	
Transcription regulator	2	
Phasin		1
Heavy metal dependent phosphohydrolase		1
Phospholipase	1	
NosD	1	
NosZ	1	
NafY	1	
Reductase, DMSO type	1	
Enoyl CoA hydratase	2	

Table 10. Candidates for gene expansion in the *D. aromatica* genome.
Proteins within the genome that show evidence of recent gene duplication are tabulated by general functional group..Duplicates, triplicates and quadruplicates are determined by adjacent clustering of the *D. aromatica* proteins on a phylogenomic tree profile. The percent identity between adjacent proteins is higher than identity to other species' protein candidates, indicating a possible recent gene family expansion event.