

Sequence- and activity-based screening of microbial genomes for novel dehalogenases

Wing Yiu Chan,^{1,2} Max Wong,³ Jennifer Guthrie,⁴
Alexei V. Savchenko,⁴ Alexander F. Yakunin,⁴
Emil F. Pai^{1,2,5**} and Elizabeth A. Edwards^{3*}

¹Department of Biochemistry, University of Toronto,
1 King's College Circle, Toronto, Ontario M5S 1A8,
Canada.

²Division of Cancer Genomics and Proteomics, Ontario
Cancer Institute, University Health Network, 101 College
Street, MaRS/TMDT, Toronto, Ontario M5G 1L7,
Canada.

³Department of Chemical Engineering and Applied
Chemistry, University of Toronto, 200 College Street,
Toronto, Ontario M5S 3E5, Canada.

⁴Structural Proteomics in Toronto, Banting and Best
Department of Medical Research, University of Toronto,
C.H. Best Institute, 112 College Street, Toronto, Ontario
M5G 1L6, Canada.

⁵Departments of Medical Biophysics and Molecular
Genetics, University of Toronto, 1 King's College Circle,
Toronto, Ontario M5S 1A8, Canada.

Summary

Dehalogenases are environmentally important enzymes that detoxify organohalogenes by cleaving their carbon-halogen bonds. Many microbial genomes harbour enzyme families containing dehalogenases, but a sequence-based identification of genuine dehalogenases with high confidence is challenging because of the low sequence conservation among these enzymes. Furthermore, these protein families harbour a rich diversity of other enzymes including esterases and phosphatases. Reliable sequence determinants are necessary to harness genome sequencing-efforts for accelerating the discovery of novel dehalogenases with improved or modified activities. In an attempt to extract dehalogenase sequence fingerprints, 103 uncharacterized potential dehalogenase candidates belonging to the α/β hydrolase (ABH) and haloacid dehalogenase-like hydrolase (HAD) superfamilies were screened for

dehalogenase, esterase and phosphatase activity. In this first biochemical screen, 1 haloalkane dehalogenase, 1 fluoroacetate dehalogenase and 5 L-2-haloacid dehalogenases were found (success rate 7%), as well as 19 esterases and 31 phosphatases. Using this functional data, we refined the sequence-based dehalogenase selection criteria and applied them to a second functional screen, which identified novel dehalogenase activity in 13 out of only 24 proteins (54%), increasing the success rate eightfold. Four new L-2-haloacid dehalogenases from the HAD superfamily were found to hydrolyse fluoroacetate, an activity never previously ascribed to enzymes in this superfamily.

Introduction

Dehalogenases catalyse the cleavage of carbon-halogen bonds and are critical to the detoxification and biodegradation of organohalogenes. Of all types of dehalogenases discovered over the past few decades (Fetzner and Lingens, 1994; Hill *et al.*, 1999; de Jong and Dijkstra, 2003), the hydrolytic dehalogenases are particularly attractive research targets (and have thus been studied the most) because they are cytosolic proteins, generally do not require cofactors, and use water as the sole co-substrate. A large number of hydrolytic dehalogenases belong to two large enzyme superfamilies: the α/β hydrolases (ABH; 22 133 sequences in databases, July 2009) and the haloacid dehalogenase-like hydrolases (HAD; 25 048 sequences). The ABH superfamily includes a variety of hydrolytic enzymes with carboxylesterase, thioesterase, peptidase, haloalkane dehalogenase and fluoroacetate dehalogenase activities (Holmquist, 2000). The vast majority of HAD members have unknown function; those with known function display one of five activities: ATPase, phosphatase, phosphoglucomutase, phosphonohydrolase or L-2-haloacid dehalogenase (Koonin and Tatusov, 1994; Allen and Dunaway-Mariano, 2004). Collectively, the dehalogenases from the ABH and HAD superfamilies transform a relatively broad variety of substrates, which is an advantageous property for bioremediation as well as industrial biocatalysis (Table 1) (Swanson, 1999). Interestingly, despite the absence of any sequence similarities between the ABH and HAD dehalogenases, they all employ a similar two-step

Received 24 August, 2009; revised 21 September, 2009; accepted 24 September, 2009. For correspondence. *E-mail elizabeth.edwards@utoronto.ca; Tel. (+1) 416 9463506; Fax (+1) 416 9788605; **E-mail pai@hera.med.utoronto.ca; Tel. (+1) 416 5817545; Fax: (+1) 416 5817545.

Table 1. Substrate specificities of the ABH and HAD dehalogenases.

Classification	Substrate range	Reference
ABH superfamily		
Haloalkane dehalogenase		
Type I (3R)	Short, primary, aliphatic alkyl chlorides and bromides	a
Type II (3B)	Long (> 3 carbon atoms), primary and secondary, aliphatic (including cyclic) alkyl chlorides and bromides	b
Fluoroacetate dehalogenase	Haloacetates, highest activity on fluoroacetate	c
HAD superfamily		
L-2-Haloacid dehalogenase	S-2-chloro- and bromo-aliphatic carboxylic acids	d

References: a: Keuning *et al.*, 1985; Pries *et al.*, 1994; b: Kmunicek *et al.*, 2005; c: Liu *et al.*, 1998; Kurihara *et al.*, 2003; d: Hasan *et al.*, 1991; van der Ploeg *et al.*, 1991; Liu *et al.*, 1994.

dehalogenation mechanism that involves an aspartate nucleophile and a covalent ester intermediate (Fig. 1) (Verschuieren *et al.*, 1993; Liu *et al.*, 1998; Schmidberger *et al.*, 2007).

Proteome analysis of sequenced microbial genomes revealed that many organisms harbour multiple (occasionally more than 20) genes that share amino acid sequence similarity to ABH and HAD dehalogenases. Unfortunately, their high sequence diversity (e.g. less than 30% sequence identity for HADs) and the lack of robust sequence motif information prevent confident separation of dehalogenases from the much more commonly occurring family members such as esterases, peptidases and phosphatases (Heikinheimo *et al.*, 1999; Nardini and Dijkstra, 1999; Holmquist, 2000; Allen and Dunaway-Mariano, 2004). To our knowledge, only the following hydrolytic dehalogenases have been sequenced and biochemically characterized to date: 9 ABH haloalkane dehalogenases (Keuning *et al.*, 1985; Newman *et al.*, 1999; Jesenska *et al.*, 2002; Jesenska *et al.*, 2005; Kmunicek *et al.*, 2005; Sato *et al.*, 2005; Prudnikova *et al.*, 2009); 2 ABH fluoroacetate dehalogenases (Kawasaki *et al.*, 1992; Kurihara *et al.*, 2003); at least 13 HAD L-2-haloacid dehalogenases (Schneider *et al.*, 1991; van der Ploeg *et al.*, 1991; Jones *et al.*, 1992; Kawasaki *et al.*, 1992; 1994; Murdiyatmo *et al.*, 1992; Liu *et al.*, 1994; Cairns *et al.*, 1996; Hill *et al.*,

1999; Rye *et al.*, 2009); 2 D- and 9 DL-2-haloacid dehalogenases (Barth *et al.*, 1992; Liu *et al.*, 1994; Brokamp *et al.*, 1996; Cairns *et al.*, 1996; Nardi-Dei *et al.*, 1997; Stringfellow *et al.*, 1997; Hill *et al.*, 1999); 1 *cis*- and 1 *trans*-3-chloroacrylate dehalogenase (de Jong *et al.*, 2004; 2007); and 3 4-chlorobenzoyl-CoA dehalogenases (Schmitz *et al.*, 1992; Benning *et al.*, 1996; Chae *et al.*, 1999).

To facilitate the discovery of novel dehalogenases with potential for broader applicability, we selected for biochemical characterization 163 potential ABH and HAD dehalogenase candidates from the sequenced genomes of five common soil bacteria using a sequence-similarity criterion. Improved dehalogenase prediction criteria were developed and validated using a second screen on potential dehalogenase candidates from additional microbial genomes. Finally, we explore the use of *E*-values from database searches as an additional tool for assisting in the identification of novel microbial dehalogenases.

Results and discussion

Selection of dehalogenase genes from sequenced genomes

The potential dehalogenase candidates in the initial screen were selected from sequenced genomes of five common soil microorganisms with direct environmental relevance (*Pseudomonas aeruginosa* PAO1, *Pseudomonas putida* KT2440, *Rhodospseudomonas palustris* CGA009, *Streptomyces avermitilis* MA-4680 and *Sinorhizobium meliloti* 1021). A total of 110 genes from the ABH superfamily were selected using the described procedures (Table S1A). For the HAD superfamily, a total of 53 target genes were selected (Table S2A). In the second screen, we selected 30 additional putative dehalogenases from 14 different microbial genomes (see *Experimental procedures*) using refined search criteria identified from the initial screen (Tables S1B and S2B).

Overexpression in *E. coli* and protein purification

Of the 163 gene sequences (110 ABH and 53 HAD) selected in the initial screen, 155 (95%) were cloned successfully, 111 (68%) expressed well and 85 clones

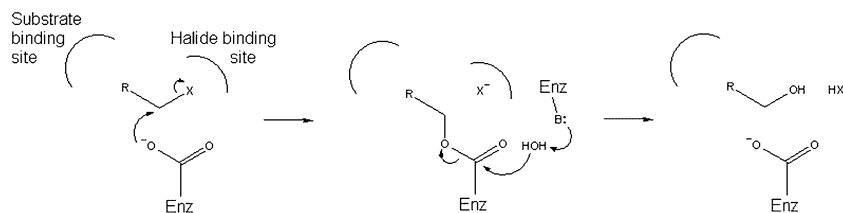
**Fig. 1.** Generalized two-step dehalogenation mechanism employed by several hydrolytic dehalogenases.

Table 2. Substrate profile of all identified dehalogenases.

Enzyme		Substrates			
		1,2-Dichloroethane	1,3-Dichloropropane	Chlorocyclohexane	1,2-Dibromoethane
Haloalkane dehalogenase					
SAV4779	(1)	–	+	+	+
JANN2620	(2)	+	+	+	+
		Fluoroacetate	Chloroacetate	S-2-Chloropropionate	Bromoacetate
Fluoroacetate dehalogenase					
RPA1163	(1)	+	+	–	–
Alr0039	(2)	+	+	Not tested	+
Bcen2424–2964	(2)	+	+	Not tested	+
Daro3835	(2)	+	+	Not tested	+
Bpro4478	(2)	+	+	Not tested	+
L-2-haloacid dehalogenase					
PA0810	(1)	–	+	–	+
RPA4199	(1)	–	+	+	+
RPA2507	(1)	–	+	+	+
SAV737	(1)	–	+	+	+
SMc00103	(1)	–	+	+	+
Adeh3811	(2)	+	+	Not tested	+
Bcen2424–2051	(2)	–	+	Not tested	+
Jann1658	(2)	–	+	Not tested	+
Bpro0530	(2)	+	+	Not tested	+
Bpro4516	(2)	+	+	Not tested	+
RSc1362	(2)	–	+	Not tested	+
RHA1_ro00230	(2)	+	+	Not tested	+
SCO3446	(2)	–	+	Not tested	+

Dehalogenases identified in the first and second screens are labelled by (1) and (2), respectively.

(52%) produced soluble recombinant protein with high yields between 5 and 110 mg l⁻¹ of bacterial culture. The remaining 70 cloned but poorly expressed and/or insoluble (i.e. primarily expressed as inclusion bodies) proteins yielded less than 5 mg l⁻¹ of bacterial culture. The purity of high-yield preparations was 75–95%, and that of the low-yield samples was less than 35% (Fig. S1). Of the 30 gene sequences selected in the second round screen, 24 clones produced soluble recombinant protein (1–180 mg l⁻¹ of bacterial culture). These results are provided in Table S3 and summarized in Table S4.

Enzymatic assays of purified proteins

Every protein preparation in the initial screen (155 samples) was assayed on every substrate regardless of protein family and sample purity. Enzyme assays were designed to test for dehalogenase, carboxylesterase, thioesterase and phosphatase activities. Proteins purified for the second screen were only assayed for dehalogenase activity. The organohalogen substrates used for the dehalogenase assays are listed in the experimental section.

Overall, the initial enzymatic screen identified seven novel dehalogenases from 155 tested protein preparations: two from the ABH and five from the HAD superfamilies. The overall success rate of the initial screen was 7% (7/103). The 52 low-yield samples with no activity were

not counted because these negative results may have been caused by insufficient levels of recombinant protein in the assays.

Specifically, of the 110 selected ABH targets, 102 were successfully cloned (and thus assayable), and only one (SAV4779) dehalogenated the haloalkanes 1,2-dibromoethane, 1,3-dichloropropane and chlorocyclohexane (Table 2). One (RPA1163) could hydrolyse fluoroacetate, and more slowly chloroacetate. Nineteen proteins hydrolysed palmitoyl-CoA (Fig. S2) and two of these were also active on *p*-nitrophenyl palmitate (both are substrates for carboxylesterases and thioesterases) (Table 3). No other protein samples displayed multiple activities and none were active in the phosphatase activity screen. Representative assay data are shown in Table S2.

Of the 53 protein preparations of the HAD candidates, five possessed haloacid dehalogenase activity and 31 displayed phosphatase activity (Tables 2 and 3). These HAD dehalogenases exhibited a substrate profile similar to that of the previously characterized L-2-haloacid dehalogenases; they can hydrolyse various 2-haloacids, but none were capable of hydrolysing fluoroacetate (van der Ploeg *et al.*, 1991; Liu *et al.*, 1994). However, the substrate specificity of PA0810 was restricted to the short haloacetates only. No HAD dehalogenases displayed phosphatase activity (and *vice versa*), and no HAD proteins were active against the tested haloalkanes and esters.

Table 3. Relative absorbance of active proteins in esterase and phosphatase activity screens.

Superfamily	Gene locus	Purity	Substrates				Identified activity
			pNP-palmitate	Palmitoyl-CoA	pNPP + Mg ²⁺	pNPP + cations	
ABH	PA5513	High	0.76	3.2	1.0	1.2	Thioesterase
ABH	PA2949	Low	2.6	2.2	0.92	0.92	Esterase
ABH	PP4164	High	0.83	2.8	0.90	0.95	Thioesterase
ABH	PP2567	High	0.88	2.4	0.95	0.99	Thioesterase
ABH	PP2083	High	0.74	2.1	0.92	0.96	Thioesterase
ABH	RPA4646	High	1.2	1.8	0.81	0.94	Thioesterase
ABH	RPA3430	Low	0.89	2.0	0.85	0.98	Thioesterase
ABH	RPA1568	High	0.86	4.3	0.87	0.89	Thioesterase
ABH	RPA0348	Low	1.7	3.1	0.88	0.94	Esterase
ABH	RPA1212	Low	0.85	2.7	0.82	0.90	Thioesterase
ABH	RPA0996	Low	0.86	2.7	0.88	0.89	Thioesterase
ABH	SAV1548	Low	0.73	1.7	1.0	0.95	Thioesterase
ABH	SAV298	Low	0.89	2.4	0.94	1.1	Thioesterase
ABH	SMc03810	High	0.92	2.5	0.86	0.91	Thioesterase
ABH	SMc01273	High	0.82	2.7	0.88	0.91	Thioesterase
ABH	SMc00361	High	0.95	2.8	0.88	0.90	Thioesterase
ABH	SMc00147	Low	0.87	2.2	0.90	1.1	Thioesterase
ABH	SMc02818	High	0.87	1.8	0.90	0.92	Thioesterase
ABH	SMc04041	Low	0.86	3.0	1.1	0.94	Thioesterase
HAD	PA4458	High	0.73	1.1	1.4*	2.6*	Phosphatase
HAD	PA0335	High	0.94	1.1	2.5	2.6*	Phosphatase
HAD	PA5177	High	0.91	0.96	16	1.9	Phosphatase
HAD	PA0562	High	1.0	1.0	5.1	1.7	Phosphatase
HAD	PA3172	High	1.0	0.98	16	3.2	Phosphatase
HAD	PA2974	High	1.0	1.1	16	17	Phosphatase
HAD	PA0608	High	0.96	1.1	7.0	2.2	Phosphatase
HAD	PA3886	High	1.1	1.1	16	17	Phosphatase
HAD	PA4960	High	0.92	1.1	16	17	Phosphatase
HAD	PP0956	High	0.90	0.39	6.3	3.2	Phosphatase
HAD	PP4761	High	0.87	1.1	16	17	Phosphatase
HAD	PP0094	High	0.86	0.86	13	4.8	Phosphatase
HAD	PP1721	Low	0.80	1.0	2.5	1.9	Phosphatase
HAD	PP5147	High	0.86	1.3	9.7	2.5*	Phosphatase
HAD	PP0259	Low	0.79	1.3	16	7.5	Phosphatase
HAD	PP1764	High	0.80	0.91	16	10	Phosphatase
HAD	PP1907	Low	0.56	0.95	3.6	2.0	Phosphatase
HAD	PP1789	High	0.84	1.2	1.8	1.4*	Phosphatase
HAD	PP5231	Low	0.73	0.8	2.8	1.3*	Phosphatase
HAD	PP0416	Low	0.85	1.1	16	7.4	Phosphatase
HAD	PP4909	Low	0.54	0.92	16	11	Phosphatase
HAD	RPA4522	High	1.1	1.3	2.3	1.7	Phosphatase
HAD	SAV4726	High	1.1	1.2	2.0*	1.4*	Phosphatase
HAD	SAV6660	High	1.0	0.95	1.7	3.4*	Phosphatase
HAD	SMc00081	High	0.86	1.2	16	10	Phosphatase
HAD	SMc01276	Low	0.83	1.1	7.1	6.6	Phosphatase
HAD	SMc00079	High	0.88	1.1	8.3	2.3	Phosphatase
HAD	SMc01617	High	0.96	1.2	1.6	1.7*	Phosphatase
HAD	SMc00910	High	1.0	1.1	16	17	Phosphatase
HAD	SMc01494	Low	0.99	1.2	9.1	3.9	Phosphatase
HAD	SMc04299	High	0.97	0.99	1.5*	0.94	Phosphatase

The values represent the ratio of the measured absorbance for each protein relative to the no protein control, at the 3-hour time point. For the identification of phosphatase and carboxylesterase activities, a ratio of 1.3 was used as cut-off. For confirmation of thioesterase activity, a more stringent cut-off of 1.7 was used because contaminating expression host proteins in low-purity samples appeared to contribute to background signals. Signals emerging only at the 24-hour time point are marked by *. The high (75–95%) and low (less than 35%) protein purities correspond to that of the high- and low-yield protein preparations respectively.

Refined criteria for identification of true dehalogenase sequences

Using the results from these enzymatic assays, additional search criteria were sought by comparing the

sequence alignments of biochemically confirmed dehalogenases with those of esterases or phosphatases. A detailed analysis of all sequences is presented in the Supporting Information; the main conclusions are provided below.

(A)

Haloalkane dehalogenases (Type I)	
DhmA	52 L H G E P T 120 F V Q D W G S L 146 A N G F L 160 A F Y A W R 211 A F P Q L V P 249 R D P - - - - I L 278 S H F I
DmbB	52 L H G E P T 120 F V Q D W G S L 146 A N G F L 160 P F Y V W R 211 A F P R L V P 249 R D P - - - - I L 278 S H F I
DhmA	52 M H G E P S 120 F C Q D W G G L 146 S N T G L 159 G F E A W L 210 I F P A L V P 248 A D P - - - - I T 277 G H F V
DhIA	53 L H G E P T 121 V V Q D W G G F 147 M N A C L 171 G F T A W K 221 K F P K M V A 259 K D K - - - - L L 288 G H F V
Haloalkane dehalogenases (Type II)	
LinB	35 Q H G N P T 105 V V H D W G S A 131 M E A I A 150 L F Q A F R 206 S W P R Q I P 243 A E P G - - - - A L T 271 A H F I
DmbA	36 Q H G N P T 106 V L H D W G S A 132 M E A I V 151 V F Q G F R 207 S W P R N L P 244 A E P G - - - - A I I 272 V H F V
* SAV4779	25 L H G N P T 89 V G H D W G G A 115 T E T I V 134 L F R A I K 190 R W P R S M P 228 F A P G P G A M M 261 G H H T
DbjA	35 L H G N P T 100 V A Q D W G T A 126 M E F I R 156 V F R K F R 212 A F P R E L P 249 G E P G - - - - A L V 279 L H Y L
DmIA	42 L H G N P T 107 V A Q D W G T A 133 M E F I R 154 M F K A L R 210 R L P R E M P 247 G D P G - - - - A L I 277 A H Y L
DhaA	38 L H G N P T 103 V I H D W G S A 129 M E F I R 148 T F Q A F R 204 R F P N E I P 241 G T P G - - - - V L I 271 L H Y L
Fluoroacetate dehalogenases	
Fac-Dex FA1	31 L H G F P Q 101 V G H D R G G R 127 L D I I P 146 A Y W H W Y 203 A A I H G S C 240 G S A G - - - - L M 270 G H F F
DehH1	32 L H G F P Q 102 V G H D R G G R 128 M D I V P 147 S Y W H W Y 204 A M I H G S C 241 G S K G - - - - Q M 271 G H F F
* RPA1163	37 L H G F P Q 107 A G H D R G A R 133 L D I L P 152 K I Y H W S 210 M R R H V M C 249 G A S G - - - - I A 279 G H F L
Esterases	
* PP2567	35 L H G W L D 101 L A H G M G T G 127 D G M G A 148 Y R L K R M 206 D G Y C W R H 245 G R Q G - - - - A 274 G H H F
* PP4164	33 L H G W L D 99 L G H S L G A I 125 I D G V I 146 L Q A Q L R 195 G V Y S W R S 234 A A D G - - - - M 259 G H H L
* SAV298	35 L H G S A S 101 A G N S L G G N 127 V N A T G 132 - - - - Y P 186 H Q L M S R P 225 S - - - - - A 249 G H L L
* RPA0996	83 I H G G K D 149 I G H S M G G M 175 L D G V T 189 V H E R M I 247 G S Y S W K F 286 A G E S - - - - F 315 H W L
* RPA1568	29 I H G A G F 94 I G H S M G S L 120 I G - - - - 122 - - - - T T S 177 E A N R P G V 213 G E R D - - - - M M 243 G H T M
* SAV1548	30 I H G W T A 93 V G H S M G G M 119 V N S I S 124 - - - - R M 168 I R A S S A T 207 G Y Y D - - - - I 236 G H E L
* PA2949	68 I H G F G A 134 A G N S M G G H 160 I D N A G 165 - - - - - V 218 A V A A S A F 255 G D R D - - - - R V 285 G H V P
* PA5513	32 L H G N G F 101 C G H S F G G V 127 L D P V L 143 S E T L G L 201 D G G V E L K 240 G D R T - - - - - Y 269 G H C F
PFE	26 S H G W L L 92 V G F S M G G G 119 L G A V T 128 - - - - - Q 181 L Q I A L L A 220 G D T - - - - Q I V 251 P H G F
* RPA3430	25 I H G F A S 96 M G Y S M G A R 122 G - - - - 123 - - - - - 162 F A D Q T R S 201 G T D D - - - - - V 229 D H - -
* RPA0348	91 I M G L G G 157 L G Y S M G S T 181 - - - - - 181 - - - - - 212 A T T H W K T 235 G T A D N - - - - - V 265 T H - -
* SMC03810	27 L H G L C G 92 G I S M G A A 118 A R P - - - - 121 - - - - - A W L T 177 F A R E P I E 216 H E R D - - - - - L 245 K A E D
* RPA4646	24 V H G F T G 92 A G L S M G G T 118 I N A A V 126 - - - - - 157 L A Y D Q V P 193 S R V D - - - - - H V L 224 S Y H V
* SMC00147	90 A H G W G S 154 - - - - - G H S F G G A 180 K L V L I 191 - - - - - 221 - - - - - 244 A E D D - - - - - K E V 273 L G H R
* SMC04041	49 L Q G R N E 125 V A H S M G A L 151 L A P F V 160 Q A I G E R 218 Q L R V G P P 257 P T A D K - - - - - L V 288 H E L L
* SMC02818	34 S H G L A E 111 F G H S M G L 137 W N S N F 151 - - - - - Q L L L A 208 C G F D P T V 247 G S A D P - - - - - S T L 279 H E G M
* SMC00361	50 L G G Y R S 118 I G S M G - - - - - 140 H G E - - - - 143 - - - - - 186 P S E Y S P E 225 G M R D P - - - - - D V 256 G D H R
* SMC01273	43 R P V L W Y 144 F G H S M G G H 170 F A P T V 181 - - - - - 209 - - - - - 220 G K A D - - - - - S F L 251 H E R Y
* RPA1212	61 I H S V N A 132 L A L S L A S E 158 I S P T G 172 - - - - - T P G T 228 T T H Q P G A 267 G V R G - - - - - 295 G A F P
* PP2083	61 L H G S F S 138 V G H S L G G T 167 F G T Q I 174 - - - - - V Y W P 230 K Q N D W W A 269 G G E R - - - - - 297 A A Q V

(B)

Type I Haloalkane dehalogenase	Type II Haloalkane dehalogenase	Fluoroacetate dehalogenase	Legend
L H G E P T V V Q D W G G F L M N A G F T A W K K D K G H F	L H G N P T V G H D W G G A L T E T L F R A I K F A P G H H	L H G F P Q A G H D R G A R V L D I K I Y H W S G A S G H F	Catalytic triad
H G P Small-x-Nuc-x-G Acid Acid His			Primary halide binding
			Secondary halide binding
			Acyl binding
			Structural motifs
			Confirmed/putative active site residues

Fig. 2. Sequence analyses of ABH targets in the initial screen. The residues' functions were inferred from the crystal structures of the haloalkane dehalogenases DhIA, LinB and DhaA, and the aryl esterase PFE. A. The sequence alignment of ABH targets identified in the initial screen (*). The catalytic triad (Nuc-Acid-His) and several structural motifs are well conserved. In haloalkane and fluoroacetate dehalogenases, the nucleophile is a fully conserved Asp, while it is generally a Ser in esterases. The catalytic acid, which is either an Asp or Glu, is located either 24 residues downstream of the catalytic nucleophile, or approximately 29 residues upstream of the catalytic His base found near the C-terminus respectively. The numbers mark the residue number for the residue to their right. B. The key catalytic motifs extracted from the sequence analysis.

For the identification of haloalkane dehalogenases in the ABH superfamily, the Asp catalytic nucleophile serves as an important selection criterion because it is fully conserved and essential for dehalogenation (Fig. 2). The pair of primary halide-binding residues can act as an additional haloalkane dehalogenase indicator. It comprises: a Trp at the nucleophile + 1 position; and either a Trp that aligns with Trp175 of DhIA (a type I haloalkane dehalogenase) or an Asn equivalent to Asn28 of SAV4779 (a type II haloalkane dehalogenase) within the upstream His-Gly-Asn-Pro motif.

For the identification of fluoroacetate dehalogenases in the ABH superfamily, only a few active site residues are considered for distinguishing their activity. Such a restriction is imposed in order to reduce the number of false

negatives during sequence screening given our current rather limited level of understanding. The residues comprise an Asp acting as nucleophile and two Arg at the nucleophile + 1 and nucleophile + 4 positions (Fig. 2).

For the identification of L-2-haloacid dehalogenases in the HAD superfamily, two motifs could serve as primary identifiers owing to the distinct functional requirements of such enzymes (Fig. 3). At the N-terminus, the dehalogenases harbour Asp-X-[Tyr/Phe] instead of the Asp-X-[Asp/Glu] found in phosphatases. Near the C-terminus, the dehalogenases possess the oxyanion hole motif Ser175-Ser176-Asn177 (L-Dex YL) or Ala-Ala-His, which contrasts with the phosphatase Mg²⁺-binding pair of carboxylates ([Asp/Glu]-Asp, Gly-[Asp/Glu]-X₃-Asp or Gly-[Asp/Glu]-X₄-Asp); the aligning residues are under-

(A)

L-2-Haloacid dehalogenases

DehIVa	8	C V	F D A Y G T L L	D V H	39	M L W R Q R Q	116	A I L S N G N D	174	F V S S N - A W D L G
HadL	7	I V	F D L Y G T L Y	D V H	38	R L W R Q K Q	115	G I L S N G S H	173	F V S S N - A W D A S
LDexYL	7	I A	F D L Y G T L F	D V H	38	A L W R Q K Q	115	A I L S N G S P	173	F V S S N - A W D A T
DehH2	7	I A	F D M Y G T L Y	D V H	38	V L W R Q K Q	115	A I L S N G S T	173	F V S S N - P W D V S
DhIB	5	V V	F D A Y G T L F	D V Q	36	Q V W R Q K Q	111	A I L S N G A P	169	F V S S N - G F D V G
* RPA2507	12	L L	F D V F G T L V	D W R	49	D V W R Q A Y	128	A P L S N G N V	184	L V A A H - N Y D L R
* PA0810	4	I L	F D V F G T L V	D W R	41	D R W R Q Q Y	118	A A L S N G N T	174	L C A A H - N Y D L K
* SMC00103	6	Y V	F D A Y G T L F	D V H	37	E L W R A K Q	113	A I L S N G S P	171	F Q S S N - R W D I A
* RPA4199	5	V V	F D A Y G T L Y	D I Q	36	Q I W R I K Q	111	A I L S N G S P	169	F V S S N - P W D V A
* SAV737	44	V V	F D V L G T L V	D E P	80	S L W Q Q H I	156	I G L S N A S R	212	M V A A H - A W D L R

Phosphatases

* PA4960	218	A V	F D M D S T L I E - A	243	E K V A A I T	316	A R Q L Q A K L	374	A V G D G - A N D L P
* PP4909	204	A V	F D M D S T L I E - A	229	E Q V A A I T	302	A R Q V Q A R L	360	A V G D G - A N D L P
* SMC01494	83	L I A	D M D S T M I G - Q	108	E K V A A I T	182	T G P I A E K L	240	A V G D G - A N D L G
* PP1764	6	V L	F D M D G T L L D T A	32	R G L P A V D	107	G V V T N K P V	165	F V G D D - L R D I E
* PA3172	9	V L	F D M D G T L L D T A	35	H G L P P V D	110	G V V T N K P V	168	F I G D D - L R D I E
* PP0956	15	A V	F D V D G V L T D G R	41	N T L D G Q G	69	V E R R A R N L	108	Y L G D D - L P D L P
* PA4458	20	A I	F D V D G V L T D G K	46	N T L D G H G	74	V E R R A R N L	113	Y L G D D - L P D L P
* PA0608	16	V M	F D L D G T L V D S V	42	F G R P P A G	121	A L I T N K P E	179	F V G D S - R N D V L
* PP0416	16	V M	F D L D G T L I D S V	42	L G R P P A G	121	A L I T N K P E	179	F V G D S - R S D V Q
* SAV4726	6	V I	F D L D G V L V D S E	32	L G H P T S Y	106	C V A S S G S H	166	V V E D S - P L G V Q
* SMC00079	7	I I	F D C D G V L V D S E	33	A G Y P I S I	104	C I C S N S T S	165	V I E D S - V H G I H
* PA0562	8	L I	S D C D G V L V D S E	34	Y A P L E E L	104	A V A S N S R R	163	V V E D S - S T G T R
* PA2974	7	L V	F D W D G T L V D S I	33	T A L P E R G	108	A V A T G K S R	165	M I G D S - P F D L E
* PP1907	8	L I	F D W D G T L A D S I	34	A G E A Q S S	108	A V A T G K A R	165	M V G D S - A F D L Q
* SMC01276	4	V L	F D C D G T L V D S A	30	F D L D R P T	109	G A V T G K S R	166	V I G D A - I Y D M Q
* PP0094	6	I L	F D L D G T L T D P R	32	L G I D E P D	105	Y I A T S K P W	163	M I G D R - K H D L I
* PP5231	6	I T	F D L D D T L W D T A	45	E H L F A I R	125	G V V T N G N A	178	H I G D H P G D D I A
* PP4761	9	W V	F D M D G T L T V A V	33	- - - I P A E	90	G I L T R N A R	149	M V G D Y - R F D L D
* RPA4522	6	L I	F D V D G T L A E T E	33	Q L P W H W D	114	G I A T T S L	175	A F E D S - T N G V R
* PP5147	4	A L	F D L D N T L L G G D	29	G I L D P V A	108	V I I T A T N R	164	M L E N G - - F D L E E
* PA0335	4	A L	F D L D N T L L A G D	29	G L V D A A E	108	V I I T A T N R	164	L D E N G - - L D L E E
* PP1721	4	A I	F D L D E T L I H G D	29	G W V D G K D	108	L I I S A S G T	164	L D Q E Q - - E N L E E
* SAV6660	10	V V	L D I E G T T S A T G	46	D P E V A R A	127	H V Y S S G S V	187	F L S D R - P G E L D
* PP0259	12	V L	L D M D G T L L D L H	38	Y A E L H G V	112	V M I T N A H R	170	F I D D T - L A I L R
* PA5177	12	V L	L D M D G T L L D L H	38	Y A E R N G L	112	A L I T N A H R	170	F I D D S - L P I L R
* SMC04299	6	L A T	D Y D G T L A D Y G	44	R E L P D L K	132	I F N K G A V M	193	A L K D T A D V V L E
* SMC00081	9	I V	F D I G K V L I H Y D	36	R S W F A N	109	T M L T N F A S	167	F I D D S - A A N V E
* SMC01617	20	F L V	D Q Y G V L R D I G R	59	G K R S A E N	164	P C L C T N P D	226	A I G D S V E H D I A
* SMC00910	18	V L C	D V W G V L H N G I	57	P R P H P G V	155	P F I C A N P D	221	A I G D G M P T D V K
* PA3886	34	F L L	D L W G V V M D G T	73	S S S V A E M	173	P F L C A N P D	235	F V G D S L V T D V P
* PP1789	14	L V	F D C D G V L D D S N	44	A A Q A L A D	118	L I V S G G D Q	171	F I G D S - K Y D F K

F D x Ar G T L x D
D

(B)

Haloacid Dehalogenase	F D A Y G T L L	L W R A	L S N G	Q S S N - R W D I	V A A H - N Y D L
Phosphatase	F D M D G T L	- - - -	- - - -	I D D S - A A N V	V E D S - P L G V
				V G D G - A N D L	V G E G - M N D V
				V G D S L V T D V	

Legend

- Catalytic nucleophile
- Catalytic base and oxyanion hole
- Catalytic base
- Oxyanion hole
- Halide binding
- Carboxylate binding
- General acid-base
- Mg binding
- Some conserved residues

Fig. 3. Sequence analyses of HAD targets in the initial screen.

The conserved residues in both enzyme groups are presented.

A. The sequence alignment of HAD targets identified in the initial screen (*). The overall sequence similarity is higher among the L-2-haloacid dehalogenases. The catalytic Asp nucleophile is located at the N-terminal motif conserved in both enzyme groups. At position nucleophile + 2, the L-2-haloacid dehalogenases generally possess an aromatic residue while most phosphatases carry a second Asp. Near the C-terminus, the dehalogenases and phosphatases have distinct sequence motifs containing catalytically important residues. The numbers mark the residue number for the residue to their right.

B. The key catalytic motifs extracted from the sequence analysis. The conserved motif variations within each subfamily are also shown.

lined. These two carboxylate pair motifs may be applied to reliably filter out probable phosphoryl transferases before further screening with dehalogenase-specific residues, because they were found in numerous phosphoryl transferases (Koonin and Tatusov, 1994; Burroughs *et al.*, 2006; Kuznetsova *et al.*, 2006).

Retrospective analysis of sequences from the initial screen, using refined search criteria

Applying the Asp nucleophile and either pair of halide-binding residues (Trp and Trp/Asn) to filter the initial list of ABH sequences would have narrowed the list to only four

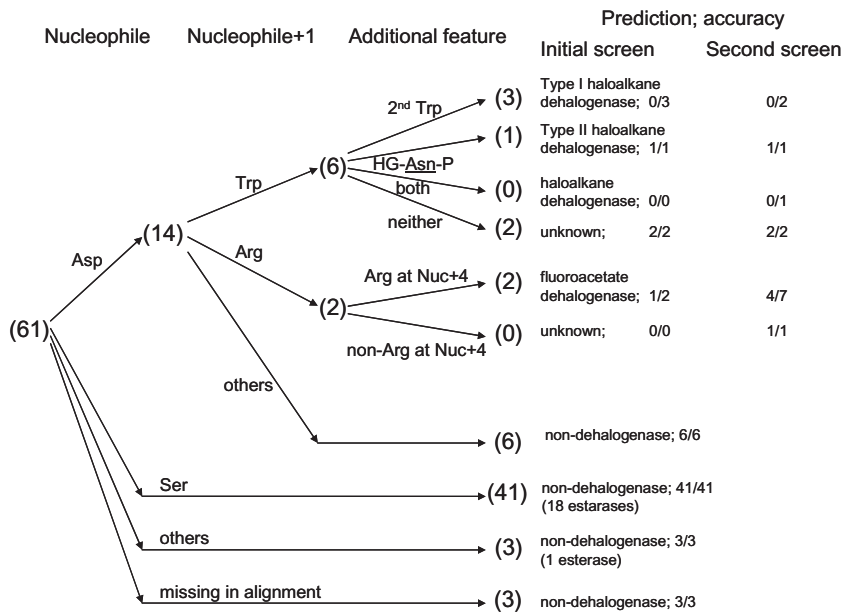


Fig. 4. Decision tree for ABH sequences. The selection process for ABH dehalogenases is summarized. For haloalkane dehalogenases, it comprises the Asp nucleophile and a pair of halide-binding residues. For fluoroacetate dehalogenases, it comprises the Asp nucleophile and a pair of Arg. Numbers in parentheses denote the outcome from the retrospective analysis of targets in the initial screen, accounting only for high-yield and/or active proteins. Results from the validation screen are also presented.

of the initial 110 candidates as true potential haloalkane dehalogenases (Fig. 4; only 61 sequences are shown because they correspond to high-yield and/or active protein samples), increasing the chances of finding SAV4779 to 1 in 4. However, the absence of dehalogenase activity in the remaining three (which all carry the halide-binding Trp pair of type I haloalkane dehalogenases) indicates that these features alone are not sufficient to pinpoint haloalkane dehalogenases. While the search may be refined by incorporating residues that line the active site or serve as supplementary halide binders, the low sequence conservation and the possibility of misalignment of true structural equivalents of the more diverse ABH proteins may lead to the dismissal of genuine dehalogenases. Nevertheless, the haloalkane dehalogenase search space could already be greatly reduced just by considering these catalytic residues. The criteria may be further customized to vary their stringency, such as by accommodating functional mutations as demonstrated in (Kennes *et al.*, 1995; Damborsky *et al.*, 1998).

Applying the Asp nucleophile and the Arg pair at positions nucleophile + 1 and nucleophile + 4 as additional fluoroacetate dehalogenase search criteria to filter the initial list of ABH sequences would have narrowed the list to only 2 of the initial 110 candidates as true potential fluoroacetate dehalogenases, increasing the hit rate for RPA1163 from 1 in 110 to 1 in 2 (Fig. 4). Only PA2086 would be falsely recognized as a potential fluoroacetate dehalogenase among all selected ABH sequences; further investigation should identify additional functionally important residues suitable to refine the search criteria for these enzymes to achieve even higher prediction accuracy.

Applying reliable HAD phosphoryl transferase sequence motifs to filter the initial list of HAD sequences would have narrowed the list to only 6 of the initial 53 candidates (Fig. 5, sum of the last two entries), eliminating the bulk of non-dehalogenase HAD sequences; only the 42 sequences that correspond to high-yield and/or active protein samples are shown in Fig. 5. The search space could be further reduced to 5 by dismissing the

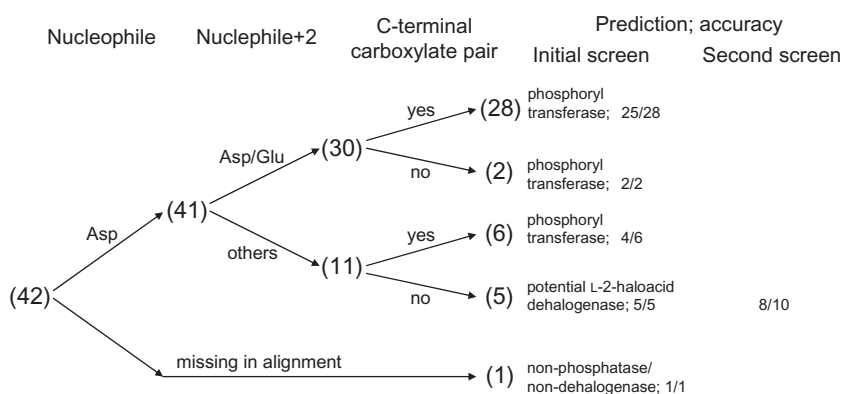


Fig. 5. Decision tree for HAD sequences. The selection process for HAD dehalogenases is summarized. The elimination of probable phosphoryl transferases (carrying at least one of the two carboxylate pair motifs) was sufficient for obtaining L-2-haloacid dehalogenases with high accuracy. Numbers in parentheses denote the outcome from the retrospective analysis of targets in the initial screen, accounting only for high-yield and/or active proteins. Results from the second screen are also presented.

proteins lacking the Asp nucleophile. Interestingly, all sequences remaining after filtering for phosphoryl transferase sequence motifs emerge as genuine L-2-haloacid dehalogenases even before any dehalogenase fingerprints are applied. In principle, one may further screen these sequences using the conserved dehalogenase motifs. However, this may discard novel dehalogenases with undiscovered sequence variations such as the Ala-Ala-His oxyanion motif observed in PA0810, RPA2507 and SAV737 (Fig. 3). Therefore, it may be more beneficial to apply the more stringent criteria in cases where a large number of sequences still remain after eliminating the putative phosphoryl transferases.

Second screen to validate new search criteria

To further demonstrate the effectiveness of the identification criteria developed in this work, they were applied to screen 14 additional microbial genomes. The procedure began with the same BLASTP sequence similarity searches, followed by a screen using some of the identified sequence fingerprints: an Asp nucleophile and a Trp at position nucleophile + 1 for haloalkane dehalogenases; an Asp nucleophile and an Arg at position nucleophile + 1 for fluoroacetate dehalogenases; and the absence of both carboxylate pair motifs for L-2-haloacid dehalogenases. This process resulted in 30 candidates, 13 of which showed dehalogenase activity in enzymatic screens (Table 2), comprising 1 haloalkane dehalogenase, 4 fluoroacetate dehalogenases and 8 L-2-haloacid dehalogenases. Accounting only for the 24 targets that yielded soluble proteins, the hit rate for haloalkane dehalogenases, fluoroacetate dehalogenases and L-2-haloacid dehalogenases were 1 in 6 (Fig. 4, sum of the first four entries), 4 in 8 (Fig. 4, sum of the fifth and sixth entries) and 8 in 10 (Fig. 5), respectively. This trend resembles that of the retrospective sequence screens, where the accuracy improves significantly from haloalkane dehalogenases (Table S5) to fluoroacetate dehalogenases (Table S6) to L-2-haloacid dehalogenases.

Novel defluorinating L-2-haloacid dehalogenases in the HAD superfamily

Intriguingly, 4 novel L-2-haloacid dehalogenases (Adeh3811, Bpro0530, Bpro4516 and RHA1_ro00230) were able to hydrolyse fluoroacetate. This is the first report of HAD dehalogenases with such capabilities, as the direct hydrolytic cleavage of the aliphatic carbon-fluorine bond was previously thought to be exclusive to the fluoroacetate dehalogenases of the ABH superfamily. With respect to sequence similarities, the HAD defluorinases did not appear to be distinct from the non-defluorinating HAD members (Table S7). Investigations

are currently underway to elucidate their structural features that enable defluorination.

Evaluation of E-values obtained from BLASTP

This work has developed sequence screening procedures for reliably discriminating against non-dehalogenases. Obviously, however, dehalogenases can only be discovered if they exist within the genomes of interest. *E*-values measure the statistical significance of the hits returned in a BLASTP search and are calculated from their overall sequence similarity with query sequences. They were thus examined in an attempt to identify prominent signs that may indicate the presence of dehalogenases within a list of BLASTP results. A detailed discussion is presented in Supporting Information.

In the genome-wide BLASTP searches for haloalkane dehalogenases, the true positives possessed highly significant *E*-values (less than 10^{-60}) compared with proteins that were inactive against the screening substrates (greater than 10^{-25}) (Fig. 6A). This suggests that *E*-values may serve as reliable indicators for the activity of these enzymes.

In the fluoroacetate dehalogenase searches, the *E*-values were also clearly divided into two clusters: they were either highly significant (smaller than 10^{-45}), or much less significant (larger than 10^{-20}) (Fig. 6B). All five identified fluoroacetate dehalogenases belong to the former group. However, an additional five proteins with similar *E*-values were inactive on the tested haloacetates. Meeting the very small *E*-value requirement thus appears to be necessary but not sufficient for confirming the presence of fluoroacetate dehalogenases in microbial genomes.

In the L-2-haloacid dehalogenase searches, 8 of the confirmed enzymes possessed highly significant *E*-values (less than 10^{-25}) while the remaining 5 had *E*-values (from 10^{-14} to 10^{-5}) that were only marginally smaller than those of the non-dehalogenases (Fig. 6C). Such *E*-value distribution suggests that there are at least two sequence subfamilies among the HAD dehalogenases, which appears to correlate with their oxyanion hole motifs (Ser-Ser-Asn versus Ala-Ala-His). However, the *E*-values of the four defluorinating members are not distinct from those of the non-defluorinating members. Taken together, our results show that *E*-values may serve as reliable indicators for detecting L-2-haloacid dehalogenases, and the analyses should be made using sequences from different subfamilies as search query.

Dehalogenase kinetics

The kinetic parameters for several dehalogenases were determined for comparison with previously characterized

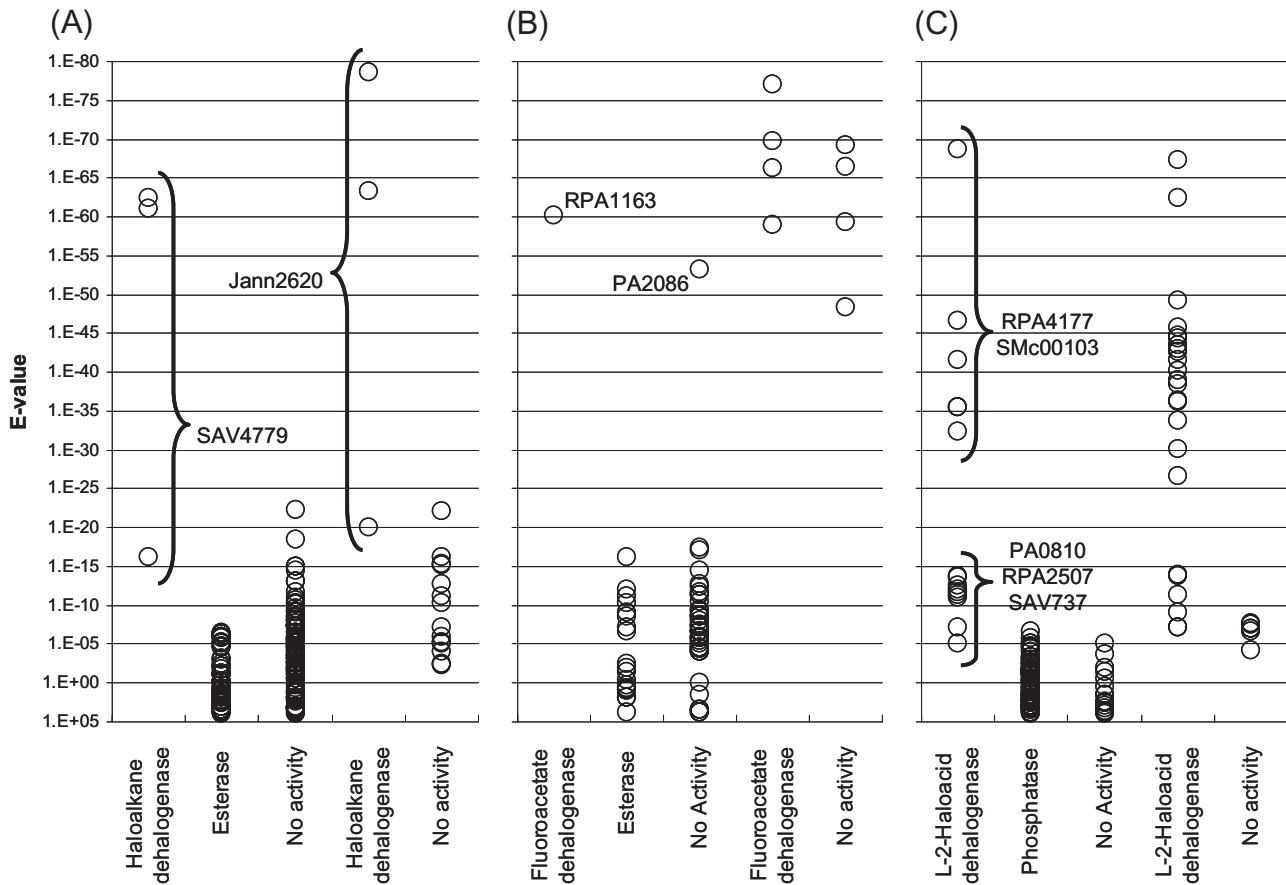


Fig. 6. *E*-value analyses of selected targets.

The *E*-values of the target genes were categorized by their identified enzymatic activity (*x*-axis label). In each panel, the first three columns contain *E*-values from the initial screen and the last two columns display results from the second screen.

A. ABH sequences searched by haloalkane dehalogenases DhIA, LinB and DhaA.

B. ABH targets obtained from the fluoroacetate dehalogenase DehH1.

C. HAD targets retrieved using L-2-haloacid dehalogenases DhIB, HadL and DehH2.

The *E*-values of the dehalogenases are generally much smaller than those of the non-dehalogenases. However, in the fluoroacetate dehalogenase searches (B), 5 non-dehalogenases (PA2086 from the initial screen and 4 proteins from the validation screen) emerged with significant *E*-values that are comparable to the active enzymes. Moreover, in the L-2-haloacid dehalogenase searches (C), 5 dehalogenases (3 initial targets and 2 validation targets) were retrieved at considerably less significant *E*-values. These sequences were also the best hits within that particular BLASTP search. The *E*-values distribution among the HAD dehalogenases suggests that there are at least two subtypes differing in overall protein sequence.

enzymes (Table 4). The novel fluoroacetate dehalogenases displayed significantly lower maximum specific activity for both fluoroacetate and chloroacetate, compared with DehH1 and FAc-Dex FA1. However, the substrate affinity constants (K_M) for fluoroacetate of RPA1163 and FAc-Dex FA1 were of the same order of magnitude.

The specific activities of the novel L-2-haloacid dehalogenases towards chloroacetate, which ranged from 1.5 to 100 $\mu\text{mol min}^{-1}$ per mg of protein, were comparable to those of the previously studied family members. The K_M values for chloroacetate were generally in the μM range. For those that exhibited the novel defluorination capability, a strong preference (i.e. higher specific activity as well as lower K_M) for chloroacetate hydrolysis was generally observed. Compared with the novel ABH fluoroacetate

dehalogenases, these novel HAD defluorinases possessed a much higher K_M for fluoroacetate even though their maximum specific activities were in the same order of magnitude (Table 4).

Conclusions

This work has significantly increased the sequence pool of microbial hydrolytic dehalogenases through the biochemical confirmation of 20 novel dehalogenases from diverse organisms. The discovery of four HAD defluorinases highlights the importance of carrying out broader functional screens to mine natural diversity. Additionally, the identification of numerous esterases and phosphatases should assist genomic annotations of sequences in the ABH and

Table 4. Quantitative dehalogenase activity assays.

Enzyme	Substrates							Relative rate ClAc/FAc	
	Fluoroacetate			Chloroacetate					
	Specific activity ($\mu\text{mol min}^{-1} \text{mg}^{-1}$)	K_M (μM)	Note	Specific activity ($\mu\text{mol min}^{-1} \text{mg}^{-1}$)	K_M (μM)	K_i (mM)	Note		
Fluoroacetate dehalogenase									
RPA1163	(1)	0.20 \pm 0.01	3 300 \pm 200	a,f	0.041 \pm 0.002	1 400 \pm 400		a,f	0.21
Alr0039	(2)	0.108 \pm 0.008	n.d.	c,d,e	0.12	n.d.		c,d,e	1.1
Bcen2424–2964	(2)	0.102 \pm 0.008	n.d.	c,d,e	n.d.	n.d.			
Bpro4478	(2)	0.38 \pm 0.04	n.d.	c,d,e	0.041	n.d.		c,d,e	0.11
DehH1		86.40		p	18			p	0.21
FAc-DexFA1		11.00	5 100	q	0.43			q	0.04
L-2-haloacid dehalogenase									
RPA4199	(1)	n.a.			25	28	1	a,f	
SMc00103	(1)	n.a.			52	88	40	b,g	
					9.8	17		b,g,z	
Adeh3811	(2)	0.401 \pm 0.008	13 500 \pm 300	a,f	25 \pm 3	10 \pm 4		c,j	62
Bcen2424–2051	(2)	n.a.			12.0 \pm 0.3	17 \pm 2		c,f	
Jann1658	(2)	n.a.			10 \pm 3	7 \pm 4		c,g	
Bpro0530	(2)	0.48 \pm 0.06	n.d.	c,d,e	29 \pm 1	4 \pm 1		c,f	61
Bpro4516	(2)	0.16 \pm 0.02	19 000 \pm 2000	a,f	64 \pm 4	11 \pm 2		c,h	400
RSc1362	(2)	n.a.			100 \pm 10	34 \pm 9		c,i	
RHA1_ro00230	(2)	0.678 \pm 0.004	20 000 \pm 2000	a,f	1.5 \pm 0.2	19 000 \pm 2000		a,f	2.2
DhIB		n.a.			55.5			r	
LDexYL		n.a.			78	1 100		s	

The activities for selected dehalogenases were quantified using different techniques; the specific activities represent the maximum values unless indicated otherwise. The enzymes identified in the initial and second screens are labelled by (1) and (2) respectively.

Notes: a, calorimetry; b, spectrophotometry; c, ion chromatography; d, rate estimation using 10 mM substrate; e, HEPES pH 8; f, Tris-SO₄ pH 8.5; g, Tris-NO₃ pH 8.5; h, CHES pH 9.5; i, CAPS pH 10; j, CAPS pH 10.5; p, Liu *et al.*, 1998; q, Kurihara *et al.*, 2003; r, van der Ploeg *et al.*, 1991; s, Liu *et al.*, 1994; Kurihara *et al.*, 1995; z, dechlorination activity on S-2-chloropropionate.

n.a., no detectable activity; n.d., not determined.

HAD superfamilies. Most importantly, the developed sequence-based selection methods will facilitate the identification of hydrolytic dehalogenases from these superfamilies of proteins. These findings will accelerate the discovery of novel and potentially more powerful dehalogenases from the rapidly expanding genome sequence database. Knowledge of the ability and diversity of dehalogenating enzymes is vital for developing effective detoxification methods at organohalogen-contaminated sites, and for guiding regulations for industrial production and use of organohalogenes.

Experimental procedures

Strains, reagents and chemicals

The bacterial strains and genomic DNA were either purchased from ATCC (Manassas, VA) or obtained directly from research laboratories. The PCR primers were synthesized by Integrated DNA Technologies (Coralville, IA). Terrific Broth pellets were purchased from EMD Chemicals (Gibbstown, NJ) and the Bradford assay reagent from Bio-Rad (Hercules, CA). Lysozyme, proteinase K, agarose, glycerol, ampicillin, kanamycin, SDS and IPTG were obtained from BioShop (Burlington, ON, Canada) and all other chemicals from Sigma-Aldrich (St. Louis, MO). Ni-NTA resin and the QIAquick PCR

Purification Kit were purchased from Qiagen (Mississauga, ON, Canada), the In-Fusion PCR Cloning Kit was purchased from Clontech (Palo Alto, CA). All chemicals were of the highest grade commercially available and all concentrations reported were the final concentrations in the reaction mixtures unless stated otherwise. The commercially available kits were used according to the manufacturers' instructions.

Target selection

In the initial screening study, five microbial genomes were screened for dehalogenase sequences: *Pseudomonas aeruginosa* PAO1, *Pseudomonas putida* KT2440, *Rhodospseudomonas palustris* CGA009, *Streptomyces avermitilis* MA-4680 and *Sinorhizobium meliloti* 1021 (chromosomal genes only). The selected gene loci were prefixed by PA, PP, RPA, SAV and SMC respectively. The following dehalogenases were used as BLASTP (Altschul *et al.*, 1990) search queries for each superfamily against the UniProt Knowledgebase: the ABH superfamily was searched with the haloalkane dehalogenases DhIA (UniProt accession number: P22643), LinB (P51698) and DhaA (P59336) and the fluoroacetate dehalogenase DehH1 (Q01398); and the HAD superfamily with the L-2-haloacid dehalogenases DhIB (Q60099), HadL (Q52087) and DehH2 (Q01399). The E-value cut-offs ranged from 10⁻³ to 10⁻¹ depending on the overall alignment quality. An additional text search in the PEDANT database (Walter *et al.*, 2009) was performed using 'haloalkane dehalogenase',

'haloacetate dehalogenase' and 'haloacid dehalogenase' to append several more distantly related superfamily members to the activity screens. The target list of the initial screen, including the selection parameters, is found in Tables S1A and S2A.

In the second screening study, the following 14 genomes were screened and the parenthesized codes denote the gene loci prefixes: *Anabaena* sp. PCC7120 (Alr or All), *Anaeromyxobacter dehalogenans* 2CP-C (Adeh), *Azotobacter vinelandii* AvOP (AvinDRAFT), *Burkholderia cenocepacia* HI2424 (Bcen2424), *Burkholderia vietnamiensis* G4 (Bcep1808), *Chromohalobacter salexigens* DSM3043 (Csal), *Dechloromonas aromatica* RCB (Daro), *Jannaschia* sp. CCS1 (Jann), *Polaromonas* sp. JS666 (Bpro), *Pseudomonas fluorescens* Pf-5 (PFL), *Ralstonia eutropha* JMP134 (Reut_A), *Ralstonia solanacearum* GMI1000 (RSc), *Rhodococcus* sp. RHA1 (RHA1_ro) and *Streptomyces coelicolor* A3(2) (SCO). The BLASTP searches employed the same queries as the initial screen, but the resulting sequences were filtered using the refined criteria developed herein. Not all potential candidates were selected; the 30 targets that were chosen are listed in Table S1B and S2B.

Cloning, overexpression and purification of selected targets

The target genes were amplified by PCR using primers carrying 5'-overhangs (5'-TTG-TAT-TTC-CAG-GGC-3' on the sense and 5'-CAA-GCT-TCG-TCA-TCA-3' on the antisense primer) and cloned into a modified pET-15b vector (Novagen, Madison, WI) (Zhang *et al.*, 2001). Protein purification was carried out as described (Zhang *et al.*, 2001), with the following modifications: overexpression in Terrific Broth was induced with 1 mM IPTG when OD₆₀₀ reached approximately 1.2; no NP-40, PMSF and benzamide were added; the centrifugal supernatant was directly passed through the Ni-NTA resin, washed with 300 ml of buffer containing 25 mM imidazole and eluted with the same buffer containing 250 mM imidazole; excision of the His₆-tag by the tobacco etch virus protease and the second pass through the Ni-NTA resin were omitted. The purified proteins were concentrated, flash frozen as beads in liquid nitrogen, and stored at -80°C.

Enzymatic screening assays

Screening was performed using 20 µg of protein in a 200 µl volume at 30°C using 96-well plates. All proteins were assayed twice. The assays were carried out as described (Holloway *et al.*, 1998; Kuznetsova *et al.*, 2005); they are summarized below.

The dehalogenase activity screen employed the following organohalogen substrates: 1,2-dichloroethane, 1,3-dichloropropane, chlorocyclohexane, 1,2-dibromoethane, fluoroacetate, chloroacetate, *S*-2-chloropropionate and bromoacetate. The reaction mixture was buffered at pH 7.0 using 1 mM HEPES. Detection was based on a pH decrease when hydrohalic acid is produced from the halogenated substrates (Holloway *et al.*, 1998), and the colour transition of the pH indicator phenol red (from pink to orange to yellow) was monitored visually.

Two separate assays were developed in our laboratories to test for general esterase activity (Kuznetsova *et al.*, 2005): one uses *p*-nitrophenyl palmitate (pNP-palmitate) to identify carboxylesterases, lipases and certain thioesterases through the release of *p*-nitrophenolate (pNP) detected by absorbance at 405 nm; the other uses palmitoyl-coenzyme A to identify thioester-specific esterases and certain carboxylesterases through the reaction between the released coenzyme A and Ellman's reagent monitored by absorbance at 412 nm.

The general phosphatase assay also developed in our laboratories uses *p*-nitrophenyl phosphate (pNPP) in the presence of divalent cations (either 5 mM Mg²⁺ alone or a cocktail containing 5 mM Mg²⁺ and 0.5 mM each of Co²⁺, Mn²⁺ and Ni²⁺) (Kuznetsova *et al.*, 2005). The hydrolytic activity is detected through the release of pNP by measuring the absorbance at 405 nm.

Dehalogenase activity was determined by visual inspection. All other activities were quantified spectrophotometrically 3 and 24 h after initiating the reaction using the plate reader SpectraMax M5 (Molecular Devices, Sunnyvale, CA). The absorbance values were converted into ratios by dividing them with the absorbance value of the control without enzyme.

Quantitative dehalogenase assays

Dehalogenase activity was confirmed and quantified using alternative techniques for a number of novel dehalogenases. The reaction was performed in 100 mM of different buffers as listed in Table 4. Turnover at 30°C was detected through either halide release (by spectrophotometry (Iwasaki *et al.*, 1952) or ion chromatography coupled with conductivity) or reaction enthalpy (i.e. calorimetry) (Todd and Gomez, 2001). The steady-state kinetic parameters were extracted by non-linear regression using the software package GraphPad Prism (GraphPad Software Inc., La Jolla, CA). The procedures are described in detail in Supporting Information.

Sequence alignments

The sequences within an activity group were aligned using the CLUSTAL algorithm (Thompson *et al.*, 1994) in the Biology Workbench at the San Diego Supercomputer Center (<http://workbench.sdsc.edu/>) (Subramaniam, 1998), with the default parameters. Manual adjustments were made for sequences that did not align properly due to insertions or deletions of large stretches of sequence. Comparisons between different sequence groups within the same protein families were made by using the profile alignment function in the Biology Workbench.

E-value analyses

For each category of BLASTP search (e.g. the search for haloalkane dehalogenases), the retrieved sequences were grouped according to their measured activities (e.g. dehalogenase, esterase or no activity) and their corresponding *E*-values were plotted. The *E*-values from the first and second screen were analysed separately. For target genes possessing several *E*-values (because they could be retrieved by

multiple search queries), all *E*-values were plotted individually without averaging. Targets with low protein yield that displayed no activity were excluded from this analysis because it was not known whether the inactivity was genuine, or caused by insufficient amounts of protein assayed.

Acknowledgements

We gratefully acknowledge support from the National Science and Engineering Research Council – Strategic Grant Program for this project. We thank the following scientists for kindly providing us with the respective bacterial strains or genomic DNA: Thomas Beatty (University of British Columbia) for *Rhodopseudomonas palustris* CGA009; Turlough Finan (McMaster University) for *Sinorhizobium meliloti* 1021; Stephanie Curtis (North Carolina State University) for *Anabaena* sp. PCC7120; Robert Sanford (University of Illinois) and Frank Löffler (Georgia Institute of Technology) for *Anaeromyxobacter dehalogenans* 2CP-C; Christina Kennedy (University of Arizona) for *Azotobacter vinelandii* AvOP; James Tiedje (Michigan State University) for *Burkholderia cenocepacia* HI2424 and *Burkholderia vietnamiensis* G4; Kathleen O'Connor (Purdue University) for *Chromohalobacter salexigens* DSM3043; John Coates (University of California) for *Dechloromonas aromatica* RCB; Mary Ann Moran (University of Georgia) for *Jannaschia* sp. CCS1; James Gossett (Cornell University) for *Polaromonas* sp. JS666; Alejandro Rooney (United States Department of Agriculture) for *Pseudomonas fluorescens* Pf-5; Roberta Fulthorpe (University of Toronto) for *Ralstonia eutropha* JMP134; Christian Boucher (Institut National de la Recherche Agronomique) for *Ralstonia solanacearum* GMI1000; Lindsay Eltis and William Mohn (University of British Columbia) for *Rhodococcus* sp. RHA1 and David Hopwood (Norwich Research Park) for *Streptomyces coelicolor* A3(2). We also acknowledge the Structural Genomics Consortium (Toronto, ON, Canada), Greg Brown, Yuriy Korniyenko, Ekaterina Kuznetsova and Mike Proudfoot for their technical assistance.

References

- Allen, K.N., and Dunaway-Mariano, D. (2004) Phosphoryl group transfer: evolution of a catalytic scaffold. *Trends Biochem Sci* **29**: 495–503.
- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol* **215**: 403–410.
- Barth, P.T., Bolton, L., and Thomson, J.C. (1992) Cloning and partial sequencing of an operon encoding two *Pseudomonas putida* haloalkanoate dehalogenases of opposite stereospecificity. *J Bacteriol* **174**: 2612–2619.
- Benning, M.M., Taylor, K.L., Liu, R.Q., Yang, G., Xiang, H., Wesenberg, G., et al. (1996) Structure of 4-chlorobenzoyl coenzyme A dehalogenase determined to 1.8 Å resolution: an enzyme catalyst generated via adaptive mutation. *Biochemistry* **35**: 8103–8109.
- Brokamp, A., Happe, B., and Schmidt, F.R. (1996) Cloning and nucleotide sequence of a D,L-haloalkanoic acid dehalogenase encoding gene from *Alcaligenes xylosoxidans* ssp. *denitrificans* ABIV. *Biodegradation* **7**: 383–396.
- Burroughs, A.M., Allen, K.N., Dunaway-Mariano, D., and Aravind, L. (2006) Evolutionary genomics of the HAD superfamily: understanding the structural adaptations and catalytic diversity in a superfamily of phosphoesterases and allied enzymes. *J Mol Biol* **361**: 1003–1034.
- Cairns, S.S., Cornish, A., and Cooper, R.A. (1996) Cloning, sequencing and expression in *Escherichia coli* of two *Rhizobium* sp. genes encoding haloalkanoate dehalogenases of opposite stereospecificity. *Eur J Biochem* **235**: 744–749.
- Chae, J.C., Kim, Y., Min, K.H., Kim, Y.C., and Kim, C.K. (1999) Cloning and sequencing of the *fcxB* gene encoding 4-chlorobenzoate-coenzyme A dehalogenase from *Pseudomonas* sp. DJ-12. *Mol Cells* **9**: 225–229.
- Damborsky, J., Bohac, M., Prokop, M., Kutý, M., and Koca, J. (1998) Computational site-directed mutagenesis of haloalkane dehalogenase in position 172. *Protein Eng* **11**: 901–907.
- Fetzner, S., and Lingens, F. (1994) Bacterial dehalogenases: biochemistry, genetics, and biotechnological applications. *Microbiol Rev* **58**: 641–685.
- Hasan, A.K.M.Q., Takada, H., Esaki, N., and Soda, K. (1991) Catalytic action of L-2-halo acid dehalogenase on long-chain L-2-haloalkanoic acids in organic solvents. *Biotechnol Bioeng* **38**: 1114–1117.
- Heikinheimo, P., Goldman, A., Jeffries, C., and Ollis, D.L. (1999) Of barn owls and bankers: a lush variety of alpha/beta hydrolases. *Structure* **7**: R141–R146.
- Hill, K.E., Marchesi, J.R., and Weightman, A.J. (1999) Investigation of two evolutionarily unrelated halocarboxylic acid dehalogenase gene families. *J Bacteriol* **181**: 2535–2547.
- Holloway, P., Trevors, J.T., and Lee, H. (1998) A colorimetric assay for detecting haloalkane dehalogenase activity. *J Microbiol Methods* **32**: 31–36.
- Holmquist, M. (2000) Alpha/beta-hydrolase fold enzymes: structures, functions and mechanisms. *Curr Protein Pept Sci* **1**: 209–235.
- Iwasaki, I., Utsumi, S., and Ozawa, T. (1952) New colorimetric determination of chloride using mercuric thiocyanate and ferric ion. *Bull Chem Soc Jpn* **25**: 226.
- Jesenska, A., Bartos, M., Czernekova, V., Rychlik, I., Pavlik, I., and Damborsky, J. (2002) Cloning and expression of the haloalkane dehalogenase gene *dhmA* from *Mycobacterium avium* N85 and preliminary characterization of *DhmA*. *Appl Environ Microbiol* **68**: 3724–3730.
- Jesenska, A., Pavlova, M., Strouhal, M., Chaloupkova, R., Tesinska, I., Monincova, M., et al. (2005) Cloning, biochemical properties, and distribution of mycobacterial haloalkane dehalogenases. *Appl Environ Microbiol* **71**: 6736–6745.
- Jones, D.H., Barth, P.T., Byrom, D., and Thomas, C.M. (1992) Nucleotide sequence of the structural gene encoding a 2-haloalkanoic acid dehalogenase of *Pseudomonas putida* strain AJ1 and purification of the encoded protein. *J Gen Microbiol* **138**: 675–683.
- de Jong, R.M., and Dijkstra, B.W. (2003) Structure and mechanism of bacterial dehalogenases: different ways to cleave a carbon-halogen bond. *Curr Opin Struct Biol* **13**: 722–730.
- de Jong, R.M., Brugman, W., Poelarends, G.J., Whitman,

- C.P., and Dijkstra, B.W. (2004) The X-ray structure of trans-3-chloroacrylic acid dehalogenase reveals a novel hydration mechanism in the tautomerase superfamily. *J Biol Chem* **279**: 11546–11552.
- de Jong, R.M., Bazzacco, P., Poelarends, G.J., Johnson, W.H., Jr, Kim, Y.J., Burks, E.A., *et al.* (2007) Crystal structures of native and inactivated cis-3-chloroacrylic acid dehalogenase. Structural basis for substrate specificity and inactivation by (R)-oxirane-2-carboxylate. *J Biol Chem* **282**: 2440–2449.
- Kawasaki, H., Tsuda, K., Matsushita, I., and Tonomura, K. (1992) Lack of homology between two haloacetate dehalogenase genes encoded on a plasmid from *Moraxella* sp. strain B. *J Gen Microbiol* **138**: 1317–1323.
- Kawasaki, H., Toyama, T., Maeda, T., Nishino, H., and Tonomura, K. (1994) Cloning and sequence analysis of a plasmid-encoded 2-haloacid dehalogenase gene from *Pseudomonas putida* No. 109. *Biosci Biotechnol Biochem* **58**: 160–163.
- Kennes, C., Pries, F., Krooshof, G.H., Bokma, E., Kingma, J., and Janssen, D.B. (1995) Replacement of tryptophan residues in haloalkane dehalogenase reduces halide binding and catalytic activity. *Eur J Biochem* **228**: 403–407.
- Keuning, S., Janssen, D.B., and Witholt, B. (1985) Purification and characterization of hydrolytic haloalkane dehalogenase from *Xanthobacter autotrophicus* GJ10. *J Bacteriol* **163**: 635–639.
- Kmunicek, J., Hynkova, K., Jedlicka, T., Nagata, Y., Negri, A., Gago, F., *et al.* (2005) Quantitative analysis of substrate specificity of haloalkane dehalogenase LinB from *Sphingomonas paucimobilis* UT26. *Biochemistry* **44**: 3390–3401.
- Koonin, E.V., and Tatusov, R.L. (1994) Computer analysis of bacterial haloacid dehalogenases defines a large superfamily of hydrolases with diverse specificity. Application of an iterative approach to database search. *J Mol Biol* **244**: 125–132.
- Kurihara, T., Liu, J.Q., Nardi-Dei, V., Koshikawa, H., Esaki, N., and Soda, K. (1995) Comprehensive site-directed mutagenesis of L-2-halo acid dehalogenase to probe catalytic amino acid residues. *J Biochem (Tokyo)* **117**: 1317–1322.
- Kurihara, T., Yamauchi, T., Ichiyama, S., Takahata, H., and Esaki, N. (2003) Purification, characterization, and gene cloning of a novel fluoroacetate dehalogenase from *Burkholderia* sp. FA1. *J Mol Catal B Enzym* **23**: 347–355.
- Kuznetsova, E., Proudfoot, M., Sanders, S.A., Reinking, J., Savchenko, A., Arrowsmith, C.H., *et al.* (2005) Enzyme genomics: application of general enzymatic screens to discover new enzymes. *FEMS Microbiol Rev* **29**: 263–279.
- Kuznetsova, E., Proudfoot, M., Gonzalez, C.F., Brown, G., Omelchenko, M.V., Borozan, I., *et al.* (2006) Genome-wide analysis of substrate specificities of the *Escherichia coli* haloacid dehalogenase-like phosphatase family. *J Biol Chem* **281**: 36149–36161.
- Liu, J.Q., Kurihara, T., Hasan, A.K., Nardi-Dei, V., Koshikawa, H., Esaki, N., and Soda, K. (1994) Purification and characterization of thermostable and nonthermostable 2-haloacid dehalogenases with different stereospecificities from *Pseudomonas* sp. strain YL. *Appl Environ Microbiol* **60**: 2389–2393.
- Liu, J.Q., Kurihara, T., Ichiyama, S., Miyagi, M., Tsunasawa, S., Kawasaki, H., *et al.* (1998) Reaction mechanism of fluoroacetate dehalogenase from *Moraxella* sp. B. *J Biol Chem* **273**: 30897–30902.
- Murdiyatmo, U., Asmara, W., Tsang, J.S., Baines, A.J., Bull, A.T., and Hardman, D.J. (1992) Molecular biology of the 2-haloacid halohydrase IVa from *Pseudomonas cepacia* MBA4. *Biochem J* **284**: 87–93.
- Nardi-Dei, V., Kurihara, T., Park, C., Esaki, N., and Soda, K. (1997) Bacterial D,L-2-haloacid dehalogenase from *Pseudomonas* sp. strain 113: gene cloning and structural comparison with D- and L-2-haloacid dehalogenases. *J Bacteriol* **179**: 4232–4238.
- Nardini, M., and Dijkstra, B.W. (1999) Alpha/beta hydrolase fold enzymes: the family keeps growing. *Curr Opin Struct Biol* **9**: 732–737.
- Newman, J., Peat, T.S., Richard, R., Kan, L., Swanson, P.E., Affholter, J.A., *et al.* (1999) Haloalkane dehalogenases: structure of a *Rhodococcus* enzyme. *Biochemistry* **38**: 16105–16114.
- van der Ploeg, J., van Hall, G., and Janssen, D.B. (1991) Characterization of the haloacid dehalogenase from *Xanthobacter autotrophicus* GJ10 and sequencing of the dhIB gene. *J Bacteriol* **173**: 7925–7933.
- Pries, F., Wijngaard, A.J., Bos, R., Pentenga, M., and Janssen, D.B. (1994) The role of spontaneous cap domain mutations in haloalkane dehalogenase specificity and evolution. *J Biol Chem* **269**: 17490–17494.
- Prudnikova, T., Mozga, T., Rezacova, P., Chaloupkova, R., Sato, Y., Nagata, Y., *et al.* (2009) Crystallization and preliminary X-ray analysis of a novel haloalkane dehalogenase DbeA from *Bradyrhizobium elikani* USDA94. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **65**: 353–356.
- Rye, C.A., Isupov, M.N., Lebedev, A.A., and Littlechild, J.A. (2009) Biochemical and structural studies of a L-haloacid dehalogenase from the thermophilic archaeon *Sulfolobus tokodaii*. *Extremophiles* **13**: 179–190.
- Sato, Y., Monincova, M., Chaloupkova, R., Prokop, Z., Ohtsubo, Y., Minamisawa, K., *et al.* (2005) Two rhizobial strains, *Mesorhizobium loti* MAFF303099 and *Bradyrhizobium japonicum* USDA110, encode haloalkane dehalogenases with novel structures and substrate specificities. *Appl Environ Microbiol* **71**: 4372–4379.
- Schmidberger, J.W., Wilce, J.A., Tsang, J.S., and Wilce, M.C. (2007) Crystal structures of the substrate free-enzyme, and reaction intermediate of the HAD superfamily member, haloacid dehalogenase DehIVa from *Burkholderia cepacia* MBA4. *J Mol Biol* **368**: 706–717.
- Schmitz, A., Gartemann, K.H., Fiedler, J., Grund, E., and Eichenlaub, R. (1992) Cloning and sequence analysis of genes for dehalogenation of 4-chlorobenzoate from *Arthrobacter* sp. strain SU. *Appl Environ Microbiol* **58**: 4068–4071.
- Schneider, B., Muller, R., Frank, R., and Lingens, F. (1991) Complete nucleotide sequences and comparison of the structural genes of two 2-haloalkanoic acid dehalogenases from *Pseudomonas* sp. strain CBS3. *J Bacteriol* **173**: 1530–1535.
- Stringfellow, J.M., Cairns, S.S., Cornish, A., and Cooper, R.A. (1997) Haloalkanoate dehalogenase II (DehE) of a *Rhizobium* sp. – molecular analysis of the gene and formation of carbon monoxide from trihaloacetate by the enzyme. *Eur J Biochem* **250**: 789–793.

- Subramaniam, S. (1998) The Biology Workbench – a seamless database and analysis environment for the biologist. *Proteins* **32**: 1–2.
- Swanson, P.E. (1999) Dehalogenases applied to industrial-scale biocatalysis. *Curr Opin Biotechnol* **10**: 365–369.
- Thompson, J.D., Higgins, D.G., and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673–4680.
- Todd, M.J., and Gomez, J. (2001) Enzyme kinetics determined using calorimetry: a general assay for enzyme activity? *Anal Biochem* **296**: 179–187.
- Verschueren, K.H., Seljee, F., Rozeboom, H.J., Kalk, K.H., and Dijkstra, B.W. (1993) Crystallographic analysis of the catalytic mechanism of haloalkane dehalogenase. *Nature* **363**: 693–698.
- Walter, M.C., Rattei, T., Arnold, R., Guldener, U., Munsterkotter, M., Nenova, K., et al. (2009) PEDANT covers all complete RefSeq genomes. *Nucleic Acids Res* **37**: D408–D411.
- Zhang, R.G., Skarina, T., Katz, J.E., Beasley, S., Khachatryan, A., Vyas, S., et al. (2001) Structure of *Thermotoga maritima* stationary phase survival protein SurE: a novel acid phosphatase. *Structure* **9**: 1095–1106.

Supporting information

Additional Supporting Information may be found in the online version of this article:

Fig. S1. Representative SDS-PAGE analysis of *P. aeruginosa* protein targets.

The target proteins in well-expressed and soluble samples appear as large round patches on the gel (e.g. lanes C8, C10,

A8 and C6). By inspection, the purity of these high-yield samples is generally above 75%. For samples that are poorly expressed or insoluble, a large number of expression host protein bands are detected. In such cases (e.g. lanes E1, E2, E5 and E7), the protein purity is approximately 35% at best. The identifiable target protein bands in low-yield samples have been circled. MW denotes molecular weight of the standards in kDa.

Fig. S2. Representative data in carboxylesterase and thioesterase screens.

The upper and lower panels display representative data from the carboxylesterase and thioesterase screens respectively. Absorbance was measured 3 h (light bars) and 24 h (dark bars) after initiation of the reaction. Data from the 24-hour time point for the carboxylesterase screen was not interpretable due to high backgrounds. The dashed lines mark the cut-off used for confirming activities.

Table S1. All ABH superfamily targets selected for biochemical characterization.

Table S2. All HAD superfamily targets selected for biochemical characterization.

Table S3. Cloning and purification results.

Table S4. Summary of cloning and purification results.

Table S5. Sequence identity and similarity of all predicted haloalkane dehalogenases.

Table S6. Sequence identity and similarity of all predicted fluoroacetate dehalogenases.

Table S7. Sequence identity and similarity of all predicted L-2-haloacid dehalogenases.

Please note: Wiley-Blackwell are not responsible for the content or functionality of any supporting materials supplied by the authors. Any queries (other than missing material) should be directed to the corresponding author for the article.