Insights into origins and function of the unexplored majority of the reductive dehalogenase gene family as a result of genome assembly and ortholog group classification†

Olivia Molenda,1,a Luz A. Puentes Jácome,1,a Xuan Cao,b Camilla L. Nesbø,ab Shuiquan Tang,a Nadia Morson,1,c Jonas Patron,b Line Lomheim,a David S. Wishart,bd and Elizabeth A. Edwards1,*ac

Organohalide respiring bacteria (OHRB) express reductive dehalogenases for energy conservation and growth. Some of these enzymes catalyze the reductive dehalogenation of chlorinated and brominated pollutants in anaerobic subsurface environments, providing a valuable ecosystem service. Dehalococcoides mccartyi strains have been most extensively studied owing to their ability to dechlorinate all chlorinated ethenes – most notably carcinogenic vinyl chloride – to ethene. The genomes of OHRB, particularly obligate OHRB, often harbour multiple putative reductive dehalogenase genes (rdhA), most of which have yet to be characterized. We recently sequenced and closed the genomes of eight new strains, increasing the number of available D. mccartyi genomes in NCBI from 16 to 24. From all available OHRB genomes, we classified predicted translations of reductive dehalogenase genes using a previously established 90% amino acid pairwise identity cut-off to identify Ortholog Groups (OGs). Interestingly, the majority of D. mccartyi dehalogenase gene sequences, once classified into OGs, exhibited a remarkable degree of synteny (gene order) in all genomes sequenced to date. This organization was not apparent without the classification. A high degree of synteny indicates that differences arose from rdhA gene loss rather than recombination.

Phylogenetic analysis suggests that most rdhA genes have a long evolutionary history in the Dehalococcoidia with origin prior to speciation of Dehalococcoides and Dehalogenimonas. We also looked for evidence of synteny in the genomes of other species of OHRB. Unfortunately, there are too few closed Dehalogenimonas genomes to compare at this time. There is some partial evidence for synteny in the Dehalobacter restrictus genomes, but here too more closed genomes are needed for confirmation. Interestingly, we found that the rdhA genes that encode enzymes that catalyze dehalogenation of industrial pollutants are the only rdhA genes with strong evidence of recent lateral transfer – at least in the genomes examined herein. Given the utility of the RdhA sequence classification to comparative analyses, we are building a public web server (http://RDaseDB.biozone.utoronto.ca) for the community to use, which allows users to add and classify new sequences, and download the entire curated database of reductive dehalogenases.

Environmental significance

Organohalide-respiring bacteria (OHRB) including Dehalococcoides mccartyi, Dehalobacter restrictus and Dehalogenimonas are important biocatalysts of a wide variety of dehalogenation reactions. An OHRB genome can harbour over 30 different copies of reductive dehalogenase-encoding genes and public databases currently contain thousands of representatives, yet fewer than 1% of all sequences have a characterized function. We used an amino acid sequence similarity-based method to classify all publicly available reductive dehalogenase sequences from closed genomes or metagenome-assembled genomes. Comparative phylogenetics combined with classification gave us new insights into origins and function of the unexplored majority of the reductive dehalogenase gene family. Given the utility of the approach, we created a public web server (http://RDaseDB.biozone.utoronto.ca) with a curated database of reductive dehalogenases and a classification tool to easily situate newly identified sequences.

Introduction

Exposure to toxic or carcinogenic halogenated organic compounds, particularly chlorinated aliphatic hydrocarbons (CAHs), via groundwater and soil is of public concern.7 While biological transformation processes proceed to some extent
naturally, these processes can be very effectively accelerated. Natural attenuation, biostimulation, and bioaugmentation using mixed microbial consortia capable of reductive dechlorination are now commonly used to remediate sites contaminated with CAHs. Biologically-mediated reductive dechlorination of chlorinated pesticides was reported as early as 1967. Dechlorination of CAHs such as perchloroethene (PCE) to less chlorinated ethenes under methanogenic conditions was first reported in 1983. The idea to deploy specific bacteria to degrade CAHs became more viable after 1987, upon the discovery of anaerobic mixed cultures that could dechlorinate the last and most toxic daughter product, vinyl chloride, from PCE stepwise dechlorination and the subsequent isolation of Dehalococcoides as the responsible organism. At present, the efficacy of in situ bioaugmentation to transform contaminants such as PCE and trichloroethene (TCE), and daughter products cis-dichloroethene (cDCE) and vinyl chloride (VC) to ethene is well established.

Organohalide respiring bacteria (OHRB) conserve energy for growth from cleavage of carbon–halogen bonds. These bacteria use halogenated organic compounds as terminal electron acceptors in their metabolism. DCB-1, later named *Desulfoomonile tiedjet*, was the first characterized OHRB (~1986) capable of growing with 3-chlorobenzoate as an electron acceptor. OHRB come from many different phyla including Deltaproteobacteria (*Geobacter, Desulfuromonas, Anaeromyxobacter, Desulfolina, Desulfomonile and Desulfovibrio*), Epsilonproteobacteria (*Sulfurospirillum*), Betaproteobacteria (*Shewanella and Comamonas*), Firmicutes (*Dehalobacter and Desulfitotubacterium*) and Chlororlexi (*Dehalococcoides, Dehalogenimonas and Dehalobium*). The success of bioremediation is largely due to the discovery of highly substrate-specific OHRB that link dehalogenation with growth, and thus whose populations increase as dehalogenation proceeds. These OHRB express respiratory reductive dehalogenases that can be categorized as part of a quinone-dependent or quinone-independent electron transport chain. The Dehalococcoidia, including Dehalococcoides and Dehalogenimonas, are quinone-independent, obligate organohalide respiring bacteria for which no other growth supporting electron acceptors are known. Characterized members of the Dehalococcoidia are capable of using more difficult-to-dechlorinate CAHs that only have one or two chlorine substituents. This ability of the Dehalococcoidia is key to complete detoxification. The enrichment culture studied herein (called KB-1) is routinely used for bioaugmentation and contains multiple organohalide-respiring bacteria (OHRB), including Dehalococcoides mccartyi, Dehalobacter restrictus and Geobacter lovleyi, which, in concert, are capable of complete dechlorination of many chlorinated ethenes, ethanes and benzenes. From the lens of geologic time, OHRB are phylogenetically diverse microbes of ancient origin that participate not only in detoxification of man-made contaminants, but also in the global halogen cycle, recycling a multitude of naturally-occurring organohalides.

Many OHRB genomes contain multiple distinct genes predicted to be reductive dehalogenases, termed *rdhA* genes sometimes accounting for a significant proportion of their genomes. For example, *D. mccartyi* have remarkably small genomes (~1.4 Mbp) yet some strains contain up to 37 *rdhA* genes. *D. mccartyi* contain a core syntenic region encoding for “housekeeping” genes such as biosynthesis of amino acids, cell components, transcription/translation, nutrient transport and energy conservation. The differences between strains in this region are no more than single nucleotide polymorphisms (SNPs) with average nucleotide identity (ANI) of 85–99%. This stable core genome is interrupted by two variable regions flanking the origin of replication commonly referred to as High Plasticity Regions (HPRs, individually HPR1 and HPR2). HPRs show signs of both lateral gene transfer and homologous recombination including repeats, duplication events, insertion sequences, genomic islands, and plasmid-related sequences. The majority of reductive dehalogenases in *D. mccartyi* are located on these HPRs. In addition, several characterized dehalogenases are within genomic islands with evidence of site-specific integration, such as the VC reductase genes *verA* and *bveA*, and TCE reductase gene *tceA*. It is generally thought that recombination events have allowed *D. mccartyi* to adapt to naturally-occurring and anthropogenic halogenated compounds.

Reducive dehalogenase genes occur in an operon containing the *rdhA*, *rdhb*, and sometimes, additional genes. The *rdhA* gene encodes the catalytically active enzyme (RdhA or RDase if characterized) and the B gene encodes a membrane anchor protein containing two or sometimes three transmembrane helices. The *D. mccartyi* RdhA and Rdhb subunits are part of a unique protein complex which is currently considered to be a fully functional stand-alone respiratory chain without quinone or cytochrome involvement. Other than RdhAB, the protein complex includes the organohalide respiration molybdoenzyme (OmeA), formally called the complex iron–sulfur cluster molybdoenzyme (CISM), found to have hydrogenase activity, its putative membrane anchor (OmeB), hydrogen uptake hydrogenase (HupX) with a [NiFe] large subunit (HupL) and iron–sulfur containing small subunit (HupS).

RdhAs are identified based on the presence of two motifs: a twin-arginine TAT membrane export sequence (RRXXKX), and an eight-iron ferredoxin cluster binding motif (CXXCXXCXXXCP). Rdases can have one or two iron–sulfur binding motifs. In certain Rdases, such as *Dehalogenimonas lykanthroporepellens* BL-DC-9 DcpA, a cobalamin binding motif can be identified (DXHXXGSXLGG). A corrinoid molecule is a critical co-factor for folding and function of RdhA proteins. However, outside of motif regions, RdhAs can have as little as 30% full-length pairwise identity (PID) even when found in the same genome. While many putative dehalogenases sequences can be found in any particular genome, only a select few have been found expressed in response to different CAHs. As a result, without a universal functioning heterologous expression system, only a few dehalogenases have been characterized out of the thousands of sequences in the National Center for Biotechnology Information (NCBI) database. To help classify newly found putative dehalogenases with unknown functions, Hug et al. developed a classification system in 2013, grouping RdhAs with upwards of 90% amino acid PID into sets.
of orthologous sequences (i.e. Ortholog Groups, or OGs) which most likely share similar function.\textsuperscript{39,37} Since 2013, two crystal structures of RdhA proteins have been solved,\textsuperscript{36,37} confirming previous features and providing more resolution to the critical substrate-binding domain.

The first objective of this study was to identify and sequence the different strains of OHRB in a set of related KB-1 subcultures amended with different chlorinated electron acceptors. We had previously noted that different electron acceptors enriched for different strains of \emph{D. mccartyi}.\textsuperscript{38} From three cultures, 8 new \emph{D. mccartyi} genomes were closed, increasing the total number of publicly available \emph{D. mccartyi} genomes to 24. As well, one new \emph{Dehalobacter restrictus} genome was closed increasing the number publicly available to 4 (as of December 2019). The second objective was to compare these genomes to identify key trends in genome evolution, homologous recombination, and lateral gene transfer. We first specifically investigated the evolution of reductive dehalogenase genes in \emph{D. mccartyi} since most genomes were available for comparison. We then tried to extend the analysis to \emph{Dehalogenimonas}, a close relative of \emph{Dehalococcoides}, and finally to \emph{Dehalobacter} – a distinct genus in the Firmicutes. The Ortholog Group classification method was extremely useful in this analysis, which led us to the development of a public web server to make it easier for the community to compare and classify new sequences (http://RDaseDB.biozone.utoronto.ca). The user can also download part of or the entire manually curated reductive dehalogenase sequence database for further analyses.

**Methods**

**Enrichment cultures analysed**

The KB-1 cultures were originally enriched from aquifer materials at a TCE contaminated site in southern Ontario as previously described.\textsuperscript{8,39} A commercial version of the culture is maintained at SiREM (Guelph, ON) for bioaugmentation. The parent enrichment culture maintained at the University of Toronto, referred to as KB-1/TCE-MeOH, has been maintained since 1998 in a 2 L glass bottle (~1.3 L culture with 0.7 L headspace) in anaerobic mineral medium amended with 100 mg L\textsuperscript{-1} TCE as electron acceptor and methanol (MeOH) as electron donor, added at 5× the electron equivalents (eqx.) required for complete dechlorination, as previously described.\textsuperscript{39,41} TCE is completely dechlorinated to ethene prior to refeeding, approximately every 2–4 weeks. Acetogens in the mixed culture fermented added methanol to hydrogen and acetate required by \emph{D. mccartyi}. In 2001, a sub-culture was created with a 2% transfer into pre-reduced anaerobic mineral medium and maintained on VC and hydrogen as described in\textsuperscript{40} and is referred to as KB-1/VC-H\textsubscript{2}. This 200 mL VC enrichment was maintained in a 250 mL glass bottle sealed using a Teflon mininert cap and was amended with 55 mg L\textsuperscript{-1} VC (supplied as 5 mL of pure VC gas) and 5× eqx. hydrogen gas (supplied as 5 mL 80% H\textsubscript{2} : 20% CO\textsubscript{2}, Praxair). The KB-1/VC-H\textsubscript{2} enrichment bottle was also amended with 0.5 mM sodium acetate as a carbon source every ten feedings. In 2003, a 1,2-dichloroethane (DCA)-fed enrichment culture was created (KB-1/1,2-DCA-MeOH culture) fed with 250 mg L\textsuperscript{-1} 1,2-DCA and 5× eqx. MeOH, in a 2 L glass Pyrex bottle containing 1.3 L culture with 0.7 L headspace. Approximately every 6 months, half of each culture was removed and substituted with fresh pre-reduced anaerobic mineral medium to replenish vitamins and buffer. In all enrichment cultures, chlorinated ethenes/ethanes, methane and ethene concentrations were monitored using gas chromatography with an FID detector (Agilent 7890A GC system, G1888 auto-sampler, helium used as carrier gas, packed inlet, Agilent GSQ-Plot column 0.53 mm × 30 mm), calibrated with external standards.

**Metagenomic sequencing and genome assembly**

DNA for metagenome sequencing was extracted from larger samples (40–615 mL) taken from the three functionally-stable enrichment cultures described above: KB-1/VC-H\textsubscript{2} (40 mL sample), KB-1/TCE-MeOH (500 mL sample), KB-1/dDCE-MeOH (300 mL sample) and KB-1/1,2-DCA-MeOH (615 mL sample). Extractions were conducted between February and May 2013. Cultures were filtered using Sterivex™ filters (Millipore 0.2 μm) and the DNA was extracted using the CTAB method (JGI bacterial genomic DNA isolation using CTAB protocol v.3). DNA was sequenced at the Genome Quebec Innovation Sequencing Centre using Illumina HiSeq 2500 technology. Paired-end sequencing with an insert size of ~400 bp and read length of ~150 bp provided roughly 50 million reads per culture. Additional mate-pair sequencing with insert size of ~8000 bp and read length of ~100 bp was conducted for the KB-1/TCE-MeOH and KB-1/1,2-DCA-MeOH cultures where we had more DNA (>10 μg each). For metagenomic sequencing using short-read Next Generation Sequencing (NGS), we have demonstrated the utility of integrating long-insert mate-pair data to resolve challenges in the metagenomic assembly, especially those related to repeat elements and strain variation.\textsuperscript{42} Raw sequences were trimmed with Trimmomatic\textsuperscript{43} to remove bases of low quality and to remove adapters.

The \emph{D. mccartyi} genomes were assembled in six steps as described in supplemental methods. Quantitative PCR to quantify the abundance of specific \textit{rdhA} genes and 16S rRNA amplicon sequencing for relative abundances of different organisms (methods provided in ESI\textsuperscript{†}) were also performed to guide the metagenomic assembly (illustrated in Fig. S1†). In all cases, multiple genomes could be closed (assembled into a single circular scaffold without gaps) from a single enrichment culture because the different populations of \emph{D. mccartyi} were present at different abundances (as inferred from read depth) at the time of sampling. Two complete genomes, each containing a vinyl chloride reductase gene (\textit{vcrA}), were closed from the KB-1/VC-H\textsubscript{2}. The naming convention developed identifies culture (KB-1) electron acceptor (in this case vinyl chloride or VC) and relative abundance (number 1 for highest abundance, 2 for next and so on). Thus, the \emph{D. mccartyi} strains from culture KB-1/VC-H\textsubscript{2} are named strains KBVC1 and KBVC2. Three \emph{D. mccartyi} genomes containing the vinyl chloride reductase \textit{vcrA} were closed from the KB-1/1,2-DCA enrichment culture referred to as strains KBDCA1, KBDCA2 and KBDCA3.
and one Dehalobacter restrictus genome named strain 12DCA. A D. mccartyi complete genome containing a vcrA gene was assembled from KB-1/TCE-MeOH culture, named strain KBTCE1. Additionally, two genomes containing trichloroethene reductase tceA genes were assembled from the KB-1/TCE-MeOH culture, strains KBTCE2 and KBTCE3. The genomes were annotated using the RAST44 and RASys45 servers; results were manually inspected and corrected where required and automatically annotated after submission to NCBI by PGAP.46 Additional searches for conserved domains were conducted using NCBI conserved domain search (E-value threshold of 0.01). The origin of replication was identified using OriRil in R.47

Alignments and phylogenetic trees
A concatenated gene set was created by aligning a set of 109 homologous genes found in genomes of Dehalococcoides mccartyi (22 strains), Dehalogenimonas lykanthroporepellens, D. alkenigignis, Dehalogenimonas sp. WBC-2, and a Chloroflexi out-group Sphaerobacter thermophilus. Outgroups were used to better visualize branching and to improve the estimate of the number of mutations along each branch. Outgroups were selected to be as closely related to in-group as reasonably possible without being part of the in-group. Homologous genes were identified using reciprocal BLASTp and OrthoMCL followed by manual inspection using the Get Homologues package. Each gene was aligned using MUSCLE v3.8.3.1 with default settings. The alignments were concatenated to create one long alignment (138 334 bp long alignment, 26 sequences, 83% average pairwise identity, gaps included). A maximum likelihood (ML) tree was built using the RAxML83% average pairwise identity, gaps included). A maximum likelihoodbootstrap replicates. The best scoring ML tree was chosen as the single tree generated from all Dehalobacter reductive dehalogenase gene and protein sequences, and a third tree with a combination of previously functionally characterized RDases from Dehalococcoides, Dehalobacter and Dehalogenimonas. Alignments and tree-building were conducted using MUSCLE and RAxML as described above. FigTree 1.4.2 was used to visualize and further edit the tree to generate figures in this study (http://tree.bio.ed.ac.uk/software/figtree/).

Time since divergence analysis and detection of positive selection
The time since divergence was calculated using the number of mutations identified from phylogenetic concatenated gene alignments and OG rdhA gene alignments to date (in substitutions per site). Substitutions per site identified from rdhA OG alignments were compared to see whether rdhA could have originated before or after (1) the divergence of Dehalococcoidia into Dehalococcoides and Dehalogenimonas and (2) the divergence of Dehalococcoides into three clades: Pinellas, Cornell and Victoria.

We compared the ratio of non-synonymous (Ka) to synonymous substitutions rates (Ks) in rdhA OG alignments because the Ka/Ks ratio can identify positive selection. If Ka/Ks < 1, then non-synonymous mutations are either neutral or deleterious, whereas if Ka/Ks > 1, then positive selection occurred.53

Table 1 General features of Dehalococcoides mccartyi and Dehalobacter restrictus genomes closed from KB-1 trichloroethene (TCE), 1,2-dichloroethane (1,2-DCA) and vinyl chloride (VC) enrichment cultures

<table>
<thead>
<tr>
<th>Dehalococcoides</th>
<th>Dehalobacter</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>KBVC1</td>
</tr>
<tr>
<td>Genome size (mbp)</td>
<td>1.39</td>
</tr>
<tr>
<td>G + C content (%)</td>
<td>47.3</td>
</tr>
<tr>
<td>Protein coding genes</td>
<td>1468</td>
</tr>
<tr>
<td>Hypothetical genes (%)</td>
<td>31.1</td>
</tr>
<tr>
<td>tRNA</td>
<td>47</td>
</tr>
<tr>
<td>CRISPR-Cas genes</td>
<td>7</td>
</tr>
<tr>
<td>Nitrogen fixation genes</td>
<td>0</td>
</tr>
<tr>
<td>Sub-group/Clade</td>
<td>Pinellas</td>
</tr>
<tr>
<td>Electron acceptor provided to culture</td>
<td>VC</td>
</tr>
<tr>
<td>rdhA genes</td>
<td>22</td>
</tr>
<tr>
<td>Identifiable rdhA genes</td>
<td>vcrA, pceA</td>
</tr>
<tr>
<td>NCBI accession number</td>
<td>CP019968</td>
</tr>
</tbody>
</table>

* Identifiable rdhA indicates the presence of a rdhA gene whose protein product was characterized in a different study. It is not known whether these rdhA are expressed by these specific strains.
ratios were calculated using the PAML package.\textsuperscript{54,55} Only OGs with five or more \textit{rdhA} were used for the analysis since using fewer than five sequences is not recommended.\textsuperscript{56} Two models were run: one which described neutral or nearly neutral evolution and one which additionally accounted for adaptive evolution. The $\chi^2$ test was used to determine whether the $K_s/K_a$ gene ratios were significantly greater than 1, suggesting positive selection.

**Nucleotide sequence accession numbers**

The KB-1 sequencing bioproject can be accessed in the National Center for Biotechnology Information (NCBI) database using bioproject no. PRJNA376155. \textit{Dehalococcoides mccartyi} KB-1 closed genome nucleotide accession numbers in NCBI are: strain KBDC1 CP019867, strain KBDC2 CP019868, strain KBDC3 CP019946, strain KBVC1 CP019968, strain KBVC2 CP19969, strain KBTCE1 CP01999, strain KBTCE2 CP019865,
and strain KBTCE3 CP019866. KB-1 *Dehalobacter restrictus* 12DCA closed genome: CP046996. All metagenomic raw reads are available in NCBI short-read archive (SRA) accession no. SRR11085567-71. 16S rRNA amplicon sequences have been deposited in NCBI SRA accession no. SRP144609.

**Results & discussion**

**General features of *Dehalococcoides* and *Dehalobacter* genomes from KB-1 enrichment cultures**

Eight complete genomes of *D. mccartyi* strains and one *Dehalobacter restrictus* strain were assembled (Fig. S1† for assembly methodology) and annotated from three different enrichment cultures (Table 1) using Illumina mate-pair and paired-end metagenomic sequencing in combination with 16S rRNA amplicon sequencing and qPCR of functional *rdhA* to guide assembly (Fig. S2†). In all KB-1 enrichments, we found lower abundance, fragmented *Dehalococcoides* and *Dehalobacter* contigs which could not be assembled into complete genomes indicating that each of the three KB-1 sub-cultures must have at least one additional strain which could not be assembled totalling at least eleven strains of *Dehalococcoides mccartyi* and at least two strains of *Dehalobacter restrictus*. The strains whose genomes were assembled and closed have been named based on contaminant/electron acceptor amended and a number to indicate rank abundance with respect to other strains found in that same enrichment culture (Table 1). *D. mccartyi* strains, overall, share at least 98% nucleotide sequence identity of the 16S rRNA gene (Fig. S3†) and fall into three clades known as the Pinellas, Victoria and Cornell groupings. Most KB-1 strains (KBCV1, KBVC2, KBCDA1, KBCDA2, KBCDA3 and KBTCE1) fall into the Pinellas clade (which also contains strains CBDB1, BTF08, DCMB5, 11a5, WBC-2, GT, IBARAKI and BAV1), while only two KB-1 strains (KBTCE2 and KBTCE3) fall into the Cornell clade (containing strain 195, MB4 and CG4) (ESI Fig. S3†). All genomes have a clear GC skew with one origin of replication, and the majority of reductive dehalogenase genes are found flanking the origin of replication, primarily coded on the leading strands.

Strains KBCV1 and KBCDA3 contain complete CRISPR-Cas systems that are similar to those in published strains 11a, CBDB1, DCMB5 and GT. Cornell-clade strains KBTCE2 and KBTCE3 contain nitrogen fixation genes similar to type strain 195. The number of putative reductive dehalogenase genes (*rdhA*) among all strains varies from five to twenty-two (Table 1). Strain KBTCE3 has the smallest genome of all known *D. mccartyi* (1.27 Mb), with the fewest number of *rdhA* (only five in both KBTCE3 and KBTCE2) and has highest GC content (49.3% in KBTCE3) (Table 1). We consider all eight KB-1 *D. mccartyi* genomes to be representative of independent strains because the number of accumulated SNPs and genomic rearrangements has clearly resulted in functional changes and differential growth patterns as inferred from the observed different abundances in the same KB-1 sub-culture.

The KB-1 1,2-DCA enrichment culture contains a population of *Dehalobacter*, as determined by 16S rRNA amplicon analysis (Fig. S2†). We closed a genome from this population whose 16S rRNA gene sequence shares 99.9% nucleotide PID with *Dehalobacter restrictus* PER-K23 and 99.8% average nucleotide identity (ANI). Several rearrangements and insertions have occurred which makes the whole genome alignment of the two strains 76.9% nucleotide PID. This KB-1 *Dehalobacter* genome will be further referred to as *Dehalobacter restrictus* strain 12DCA.

**Reductive dehalogenase genes in *Dehalococcoides mccartyi* and conserved synteny**

Hug *et al.* (2013) developed a classification system for reductive dehalogenases where sequences were assigned to Ortholog Groups (OGs) in an attempt to cluster *rdhA* sequences into groups with similar activity on specific halogenated electron acceptors. Previous studies identified 32 distinct *rdhA* genes in the KB-1/TCE-MeOH culture, with two being identical to genes found previously in *D. mccartyi* CBDB1. Although 104 *rdhA* sequences were found in this study, only three new OGs could be described, and the remaining *rdhA* sequences fell into previously identified groupings (Table S2†). As more *D. mccartyi* genomes are closed, fewer new dehalogenases are discovered that do not belong to an existing OG (Fig. 2); as of December 2019, a total of 84 OGs from *D. mccartyi* have been identified. We further applied this classification to 144 *Dehalogenimonas* and 356 *Dehalobacter* reductive dehalogenases. A similar trend occurred with *Dehalobacter restrictus* *RdhAs*, but not with *Dehalogenimonas*, because too few *Dehalogenimonas* genomes are available, and those that are available, are not from the same species (Fig. 2). With the strict 90% amino acid full-length alignment identity cut off, resulting ortholog groups were found to be genus-specific for *Dehalococcoides* and *Dehalogenimonas*.

Whole genome alignments of *D. mccartyi* generally result in strong core-genome region alignments (~90% nucleotide pairwise identity), with poor alignments of two regions flanking the origin, deemed the high plasticity regions (HPRs) (~30% pairwise nucleotide identity) suggesting that HPRs are unique per strain. Analysing twenty-four closed genomes has revealed that most of the *rdhA* in HPRs are actually synteny, and that poor alignments are a result of gene loss of specific *rdhA* genes (Fig. 1 and Fig. 2).
Fig. 3  Synteny in Dehalococcoides mccartyi genomes. (a) Order of rdhA found in high plasticity region one (HPR1) in 22 D. mccartyi genomes labeled by strain name. rdhA are labeled by Ortholog Group (OG) number. rdhA of the same OG share >90% amino acid pairwise identity. Unpaired rdhA (i.e. singletons) are not included due to space limitations (see Table S2† for all genomes and rdhA). Green (leading strand) and red (lagging strand) accents indicate DNA strand clockwise from the oriC. The majority of rdhA are on the leading strand. * found in all D. mccartyi strains. (b) Order of rdhA found in high plasticity region two (HPR2) in 22 D. mccartyi genomes labeled by strain name. rdhA are labeled as in panel (a) except colors reflect different leading strands. Green (lagging strand) and red (leading strand) accents indicate DNA strand anticlockwise from the oriC. The majority of rdhA are on the leading strand. HPR2 starts at tRNA-Leu/Arg/Val at approximately 1.2 Mbp from the origin. OG 35 is only in strain 195 after tRNA-Ala, 1.3 Mbp from the origin. +OG known to be expressed during starvation.
3). Therefore, most strain HPRs are much more similar than previously thought.\(^\text{21}\) A conserved sequential order of \(\text{rdhA}\) exists in all \(D.\ \text{mccartyi}\) strains (Fig. 3a HPR1, Fig. 3b HPR2, with accession numbers Table S2\(^\dagger\)). For example, at the end of HPR2, the most common order of RdhA OGs is 5\(^{-}\)-40:30:11:10:17-3\(^{0}\). Some strains have all of these OGs (CG5, BTF08, CBDB1, KBVC1, GT, KBVC2, KBTCE1, 11a5) while certain strains only have a select few, such as IBARAK1 5\(^{-}\)-30:11-3\(^{0}\), CG4 5\(^{-}\)-10:17-3\(^{0}\), or GY50, that have 5\(^{-}\)-40:30:10:17-3\(^{0}\) (Fig. 3b). The visualization of this synteny would not have been possible without grouping of \(\text{rdhA}\) sequences, nor without the numerous complete genomes currently publicly available in NCBI.

**Trends in \(D.\ \text{mccartyi}\) rdhA: acquisition, loss and evolution**

There are two evolutionary time points which can be identified from the concatenated homologous gene alignment in the \(D.\ \text{mccartyi}\) phylogenetic tree (Fig. 4). First, an estimate of the point of divergence of \(D.\ \text{mccartyi}\) and \(D.\ \text{dehalogenimonas}\) (\(D.\ \text{dehalococcoides}\) speciation) and secondly the separation of \(D.\ \text{mccartyi}\) into three distinct clades: the Pinellas, Cornell and Victoria (clade speciation). The clade separation in \(D.\ \text{mccartyi}\) is evident not only at the 16S rRNA level (Fig. S3\(^\dagger\)) but also in each homologous \(\text{rdhA}\) gene alignment (Fig. 5). \(D.\ \text{mccartyi}\) strains that are in the same clade typically have the same SNPs. These divergence time points could also be seen in \(\text{rdhA}\) alignments within OG groups. Each OG \(\text{rdhA}\) alignment includes sequences with varying numbers of nucleotide substitutions. These substitutions are sometimes clade-specific, meaning that the OG was present prior to clade speciation, and the diversifications occurred after the clade speciation event. Other OGs include sequences with very few substitutions, which are not clade-specific, suggesting more recent acquisition (Fig. 4 and 5 with specific examples; all OG analyzed in Fig. S4 and S5\(^\dagger\)).

Reductive dehalogenase sequences belonging to an OG with a long history with \(D.\ \text{dehalococcoides}\) (before the clade speciation event, or even before the \(D.\ \text{dehalococcoides}\) speciation event, Tables S3 and S4\(^\dagger\)) occur syntenically and are typically preceded by MarR-type regulators, such as MarR-type regulator Rdh2R (cbdb1456), which was found to supress downstream \(\text{rdhA}\) expression in CBDB1.\(^{21}\) Syntenic OG alignments have small (~0) \(K_{S}/K_{A}\) ratios meaning that syntenic OG \(\text{rdhAs}\) are not currently under selective pressure (Fig. 5). The first and most common trend observed from sequence classification and analysis is a consensus order of \(\text{rdhA}\) genes which are ancestral and not found expressed or synthesized at this time\(^{15}\) (summarized in Table 2). \(D.\ \text{dehalococcoides}\) OGs whose level of divergence predates the \(D.\ \text{dehalococcoides}\) speciation event, such as OG 15 and OG 34 have homologous representatives in
Dehalogenimonas, including Dehalogenimonas sp. strains WBC-2, GP, D. alkenigignens and D. formicexedens. 

For some OGs (Table 2), there is evidence for positive selection on RdhA according to $K_a/K_s$ ratios. For example, we found that OGs 13 and 71 occur in the exactly the same position in different genomes, and that genomes contain either one or the other, but not both. OG 13 (Fig. 3) is found in both Pinellas and Victoria clades and is highly similar to OG 71 which is only in Cornell strains (Fig. 3). OG 13 dehalogenases have small $K_a/K_s$ ratios while RdhAs in OG 71 have high ratios, suggesting positive selection. The conditions experienced by Cornell strains likely led to the use and specialization of this dehalogenase. Therefore, the second trend in dehalogenases in D. mccartyi is the evolution of new dehalogenases from existing ones (Summarized in Table 2).

Some $rdhA$ are thought to have been acquired laterally by D. mccartyi, including tceA (OG 5, Fig. 4), bvCA (OG 28), and vcrA (OG 8) due to their location on genomic islands, and very high nucleotide and amino acid sequence conservation among strains from distant geographical locations and sometimes different clades. Additionally, cbrA (OG 53), mbrA (OG 52) pchA (OG 35) and pteA (OG 16) currently also appear to have very few mutations and occur in the vicinity of mobile elements. McMurdie et al. (2011) established that vcrA nucleotide polymorphisms indicate that it was acquired by Dehalococcoides approximately 1000 years ago, after the Dehalococcoides clade speciation event, meaning vcrA looks identical across all clades. All $rdhA$ predicted to occur on mobile elements also have few mutations within OGs like vcrA, suggesting that they were acquired in a similarly recent timeframe (Table S3†). Interestingly, all of these mobile OGs have members which have been biochemically characterized due to their association with dehalogenation of industrial pollutants and growth on these substrates in a laboratory setting.
Table 2 Summary of common features differentiating Ortholog Groups (OGs) of reductive dehalogenase sequences from D. mccartyi

<table>
<thead>
<tr>
<th>Type</th>
<th>General features</th>
<th>Protein information</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Present in more than one clade</td>
<td>OG 15 known to be expressed by D. mccartyi during starvation&lt;sup&gt;101,102&lt;/sup&gt;</td>
</tr>
<tr>
<td></td>
<td>Found in D. mccartyi before clades separated. Synteny within genomes conserved. OG 15 illustrated in Fig. 5</td>
<td>OG 23 found in all D. mccartyi genomes at time of analysis</td>
</tr>
<tr>
<td>2</td>
<td>Pairs of OGs which have recently diverged from one another. Location of rdhA is syntenic across all genomes. OG 13 and 71 illustrated in Fig. 5</td>
<td>No characterized members at time of study</td>
</tr>
<tr>
<td>3</td>
<td>Highly similar PIDs in different strains, regardless of clades. Often associated with mobile genes or on a genomic island. OG 5 illustrated in Fig. 5</td>
<td>5 – TceA&lt;sup&gt;103&lt;/sup&gt; 8 – VcrA&lt;sup&gt;100&lt;/sup&gt; 16 – PteA&lt;sup&gt;60&lt;/sup&gt; 28 – BvCA&lt;sup&gt;104&lt;/sup&gt; 52 – MbrA&lt;sup&gt;63&lt;/sup&gt; 53 – Chra&lt;sup&gt;105&lt;/sup&gt; 35 – Pca&lt;sup&gt;64&lt;/sup&gt;</td>
</tr>
<tr>
<td>4</td>
<td>Duplicated within the same genome. OG 34 illustrated in Fig. 5</td>
<td>No characterized members at time of study</td>
</tr>
</tbody>
</table>

<sup>a</sup> PID – pairwise identity.

Some OGs display movement within a genome, rather than between genomes. Victoria clade strains VS, CG1 and GY50 contain examples of an OG occurring twice in the same genome suggesting that a duplication event occurred. Duplicated OGs occur in different HPRs, and not in tandem (Fig. 3). In one case, pceA (OG 30) appears to have moved within a genome and without duplication. OG 30 is syntenic in HPR2 in 14 strains, with the exception of strain 195 where it occurs in HPR1 (Fig. 3). In strain 195, pceA is located near a serine recombinase which possibly mediated its movement. Additionally, the strain 195 pceA is present without its usual upstream transcriptional regulator, which was thought to be responsible for rdhA regulation.<sup>72</sup> Transcriptomic studies show that strain 195 will continue to produce high transcript levels of pceA regardless of starvation or TCE amendment.<sup>74</sup> Furthermore, strain 195 is the only strain to produce pceA in the presence of PCE, where other strains, such as CG5, transcribe pceA in the presence of multiple PCB congeners<sup>84</sup> and CBDB1 was found to transcribe pceA in the presence of 2,3-dichlorophenol (DCP).<sup>75,76</sup> Thus, the third trend we observed in dehalogenase evolution is mobility, either within or between strains (summarized as types 3 (lateral transfer) and 4 (within genome duplication or rearrangement) in Table 2).

**Gene loss and genomic streamlining in Dehalococcoides mccartyi**

*Dehalococcoides mccartyi* genomes are among the smallest genomes found in free-living bacteria (avg. 1.4 Mbp and 1451 protein-coding genes). A common theme among all small free-living prokaryotes is their high niche specialization and low-nutrient level environments.<sup>77</sup> *D. mccartyi* genomes probably required an extensive period of time to become as specialized and as small as they currently are. Wolf et al. (2013) proposed a theory that gene loss is equally or even more important than lateral gene transfer in shaping genomes.<sup>78</sup> The high level of synteny and number of mutations found in Ortholog Groups of dehalogenases supports this theory in *D. mccartyi*. Given that all dehalogenases capable of dehalogenating industrial contaminants also appear to be mobile, it is possible that anthropogenic releases of organohalides have caused *D. mccartyi* genomes to enter a period of complexification involving sharing select *rdhA*. The exchange of key reductive dehalogenases among *D. mccartyi* can be akin to the recent dissemination of antibiotic resistance genes in the natural environment.<sup>79</sup> Transfer mechanisms may even be related considering the homology observed between the integration module of the vcrA containing genomic island and the Staphylococcal cassette chromosome that carries the broad spectrum beta-lactam resistance gene.<sup>85</sup>

**Reductive dehalogenases in other obligate dechlorinating organisms: Dehalobacter and Dehalogenimonas**

We analyzed the reductive dehalogenases from other OHRB including *Dehalobacter* and *Dehalogenimonas* to extend the OG classification system beyond *Dehalococcoides*. In both cases, there are very few publicly available closed genomes, which limited our analysis. Members of the genus *Dehalobacter* are also organohalide-respiring bacteria that use hydrogen or formate as electron donor, acetate as carbon source, and organohalides as electron acceptors.<sup>83,84</sup> *Dehalobacter* spp. have been shown to respire and/or be implicated in the reductive dechlorination of chlorinated methanes,<sup>85</sup> ethenes,<sup>86–88</sup> ethanes,<sup>89,90</sup> cyclohexanes,<sup>90,91</sup> benzenes,<sup>92</sup> phenols,<sup>94</sup> biphenyls,<sup>95</sup> and phthalides.<sup>95</sup> As of November of 2019, ten *Dehalobacter* genome assemblies, corresponding to strains PER-K23 (DSM 9455), CF, DCA, UNSWDHB, TeCB1, MCB1, 12DCB1, 14DCB1, E1, and FTH1 were publicly available NCBI, but only
three of these genomes are closed (strains PER-K23, CF, DCA) (Table S6† for all genomes analyzed in this study). Four additional metagenome-assembled-genomes of Dehalobacter from KB-1 subcultures are found in the Integrated Microbial Genomes & Microbiomes (IMG/M) system, corresponding to strains TCB1 (2775506732), TCB2 (2786546105), TCB3 (2821530269), and HCH1 (2823892816). Like Dehalococcoides, many rdhA genes (~10 to 30) can be found in each Dehalobacter genome, but only a subset of genes has been characterized (see Table S5†). As part of this study, a new closed Dehalobacter restrictus genome was submitted to NCBI, strain 12DCA (CP046996).

The rdhA genes of the 15 publicly available Dehalobacter restrictus genome assemblies were translated and classified into OGs. Like Dehalococcoides mccartyi, fewer new Dehalobacter restrictus dehalogenases were classified into new OGs with an increasing number of distinct genome assemblies (Fig. 2). A total of 80 OGs resulted from the classification of 356 RdhA protein sequences from 15 genomes (or metagenome assembled genomes/binned genomes), approaching saturation, as with Dehalococcoides mccartyi sequences (Fig. 2). A full Dehalobacter RDase tree annotated by OG can be found in Fig. S6†. An initial examination of the distribution of OGs in closed Dehalobacter genomes as well as in some of the contigs from other Dehalobacter spp., (Fig. S7†) suggests that some Dehalobacter RdhA OGs also display synteny across genomes. As additional Dehalobacter complete and closed genomes become available, trends such as synteny and gene evolution will become easier to see.

With respect to Dehalogenimonas, to-date only six genomes have been closed or binned (D. lykanthroporepellans, Dehalogenimonas sp. strain WBC-2, D. alkenigignens strain IP3-3, D. alkenigignens strain BRE15M, D. formicexedens and Dehalogenimonas sp. strain GP) contributing 144 reductive dehalogenase sequences. These dehalogenases formed only 18 OGs, where 16 of the 18 come from two similar strains of Dehalogenimonas alkenigignens. Unlike Dehalococcoides mccartyi and Dehalobacter restrictus, Dehalogenimonas is a more divergent group at this time (Fig. 4 and 2) with multiple species identified. Much more sequencing is required to more clearly define the extent and organization of OGs in Dehalogenimonas.

Fig. 6  Phylogenetic amino acid tree of reductive dehalogenases which have been at least partially characterized from Dehalococcoides (Dhc, in blue), Dehalogenimonas (Dhg, in grey) and Dehalobacter (Dhb, in yellow). Best scoring maximum likelihood tree displayed. Scale indicates number of substitutions per site. Ortholog Groups (OGs) of dehalogenases with upwards of 90% amino acid identity are highlighted and identified by number. OGs containing a functionally characterized representative are annotated by dehalogenase name. rdhA sequences are coloured by genome they originated from. rdhA are named by NCBI locus tag.
We next considered only the reductive dehalogenases from *Dehalococcoides*, *Dehalobacter* and *Dehalogenimonas* that have been characterized, either biochemically or genetically, to reveal relationships between function and OG classification (Fig. 6). From *Dehalococcoides* dehalogenases, all characterized representatives in the same OG have activity on the same substrates. For example VcRA from different *D. mccartyi* (strain VS, WBC-2 or from the KB-1 consortium) are active on the same tested substrates (see Table S5† for full list of characterized dehalogenases). We found the same to be true with BvcA (*D. mccartyi* BAV1 and KB-1) and TceA (*D. mccartyi* 195 and KB-1). It is important to note that the number of substrates tested was a limited list in all cases. From *Dehalogenimonas* there are no OGs with more than one characterized representative. From *Dehalobacter*, OG 97 can be described as a 1,1,1-trichloroethane/chloroform/1,1-dichloroethane group of dehalogenases with several overlapping functions, and OG 96 includes a perchloroethene dehalogenase and a tetrachlorobenzene/trichlorobenzene dehalogenase. At present, the members of these OGs have not been extensively tested on different substrates. Future work and new protein structures will refine the definition of OGs to better associate dehalogenase function to sequence.

**New community resource: the reductive dehalogenase database web server**

A database of reductive dehalogenase gene sequences and protein sequences was previously located in a publicly accessible Google Doc.† We have transformed this resource into a web-based database where the data can be interactively viewed and downloaded. Additionally, users can check and add new sequences, classify RdhA into Ortholog Groups and visualize sequences in a tree with other RdhA sequences. The webservice can be accessed at: http://RDaseDB.biozone.utoronto.ca.

**Conclusions**

Sequencing and completely closing multiple genomes of *D. mccartyi* has provided new insights into the evolution of rdhA genes in this genus. The majority of *rdhA* found in *D. mccartyi* display a much higher degree of synteny between genomes than previously appreciated. Evolutionary analysis revealed that most *rdhA* sequences are old, and have likely been in *D. mccartyi* genomes since the time of a Dehalococcoidia common ancestor (Fig. 4 and Table S4-†). It is possible that the relatively recent anthropogenic release of specific industrial chemicals has enabled the dissemination of specific reductive dehalogenases capable of solvent dehalogenation, initiating a period of adaptation and complexification of the *D. mccartyi* genome. These genome comparisons have provided specific examples of how the complement of *D. mccartyi* *rdhA* has been shaped, revealing (1) adaptation of existing *rdhA* to new substrates, (2) lateral gene transfer of *rdhA* from the environment and (3) duplication or movement within genomes. While too few genomes are currently available from other OHRB, it is anticipated that these trends will still prevail. The classification of reductive dehalogenase sequences into Ortholog Groups has enabled a better understanding of the evolution and function of *rdhA* sequences, clearly showing that they are of pre-industrial origin. The OG classification prevents over-reporting of “novel” sequences, but rather allows for rapid situation of newly sequenced members in the greater context. To assist future studies of organohalide-respiring bacteria and their reductive dehalogenases, we have begun a public resource to assist the community in the classification of reductive dehalogenase sequences by building a high-quality curated database of all reductive dehalogenases, and maintaining a simple classification system and nomenclature.

**References**


E. R. Hendrickson, J. A. Payne, R. M. Young, M. G. Starr, M. P. Perry, S. Fahnestock, D. E. Ellis and R. C. Ebersole,


63 D. Cheng and J. He, Isolation and characterization of “Dehalococcoides” sp. strain MB which dechlorinates tetrachloroethene to *trans*-1,2-dichloroethene, *Appl. Environ. Microbiol.*, 2009, **75**, 5910–5918.


