Site-Selective C-H Halogenation using Flavin-Dependent Halogenases Identified via Family-Wide Activity Profiling

Brian F. Fisher,[1] Harrison M. Snodgrass,[1] Krysten A. Jones,[2] Mary C. Andorfer,[3] Jared C. Lewis*,[1]

Synopsis

High-throughput screening of >20,000 reactions catalyzed by 87 soluble genome-mined halogenases on 62 substrates found 39 new active halogenases for selective late-stage C−H functionalization.

Abstract

Enzymes are powerful catalysts for site-selective C-H bond functionalization. Identifying suitable enzymes for this task and for biocatalysis in general remains challenging, however, due to the fundamental difficulty of predicting catalytic activity from sequence information. In this study, family-wide activity profiling was used to obtain sequence-function information on flavin-dependent halogenases (FDHs). This broad survey provided a number of insights into FDH activity, including halide specificity and substrate preference, that were not apparent from the more focused studies reported to date. Regions of FDH sequence space that are most likely to contain enzymes suitable for halogenating small molecule substrates were also identified. FDHs with novel substrate scope and complementary regioselectivity on large, three-dimensionally complex compounds were characterized and used for preparative-scale late-stage C-H functionalization. In many cases, these enzymes provide activities that required several rounds of directed evolution to accomplish in previous efforts, highlighting that this approach can achieve significant time savings for biocatalyst identification and provide advanced starting points for further evolution.

Introduction

Enzymes can be powerful tools for the synthesis of fine chemicals, pharmaceuticals,[1] agrochemicals, and many other materials.[2] On the other hand, the very features that give rise to the selectivity and catalytic proficiency of enzymes acting on their native substrates often lead to high substrate specificity and thus poor activity on non-native substrates. The dearth of enzymes available for reactions of interest can therefore be a major impediment to implementing enzymes in synthetic routes. Many of the enzymes commonly used today (e.g. ketoreductases, transaminases, cytochromes P450, etc.) were originally identified via arduous enzymology aimed at clarifying their native biological activities. Often, many rounds of directed evolution were also required to optimize these enzymes for synthetic applications. Expanding the number of known biocatalysts, with an emphasis on exploring broad sequence diversity within an enzyme family, could therefore greatly facilitate the use of enzymes in chemical synthesis.[3, 4]

Numerous methods have been used to explore the functional diversity of naturally occurring enzymes in discrete genomes, metagenomic samples, and sequence databases.[5] Advances in DNA sequencing—in particular, metagenome sequencing— have resulted in an explosion of the size of protein sequence databases.[6] Coupled with the decreasing cost of gene synthesis,[7] mining these sequence databases for potential biocatalysts is becoming increasingly accessible to scientists.[8] Such approaches are most commonly used to identify enzymes that act on a substrate of interest, often a chromogenic probe

compound chosen more for ease of screening than synthetic utility,[9, 10] but efforts to profile the activity of entire enzyme families on a range of substrates and identify biocatalysts with a collectively-broad substrate scope are less common. Family-wide profiling efforts include investigations on phosphatases,[11] metallo-β-lactamases,[9, 12] and glutathione-S-transferases.[13] Studies on dehalogenases,[14] esterases,[15] glycosyl transferases,[16] and imine reductases[17] highlight the potential synthetic utility of enzymes identified from such efforts. Comparable genome mining efforts on enzymes that functionalize C-H bonds have not been reported.

Flavin-dependent halogenases (FDHs), which catalyze site-selective C-H halogenation of electron-rich aromatic compounds, have been studied extensively due to their potential synthetic utility.[18, 19] Late-stage functionalization,[20] sequential halogenation/cross-coupling,[21-23] and preparative-scale halogenation[24] have all been accomplished using these enzymes. Our efforts have focused primarily on RebH, an FDH that was identified in studies aimed at elucidating the biosynthetic pathway of the antitumor compound rebeccamycin.[25] In this context, RebH catalyzes site-selective chlorination of tryptophan, and it has since been shown to halogenate a range of indoles and anilines.[26] Our group has also shown that directed evolution can be used to create RebH variants with improved thermal stability,[27] high activity on large, biologically active compounds,[20] and high selectivity for different sites on target compounds.[28] While effective, these efforts required 3-8 rounds of directed evolution due to the wild-type enzyme's modest stability, low activity on large substrates, and high regioselectivity.

While additional FDHs could therefore expand the utility of these enzymes for synthesis, only a relatively narrow set of FDHs have been investigated for biocatalysis.[19] FDHs that catalyze tryptophan chlorination, such as RebH, Thal,[29] SttH,[30] and PrnA,[31] in particular are over-represented. Fungal halogenases, such as Rdc2,[32] RadH,[33] and GsfI,[34] which natively chlorinate phenol-containing substrates, have also been shown to be active and selective biocatalysts. Literature reports on the collective substrate scopes of the FDHs reported to date suggest that they prefer chloride over other halides and that they act on electron-rich compounds similar to their native substrate.[19] On the other hand, the existence of a range of complex halogenated natural products distinct from those produced by well-characterized biosynthetic gene clusters[35] implies that FDHs with unique substrate scopes might be found in less characterized halogenase subgroups.[36] We hypothesized that exploring uncharacterized FDHs found in protein sequence databases could, together with currently characterized enzymes, form a diverse starting toolkit for selective late-stage C-H halogenation.

Herein, we describe the use of a high-throughput mass spectrometry-based screen to evaluate a broad set of over one hundred putative FDH sequences drawn from throughout the FDH family. Halogenases with novel substrate scope and complementary regioselectivity on large, three-dimensionally complex compounds were identified. This effort involved far more extensive sequence-function analysis than has been accomplished using the relatively narrow range of FDHs characterized to date, providing a clearer picture of the regions in FDH sequence space that are most likely to contain enzymes suitable for halogenating small molecule substrates. The representative enzyme panel constructed in this study also provides a rapid means to identify FDHs for lead diversification via late-stage C-H functionalization. In many cases, these enzymes provide activities that required several rounds of directed evolution to accomplish in previous efforts, highlighting that this approach can achieve significant time savings for biocatalyst identification and provide advanced starting points for further evolution.

Results

Organization of Halogenase Sequence Similarity Network

A BLAST search of the UniProt sequence database using RebH as a query sequence and an *E*-value threshold of 10^{-5} generated 3,975 unique hits spanning a range of sequence and host diversity, including bacterial, archaeal, eukaryotic, and viral proteins. Nearly all (>90%) previously reported FDHs are present in this set. The dinucleotide-binding GxGxxG motif, characteristic of FAD-binding proteins, is present in 92% of the sequences, and the WxWxIP motif,[37] characteristic of FDHs but absent in flavin-dependent monooxygenases, is found in 69% of the sequences. The latter value increases to 78% when motif variants WxWxI[R,G][38] are included. Collectively, these analyses suggest that the majority of the sequences examined are likely FDHs.

Sequence similarity networks (SSNs)[39] were then used to visualize functional relationships among putative FDH sequences. In this representation, protein sequences are illustrated as nodes in a network graph that are connected by edges (lines) to other sequences that exceed a specified pairwise sequence similarity. An SSN was generated for the entire FDH sequence set with a permissive edge detection threshold (corresponding to ≈30% sequence identity) using the Enzyme Function Institute's Enzyme Similarity Tool (EFI-EST).[40, 41] Previously reported data for 129 known enzymes found among the BLAST hits were mapped onto this Level 1 SSN to explore subnetwork co-localization of enzyme properties. The clearest defining features of the individual subnetworks are host domain and compound class—indole, phenol, or pyrrole—of native substrates for known FDHs within the subnetworks (Figure 1A), the latter suggesting that the SSN might provide a framework for identifying enzymes that act on specific compound classes and for surveying regions of sequence space where substrate preference is unknown.

The largest subnetwork, comprising 2,270 sequences, contains FDHs that either natively halogenate tryptophan or have been shown to catalyze indole halogenation *in vitro*. All known tryptophan FDHs are found in this Indole Subnetwork, including tryptophan 5-, 6-, and 7-halogenases PyrH,[42] SttH,[43] and RebH.[44] BrvH, a halogenase identified from metagenomic analysis,[45] and three recently reported halogenases from *Xanthomonas campestris*[46] are also in this subnetwork. Although the native substrates of these enzymes are not known, they have been shown to halogenate a variety of small indoles. A protein whose structure has been determined as part of structural genomics efforts (PDB: 2PYX)[47] is also present in this subnetwork, although its native activity is also unknown.

The second largest subnetwork comprises 438 sequences from bacteria and fungi and includes most known phenol FDHs that, collectively, halogenate a diverse range of phenol-containing substrates. For example, the bacterial halogenase TiaM chlorinates a large macrocyclic intermediate in the biosynthesis of tiacumicin B.[48] Bacterial halogenases VhaA and Tcp21 chlorinate PCP-tethered amino acids in the biosynthesis of the NRPS glycopeptide antibiotics vancomycin and teicoplanin, respectively.[49] The bacterial iodinase CalO3 is also present in the Phenol Subnetwork,[50] showcasing that substrate diversity also extends to halide specificity. All fungal phenol FDHs that have been studied as biocatalysts on diverse substrates, including Rdc2,[32] RadH,[33] and GsfI,[34] are also contained in this subnetwork.

The third largest subnetwork, with 212 sequences, contains FDHs that are involved in chlorinated pyrrole natural product biosynthesis. The six pyrrole halogenases in the Pyrrole Subnetwork have an

average pairwise identity of 87%, and all are annotated in UniProt as PrnC, which halogenates a pyrrole small molecule intermediate in pyrrolnitrin biosynthesis.[51] The halogenase PltM natively halogenates a phenolic substrate, phloroglucinol, to produce chlorinated compounds that induce biosynthesis of a pyrrole-containing natural product, pyoluteorin.[52] Two proteins, Dox16 and Dox17, potentially halogenate phenolic moieties during the biosynthesis of pyrrolomycins,[53] pyrrole-containing compounds structurally similar to pyrrolnitrin. These observations suggest that the common ancestor to enzymes in this subnetwork diverged in substrate specificity to yield halogenases specialized for distinct roles in chlorinated pyrrole natural product biosynthesis (Figure S18).

Most of the smaller subnetworks contain only uncharacterized proteins (and are therefore simply assigned a number for reference in Figs. 1 or 2), but several include known FDHs with diverse native substrates. One small subnetwork contains several enzymes, including MibH,[54] MscL,[55] and KrmI,[56] that natively chlorinate peptidyl tryptophan sidechains in macrocyclic lanthipeptide and NRPS natural products. Other subnetworks include MalA and MalA', which are responsible for iterative chlorination in the biosynthesis of malbrancheamide,[38] ChlA, which chlorinates a phenol in DIF-1 biosynthesis,[57] and GetL, an enzyme suspected to be responsible for chlorinating PCP-tethered histidine in the biosynthesis of tetrapeptide antibiotics[58]. Halogenases responsible for chlorinating ACP-tethered pyrroles (Variant B pyrrole halogenases) such as PltA[59, 60] and Mpy16[61] occupied a subnetwork distinct from the larger subnetwork that included the Variant A pyrrole halogenase PrnC. A few subnetworks contain enzymes that are not FDHs, including the flavin-dependent monooxygenases Qhpg[62] and LodB[63] and putative geranylgeranyl reductases[64].

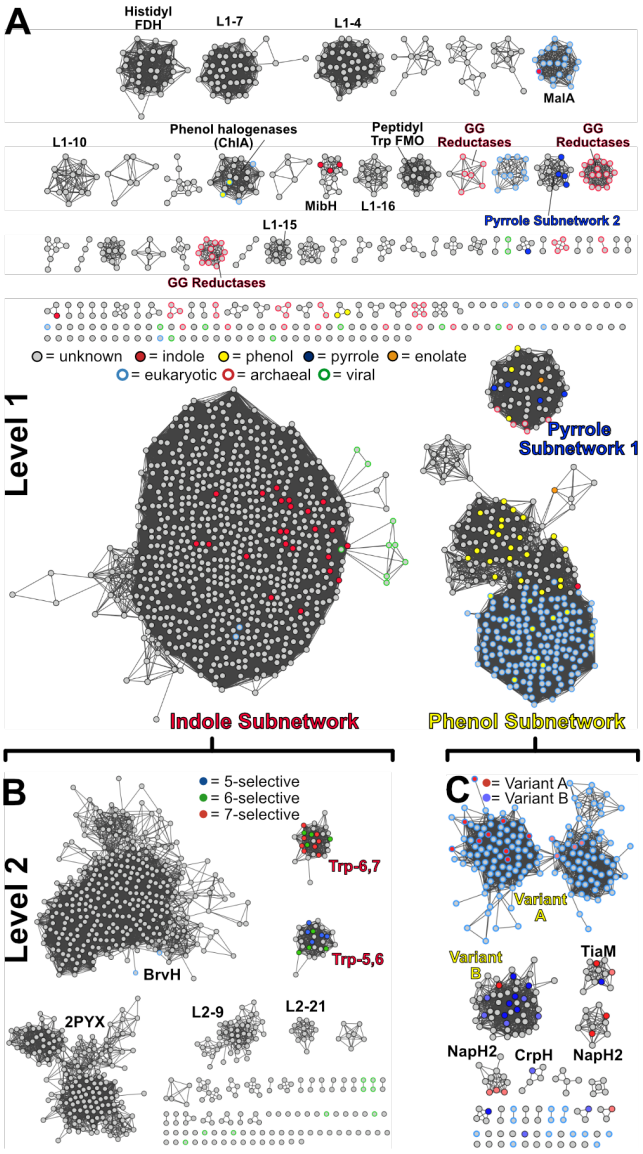

Figure 1. A) Sequence similarity network for flavin-dependent halogenases. Each circle is a representative node, grouping protein sequences with >50% sequence identity as determined by CD-HIT.[65] Edge detection threshold set at alignment score of 70 (≈30% sequence identity). Nodes are filled according to native substrate functional group of at least one sequence in the representative node; colored stroke indicates domain (thin black stroke = bacterial). Subnetworks with ≥15 sequences but without any known sequence are labelled numerically. B and C) Level 2 subnetworks formed from the Indole (B) and Phenol (C) subnetwork using a stricter alignment score cutoff of 140 (≈40% sequence identity). Level 2 subnetworks are labelled based on known sequences in the subnetwork. For Indole Subnetwork sequences, nodes containing known tryptophan halogenases are filled according to their regioselectivity, and subnetworks with ≥15 sequences are labeled numerically. For Phenol Subnetwork sequences, nodes are filled according to the halogenase variant type (A = free small molecule native substrate, B = ACP-tethered native substrate).

Subnetworks in SSNs can be explored in greater detail by increasing the stringency of the sequence similarity required for edge detection.[39] The SSN drawn with ≈30% identity cutoff for edge detection (Level 1), was examined with the identity cutoff increased to ≈40% (Level 2). Functionally distinct

subnetworks within the Level 1 Indole Subnetwork became evident in the Level 2 SSN. All known tryptophan halogenases localized to only two relatively small Level 2 subnetworks distinguished by their regioselectivity (Figure 1B). All tryptophan 5-halogenases, such as PyrH, localized into one of these, and all tryptophan 7-halogenases, including RebH and PrnA, were found in the other subnetwork. Interestingly, tryptophan 6-halogenases were found roughly evenly distributed between these two subnetworks. Only two reports describe the substrate scopes of FDHs within the largest Level 2 subnetwork in the Indole Subnetwork, which demonstrated that some enzymes in this subnetwork prefer bromination to chlorination.[45, 46] The second largest subnetwork contained the sequence of the structurally characterized but functionally uncharacterized protein 2PYX. Overall, the sparse evaluation of enzymes within the Indole Subnetwork highlights the fact that, even among proteins that are most similar to the well-characterized tryptophan halogenases, there remains a vast amount of sequence space to be explored.

Closer inspection of the Level 1 Phenol Subnetwork at the stricter Level 2 identity cutoff shows subnetworks separated on the basis of domain and whether the FDH natively halogenates a free small molecule (variant A) or an acyl carrier protein-tethered small molecule (variant B) (Figure 1C). The largest subnetwork is composed entirely of eukaryotic sequences, and all experimentally characterized proteins within the subnetwork, such as Rdc2, are variant A halogenases. The second-largest subnetwork in this group contains only bacterial sequences, many of which, including VhaA,[66] are variant B halogenases that catalyze chlorination in glycopeptide antibiotic biosynthesis.[49]

Expression of Genome-Mined Halogenases

The sequence similarity network outlined above was used as a framework to guide the selection of a diverse set of novel FDHs from each subnetwork. The Phenol Subnetwork was oversampled due to the high structural diversity of substrates natively halogenated by known enzymes in this subnetwork. Other sequences were sampled evenly from the rest of the SSN. Transcriptomic data for sequences from eukaryotes were analyzed using the JGI Mycocosm database to prioritize the synthesis of sequences in order of sequence model quality (see SI). Figure 2 depicts the SSN and treemap representations summarizing the distribution of different properties of enzymes in the different subnetworks.

A total of 128 putative halogenase sequences and RebH as a control were codon-optimized and co-expressed with chaperones from the plasmid pGro7 in *E. coli* BL21(DE3) under conditions found to be successful in expression of bacterial as well as fungal halogenases.[34] A total of 87 new enzymes were obtained in sufficient soluble concentration for functional characterization, but attempts to improve parallel expression of the remaining enzymes did not lead to significant improvements (Fig. S53). Halogenases from throughout the entire SSN could be expressed with good titers, but solubility was not evenly distributed (Figure 2D). While 68% of enzymes were soluble, the Indole Subnetwork provided a significantly higher fraction of soluble enzymes compared to others (91%, 42 total). The halogenases in the Phenol Subnetwork had much lower solubility (49% overall, 17 total), which was not significantly influenced by the domain of the source organism (45% soluble for eukaryotic, and 50% soluble for bacterial genes). An average number of Pyrrole Subnetwork halogenases were soluble (71%, 5 total), while several small subnetworks that were sampled provided no soluble halogenases under the expression conditions tested.

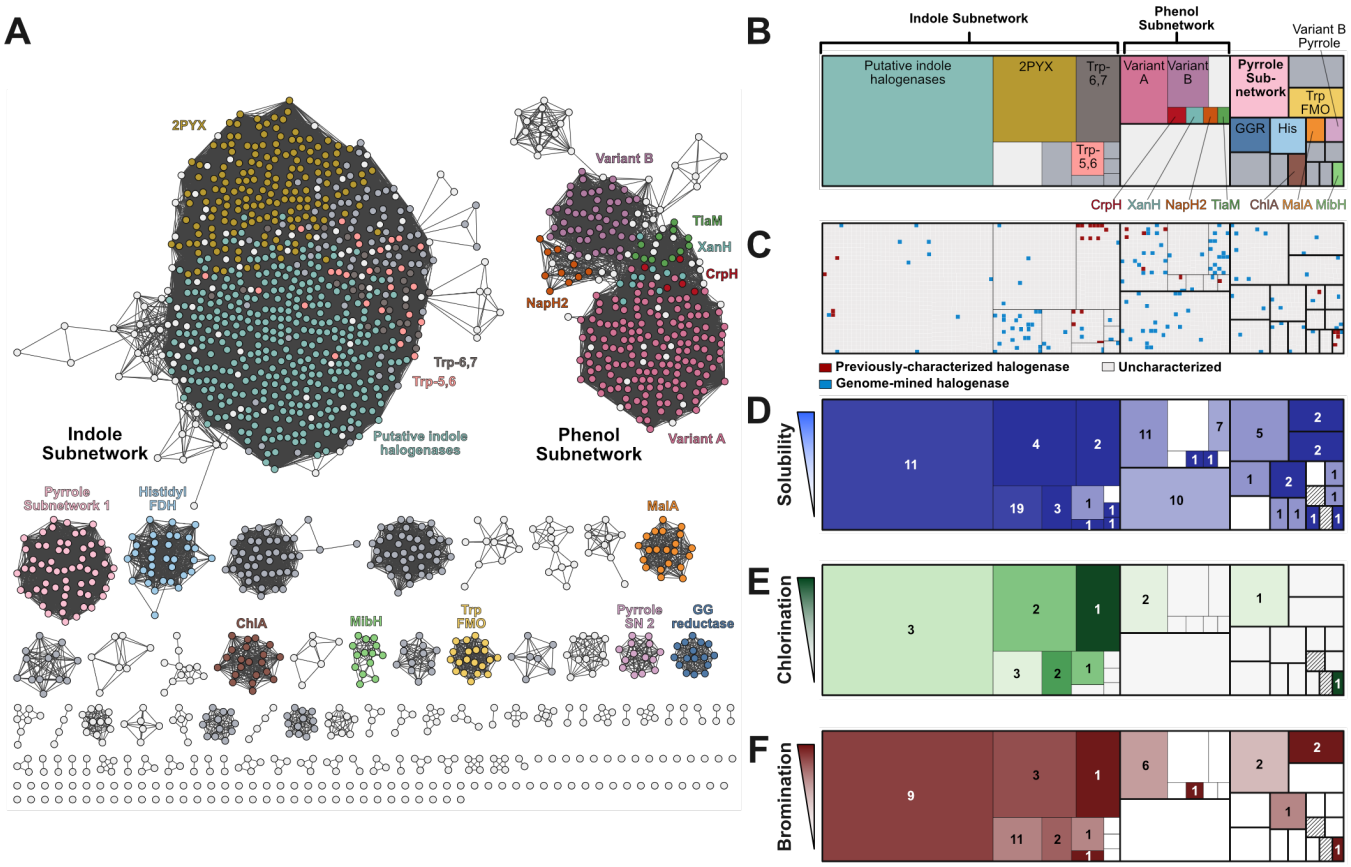

Figure 2. A) Sequence-similarity network for flavin-dependent halogenases, drawn at the less stringent edge detection threshold (≈30% identity), colored according to subnetwork. Subnetworks within the Indole and Phenol subnetworks at the more stringent threshold are colored differently. Subnetworks with fewer than 15 members are colored white; subnetworks without sequences of known or inferred function are colored light gray. B) Treemap illustrating the SSN with the same coloring as A. C) Treemap comparing FDHs previously studied as biocatalysts with FDHs investigated in this study. D) Treemap illustrating solubility of genome-mined enzymes in each subnetwork of the SSN. Color gradient represents fraction of enzymes within the subnetwork that was soluble; diagonal bars indicate subnetworks wherein no enzyme was tested. E and F) Treemaps illustrating the fraction of enzymes in each subnetwork that were capable of chlorinating (E) or brominating (F) at least one substrate in the high-throughput screen (8% conversion threshold).

Probe Substrate High-Throughput Screen

The set of 87 diverse, soluble FDHs was subjected to a high-throughput activity screen to evaluate which enzymes had detectable activity and, for active enzymes, to develop substrate activity profiles to better understand whether activity and subnetwork membership were related. For initial activity screens, a set of 12 probe substrates—4 indoles, 4 anilines, and 4 phenols—was selected from among the substrates previously found by our group to be reactive under FDH chlorination conditions (Figure 3A). The key hypothesis governing selection of these substrates was that their high inherent reactivity, reflected in their high calculated halenium affinity values,[34, 67] would lead to detectable reactivity with active enzymes even if they exhibited poor binding within FDH active sites. Structural variation within the panel was used to facilitate the identification of viable substrates,[68] and substrates with multiple potentially reactive sites

were prioritized to increase the probability that reactive binding poses could be achieved. Initial screens evaluated both chlorination and bromination activities, the two most common halogenation reactions catalyzed by FDHs. The probe substrate screens required at least 2,040 independent experiments, not including replicates or controls. This heavy screening requirement prompted us to adopt a high-throughput LC-MS-based screen (Figure 3B), which also required that viable substrates ionize well by ESI.[69-71] Using this method, analysis throughput of up to ≈11 seconds per reaction was achieved, and ultimately ≈20,000 experiments were analyzed.

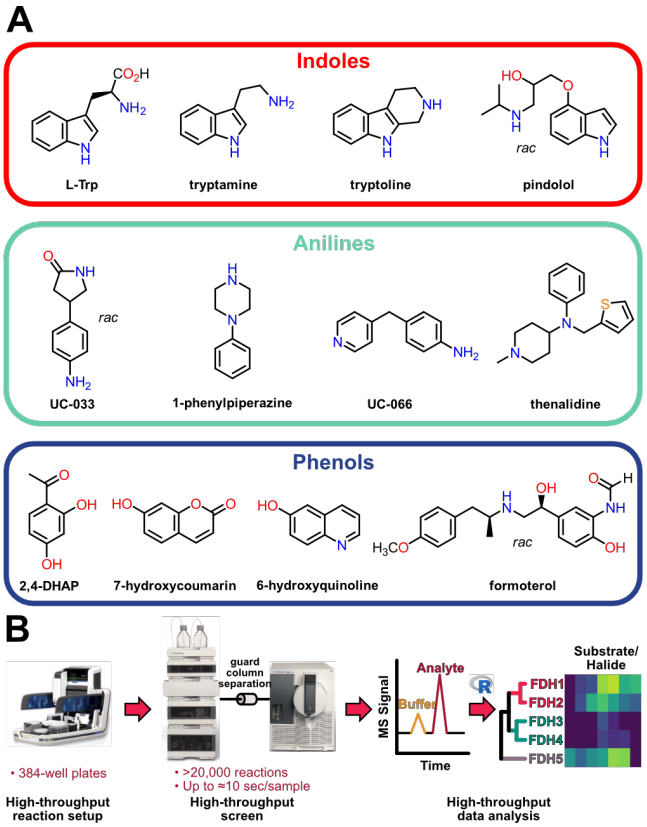

Figure 3. A) Probe substrates included in initial high-throughput screen. B) Scheme summarizing LC-MS-based high-throughput screening method employed.

A total of 39 new halogenases (45% of soluble enzymes) were able to halogenate at least one of the probe substrates. Halogenation of nearly the entire probe substrate panel was achieved by the genome-mined set of enzymes; only formoterol was not halogenated by at least one new halogenase. Overall, bromination activity was more prevalent than chlorination activity. All genome-mined enzymes that were active had brominase activity, but only 16% of the enzyme set had detectable chlorinase activity. Activity was unevenly distributed across the SSN; certain SSNs had a higher abundance of active enzymes than others (Figure 2E-F). The Indole Subnetwork had a particularly high percentage of active enzymes; of the 42 Indole Subnetwork enzymes screened, 27 (64%) were active. The fraction of active enzymes was similar for bacterial and eukaryotic proteins, with 48% of bacterial and 56% of eukaryotic enzymes screened having some activity on probe substrates. One of the three viral proteins tested was active, and none of the six archaeal proteins were active.

The high-throughput screening conversion data for each reaction were plotted as a heatmap, and hierarchical clustering analysis was used to characterize, separately, the similarity of activity profiles for substrates and for FDHs (Figure 4). Substrates tended to form clusters based on their compound class, consistent with the observed similarity of "enzyme-scope" of substrates within the same substrate class.[34] All phenols were present in two substrate clusters, one containing only chlorination reactions and the other containing only bromination reactions. Anilines and indoles were more mixed into the remaining two clusters, but still distinguishable. One of these clusters primarily included indole chlorination, dominated by the high indole chlorination activity of RebH and a highly similar enzyme, 1-B12. The other contained mostly aniline bromination reactions, high activity for which was more broadly distributed.

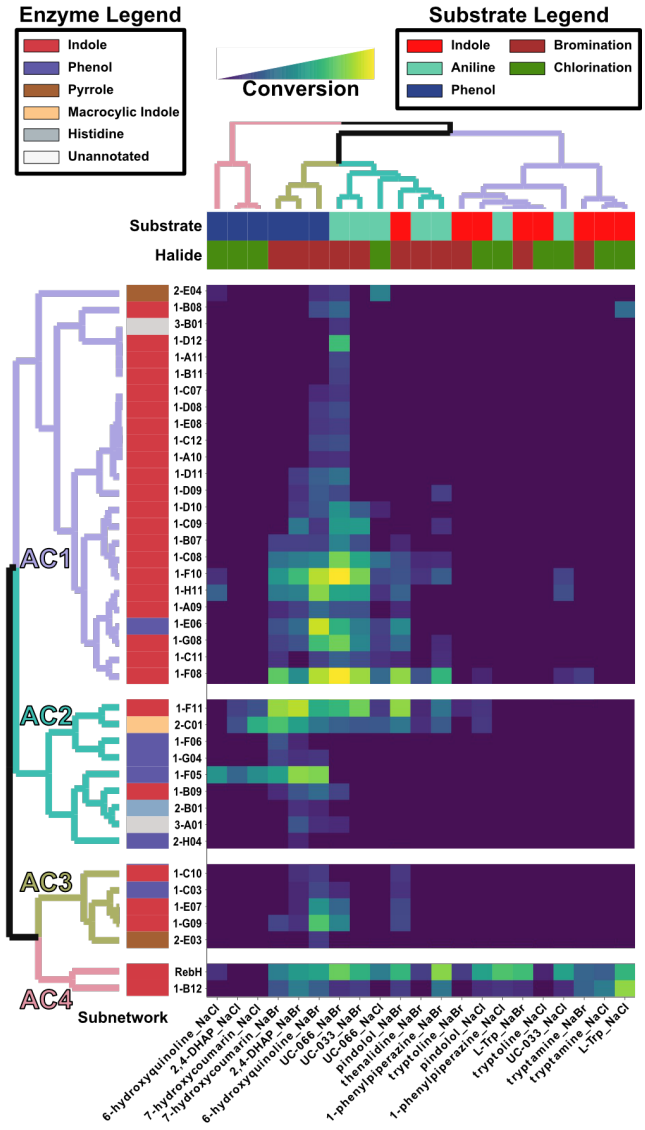

Figure 4. Heatmap of high-throughput screening results, with hierarchical clustering dendrograms for substrate/halide activity similarity and enzyme activity similarity at top and left, respectively. Substrate functional groups and halide used in the reaction are color-coded with bars at the tips of the dendrograms. Only reactions with >8% conversion, a value selected that removed false positives (see SI).

Most importantly, enzymes in the same subnetwork tended to cluster together based on their activity profiles. Four activity clusters of enzymes (AC1-4) can be distinguished from the probe substrate high-throughput screening data. AC1, at the top of the heatmap, contained almost exclusively halogenases in the Indole Subnetwork. None of the Indole Subnetwork enzymes in this AC were in either of the two Level 2 tryptophan subnetworks, however, and they were distinguished by their preference for bromination of phenols and anilines. Despite the fact that indoles are the most common substrates known to be halogenated by enzymes in the Indole Subnetwork, halogenase activity on indoles in AC1 was limited. Pindolol was the only indole halogenated by more than one enzyme, and only a single FDH, 1-F08 (34% identical to SttH), chlorinated more than one indole.

Activity cluster 2 (AC2) had similar bromination scope to AC1, but had higher breadth of phenol chlorination activity. Only two of the nine enzymes in this activity cluster were present in the Indole Subnetwork, whereas four were in the Phenol Subnetwork. Halogenase 1-F11, from an unannotated subnetwork within the Indole Subnetwork (38% identical to tryptophan-5 halogenase ClaH[72]), and 2-C01, a halogenase in the same subnetwork as the lanthipeptide indole halogenase MibH (36% identical[54]), were capable of chlorinating multiple phenols, 2,4-dihydroxyacetophenone (2,4-DHAP) and 7-hydroxycoumarin. The FDH 2-C01 was particularly versatile in halide scope. UC-066, 7-hydroxycoumarin, and 2,4-DHAP were chlorinated and brominated by 2-C01 with similar yields. Enzyme 1-F05, from the Phenol Subnetwork (49% identical to ArmH4[73]), was similarly versatile in the halides it accepted, but its activity was specific for phenolic probe substrates. It had the broadest phenol substrate scope of any enzyme tested, but did not halogenate any aniline or indole.

Activity cluster 3 (AC3) was small and populated by low-activity enzymes only having bromination activity on the substrates that were most easily halogenated. AC4 contained only two enzymes, RebH and 1-B12, that had the broadest substrate scope, particularly on indole probe substrates. The high probe substrate scope of RebH was expected by design, since the indoles and anilines of the probe panel was assembled from substrates that were known to be chlorinated by RebH. The enzyme 1-B12 has high sequence similarity to RebH (64% identical) and a strongly similar substrate activity profile.

Activity and Selectivity of Mined Halogenases Toward Complex Substrates

Based on the remarkable activity with that our genome-mined halogenases exhibited toward probe substrates, we next wondered whether they might be capable of halogenating substrates that were not selected from a set of easily halogenated compounds. A total of 50 larger and more three-dimensionally complex additional substrates were selected for these activity studies (Figure 5A). Among the compounds in this expanded substrate set were yohimbine, a compound for which we previously evolved halogenase activity from RebH,[20] and premalbrancheamide, a compound natively halogenated by the FDH MalA.[38] Most of the substrates have not been reported as FDH substrates previously, including β-estradiol 17-(β-D-glucuronide), an estrogen metabolite, and cabergoline, an ergot alkaloid. A total of 48% of the more complex substrates tested were halogenated by at least one halogenase under the non-optimized conditions used in the high-throughput screen (Figure 5B). Hierarchical clustering was performed on the reaction data as above. However, the similarities between enzymes were substantially lower than in the clustering analysis of the probe substrate data, and activity clusters were consequently less defined.

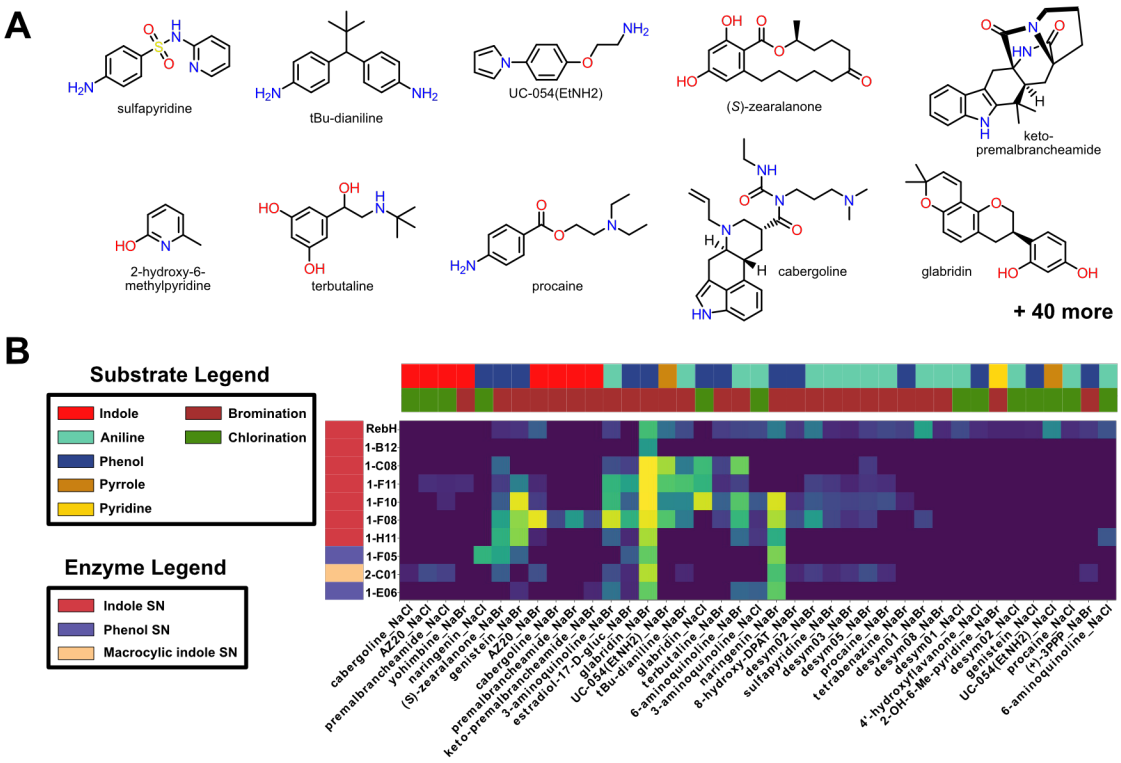

Figure 5. A) Representative compounds included in expanded high-throughput substrate screen, each of which was halogenated by at least one genome-mined FDH. B) Heatmap of expanded substrate screen data with ten of the most active enzymes from the probe high-throughput substrate screen.

Larger quantities of several of the most active genome-mined FDHs were expressed, purified, and used for preparative-scale bioconversions on a subset of the larger substrates evaluated (Figure 6). Premalbrancheamide is a compound natively dichlorinated by MalA at C5 and C6, and which has been shown to be halogenate at the C5 or C6 positions non-selectively using either chloride or bromide as halide sources.[38] The Indole Subnetwork FDH 1-F08 preferentially brominates premalbrancheamide at C5 in 51% isolated yield, and it also brominates AZ20, a selective ATR kinase inhibitor, in 28% isolated yield at the indole C3 position. A different Indole Subnetwork enzyme, 1-F11, was capable of brominating β-estradiol 17-(β-D-glucuronide), an estradiol metabolite, at the C4 position of the phenol in 57% yield and the 1-position of the carvedilol carbazole ring in 56% isolated yield.

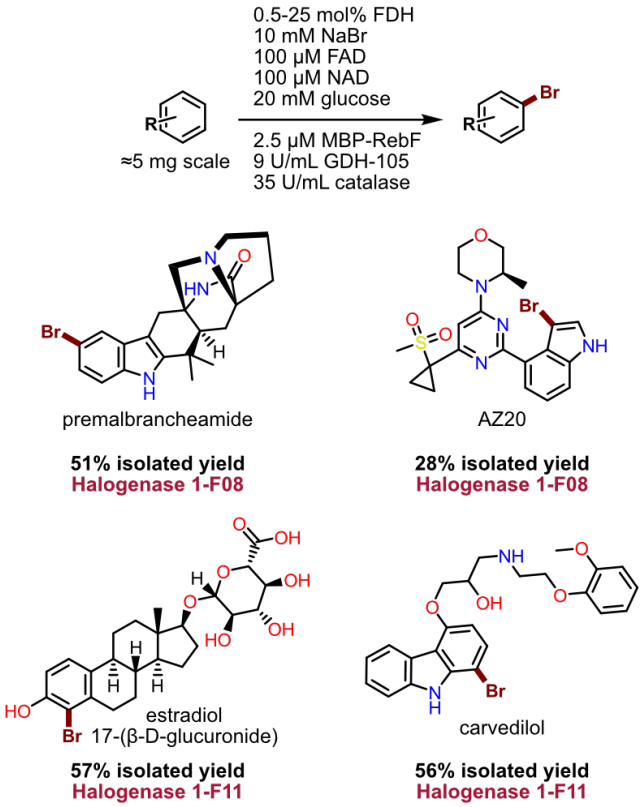

Figure 6. Preparative-scale bioconversions of larger substrates.

Many examples of reactions in which different regioisomers were formed by different enzymes were also identified (Figure 7). Pindolol, which is brominated at C7 by RebH variants,[21] is also brominated at C7 by the Indole Subnetwork FDH 1-F11. The MibH subnetwork enzyme 2-C01, on the other hand, preferentially brominates at C2. This finding is notable since C2 is less electronically activated than C7 based on its 3 kcal/mol lower halenium affinity (HalA), a metric for computationally evaluating the reactivity of different positions of a molecule toward EAS.[67] Naringenin is brominated at two different positions using 1-F11 or Phenol Subnetwork enzyme 1-F05. Despite the negligible energetic differences in HalA for C6 and C8 (0.7 kcal/mol), 1-F05 was found to be >95% selective for C8, while 1-F11 and other FDHs were found to have only minor preferences in regioselectivity for C8 or C6. Trp-6,7 halogenase subnetwork enzyme 1-B12 halogenates the indole-containing compound methylergonovine at C7, which has a halenium affinity 4.2 kcal/mol lower than C2, the most nucleophilic aromatic C-H site on this compound. The FDH 2-C01, on the other hand, brominates methylergonovine at C2.

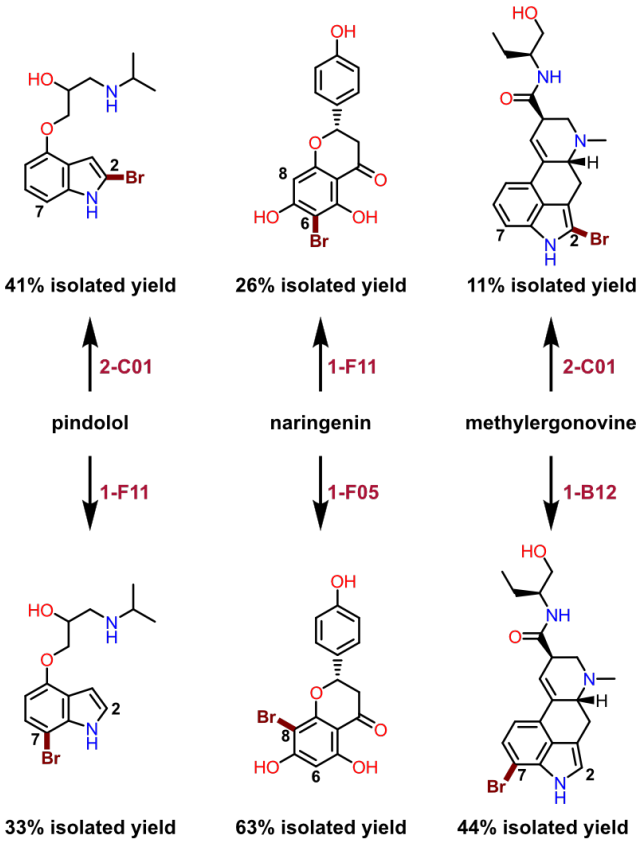

Figure 7. Regiocomplementary halogenation of large molecules.

Discussion

A Family-wide View of FDH Properties

Family-wide analysis of FDHs revealed several notable trends that are not apparent from prior studies. First, FDHs from diverse host organisms can be solubly expressed without significant optimization of expression conditions. Bacterial enzymes had the highest soluble expression success rate (76%), while a lower fraction (40%) of eukaryotic enzymes were soluble. Notably, however, the lower fraction of soluble eukaryotic FDHs reflects the poorer solubility of halogenases in the Phenol Subnetwork regardless of host organism domain. Nearly all (20/23) of the eukaryotic proteins evaluated were within the Phenol Subnetwork. Within this subnetwork, the soluble expression rate is generally low, but it is actually higher for proteins from eukaryotes (54%) relative to bacteria (44%). This finding indicates that eukaryotic FDHs can be readily expressed in *E. coli* and that genome mining efforts should be encouraged to include enzymes from eukaryotic species.[74]

Second, halogenase activity was also evenly distributed between enzymes from bacterial and eukaryotic organisms (48% and 56% active, respectively). FDH activity was not observed for any archaeal proteins evaluated, consistent with the strong possibility that most if not all archaeal sequences in the SSN are geranylgeranyl reductases. Interestingly, one viral FDH, a cyanophage auxiliary metabolic gene product,[75] was active, though its activity and substrate scope were low (conversion of <35% on only three probe substrates was observed). In general, the identification of such a high percentage of active halogenases, despite the use of non-native substrates for activity profiling and a lax homology requirement

for evaluation (*E*-value threshold of 10^{-5}), suggests that this family contains a large number of enzymes suitable for biocatalysis.

Third, bromination activity was much more widespread than chlorination activity within the FDHs surveyed. The majority of the FDH biocatalysis literature focuses on chlorination activity because most FDHs reported to date are involved in the biosynthesis of chlorinated natural products. Moore[76] has reported three flavin dependent brominases involved in the biosynthesis of brominated natural products, but these are more distantly related to enzymes comprising the SSN in the current study. These brominases have 17 ± 4% sequence identity to enzymes in the SSN; for comparison, RebH exhibits 29 ± 10% sequence identity to our genome-mined enzymes. Sewald[45, 46] reported flavin dependent halogenases (contained in the Indole Subnetwork of the FDH SSN) that prefer bromide over chloride when acting on the (presumably) non-native substrate indole. While this observation was taken to indicate specificity of these enzymes toward bromide, our findings indicate that a preference for bromination is common in FDHs. We suggest that the higher electrophilicity of bromine relative to chlorine in heteroatom-X species,[77] such as the proposed hypohalous acid or haloamine halogenating agents in FDH catalysis, leads to more facile bromination. For example, the native chlorinase RebH can brominate a greater range of non-native substrates than it can chlorinate. Preference for bromination over chlorination for non-native as well as native substrates is also observed when both Cl^- and Br^- are present in solution. In competition reactions including both NaCl and NaBr, RebH prefers bromide over chloride for L-tryptophan, 1-phenylpiperazine, pindolol, and 2,4-dihydroxyacetophenone halogenation.[64] It is therefore possible that enzymes with higher bromination than chlorination scope in our high-throughput screen could nevertheless natively catalyze chlorination reactions.

Analyzing FDH Activity using Sequence Similarity Networks and Activity Clustering

Sequence similarity networks provide an intuitive structure for exploring the protein sequence space of enzyme families. The FDH SSN contains Level 1 subnetworks comprising enzymes with similar native substrate preferences (indole vs. phenol, etc.). At a more stringent identity threshold cutoff, Level 2 subnetworks with finer functional distinction are revealed. Within the Level 1 Phenol Subnetwork, for example, different Level 2 subnetworks containing primarily either variant A or variant B halogenases, which natively halogenate free small molecules or PCP-tethered substrates respectively, are observed. The ability to distinguish such enzyme subclasses based on sequence alone is useful for focusing future genome mining efforts since our data indicate that neither of the variant B phenol halogenases examined were even soluble. Information on site selectivity could also be obtained directly from sequence information in some cases. For example, within the Level 1 Indole Subnetwork, separation of tryptophan 5- and 7-halogenases into distinct Level 2 subnetworks was apparent, though tryptophan 6-halogenases were roughly evenly distributed between these subgroups.

Only six of the Level 1 subnetworks (Fig. 1A) examined contained enzymes with measurable chlorination or bromination activity on our probe substrate set, but these subnetworks contained 78% of the FDHs within the SSN. Specifically, enzymes in the Indole Subnetwork (66%), the Phenol Subnetwork (42%), the Pyrrole Subnetwork (2/5), subnetwork 4 (2/2), subnetwork 8 (1/2), and the MibH subnetwork (1/1) were active. These findings reflect the nature of the probe substrates chosen, but given the range of substrates examined and the similarity of these substrates to pharmaceuticals and other fine chemicals, they also highlight regions of FDH sequence space most likely to be of interest for biocatalysis.

Active enzymes were found in most Level 2 subnetworks that comprise the Level 1 Indole Subnetwork (Fig. 1B). For example, all enzymes in both Tryptophan halogenase subnetworks were active, as was the only enzyme in subnetwork 21. Most enzymes in the BrvH halogenase subnetwork (84%), three out of four enzymes in subnetwork 2PYX, and two of three enzymes (1-C08 and 1-F11) in subnetwork 9 were active. The activity results within the Indole Subnetwork broadly show that a high fraction of these enzymes have potential as useful biocatalysts and highlight several underexplored regions in the FDH sequence space that merit further investigation.

Analysis of Level 2 subnetworks within the Level 1 Phenol Subnetwork also highlights regions with high potential for biocatalyst identification. The majority of the tested enzymes in the variant A subnetwork, including 1-F05, were active (66%), but both of the variant B halogenases were insoluble under the conditions examined. The only soluble genome-mined enzyme in the large phenol halogenase subnetwork containing XanH was active. The single evaluated enzyme in the NapH2 subnetwork was inactive, as were the six soluble enzymes that were either singletons or within small (<15 members) subnetworks. Overall, the variant A subnetwork within the Phenol Subnetwork shows clear promise as a source of novel biocatalysts, but further study of other subnetworks would be required to get a clearer picture of their potential.

Finally, functional characterization of enzymes across the FDH SSN demonstrated that enzymes within a Level 1 subnetwork have similar activity profiles on smaller probe substrates, but that this trend diminishes on more complex substrates. Not surprisingly, more closely-related enzymes possess more similar activity profiles. Highly similar substrate activity profiles are observed for RebH and 1-B12 (64% identical), both of which are within the Level 2 tryptophan 6,7-halogenase subnetwork, as well as for halogenases 1-H11 and 1-F10 (50% identical), both of which are in the Level 2 BrvH subnetwork. These trends suggest an approximate %ID threshold for future genome-mining of new halogenases with similar substrate scopes. Because the gene selection process of this study intentionally favored diverse sequences to maximize the breadth of our search for new halogenases, however, there are few instances of such similar enzyme pairs in which both were soluble and highly active. The average %ID for the most closely-related enzyme within the genome-mined set was 41.2 ± 12.4%, perhaps too low for similarities in activity profiles among enzymes to result in consistent trends. More thorough genome mining of subnetworks with highly active FDHs could yield more concrete activity profiles and reveal more detailed information regarding enzyme substrate preferences.

Unique Activity and Selectivity of Mined Halogenases

The selectivity of FDHs on their native substrates has driven interest in these enzymes as biocatalysts. The potential of FDHs has been explored by researchers seeking to extend their synthetic utility toward gram-scale synthesis,[24] more facile cross-coupling chemistry,[21-23]synthesis of enantioenriched products,[78, 79] and diversification of natural product biosynthetic pathways.[80-82] Other work has sought to make operation more economical, including efforts toward improved cofactor regeneration.[24, 83] Because a given FDH may not provide the selectivity required for a particular application, however, a number of labs have explored the use of targeted mutations to alter FDH selectivity. While grafting key residues from one tryptophan halogenase into another has been used to switch selectivity on tryptophan,[29] modest selectivity has generally been reported for efforts focused on non-native substrates (e.g. converting SttH from 90% 6-selective to 75% 5-selective for 3-indolepropionic acid

chlorination).[30, 84] To address this issue, our lab established that directed evolution can be used to generate FDHs with high (>90%) regioselectivity for different sites on a single substrate (tryptamine), and that the resulting enzymes also had altered selectivity on a range of other substrates.[28] Several rounds of evolution were required to achieve this goal, so accelerating the identification of FDHs with complementary regioselectivity on non-native substrates remains an important goal.

Gratifyingly, a number of enzymes identified in our family-wide survey of FDH activity exhibited regiocomplementarity on a number of structurally complex substrates. For example, the enzyme 2-C01 often provided different regiochemical outcomes than other halogenases. This FDH is present in a subnetwork along with MibH, which natively chlorinates a tryptophan indole ring in a large lanthipeptide.[54] MibH has a large, hydrophobic binding pocket in order to accommodate its native substrate. The genome-mined halogenase 2-C01 may have a similar active site, which could accommodate large substrates in distinct binding poses. This finding suggests that 2-C01 could be a promising starting point for evolving FDHs to achieve late-stage functionalization of aryl C–H bonds with distinct regioselectivity relative to other FDHs characterized to date.

Comparison of the Genome-Mined Halogenase Library with Evolved Variants

Enzymes frequently require substantial modification before they are capable of being deployed as useful catalysts for organic synthesis. As the regiocomplementarity noted above illustrates, access to a diverse pool of enzymes that can serve as starting points for directed evolution can greatly expedite biocatalyst identification. While evolving a single enzyme can take a great deal of effort and may ultimately fail to provide the desired levels of improvement, a related enzyme may be better suited initially to the task and can drastically reduce the effort required to obtain a desired biocatalyst. This point can be retrospectively illustrated by several enzymes in our genome-mined set, which perform comparably to evolved RebH variants with increased thermal stability,[27] expanded substrate scope,[20] and altered regioselectivity[28].

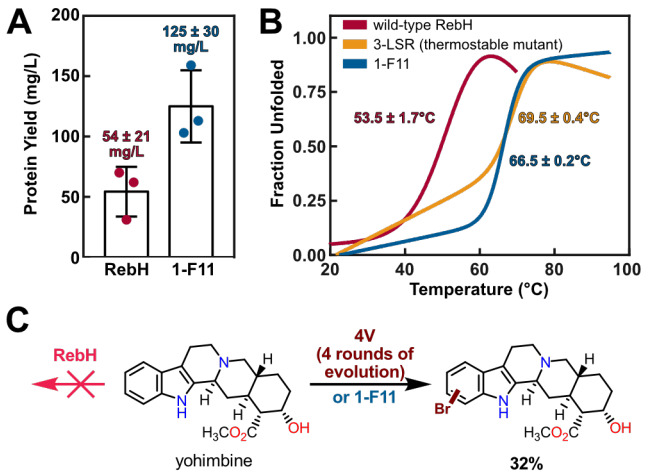

Figure 8. A) Comparison of isolated RebH and 1-F11 protein yields after Ni-NTA purification from 50 mL expression cultures. B) Comparison of CD thermal melts of RebH, thermostable RebH variant 3-LSR, and genome-mined halogenase 1-F11. Curves shown are best fit for thermal unfolding monitored at 222 nm using CDPal.[85] C) Wild-type RebH required several rounds of directed evolution before

yohimbine halogenation was detectable. Halogenase 1-F11 can halogenate yohimbine without directed evolution (HPLC conversion shown).

Halogenase 1-F11 is notable in this regard. FDH biocatalysis is often hampered by low protein expression yields,[56] therefore a more soluble starting enzyme for FDH directed evolution would be especially attractive. The expression yield of 1-F11 was 125 ± 30 mg/L from a 50 mL expression culture, higher than that of RebH, expression of which yielded 54 ± 21 mg/L enzyme under analogous expression conditions (Figure 8A). Higher halogenase activity in lysate on numerous substrates is also observed for 1-F11 compared with RebH.[64] Despite originating from a mesophilic sphingomonas species within an *A. thaliana* root microbiome, 1-F11 has comparable thermal stability (T_m = 66.5 °C) to RebH variant 3-LSR (T_m = 69.5 °C, Figure 8B), which was evolved over three rounds of directed evolution for improved thermal stability[27]. Since more stable enzymes can have a longer catalytic lifetime and can better tolerate random mutations,[86] 1-F11 provides a convenient starting point for directed evolution. 1-F11 also compares favorably in substrate scope with the RebH mutant 4V, which was evolved over four rounds of directed evolution for the late-stage C−H functionalization of yohimbine, a complex, biologically active molecule.[20] Because RebH has minimal activity on yohimbine, an substrate-walking directed evolution approach was required to evolve an enzyme that could halogenate this compound. Enzyme 1-F11, on the other hand, was capable of brominating yohimbine without any modification through directed evolution (Figure 8C). In short, 1-F11 possesses capabilities that took a total of seven rounds of directed evolution using two different approaches to accomplish, highlighting the benefits of family-wide genome mining for biocatalyst identification. Moreover, given the broad substrate scope of 1-F11, we envision it could be an ideal starting point for further directed evolution.

Conclusions

FDHs were first characterized in the mid-1990s. Since this time, most of the FDHs reported have come from either studies on individual biosynthetic pathways or genome mining efforts targeting specific organisms or metagenomic samples.[19] Figures 2C and 2E/F illustrate how these efforts have focused on a remarkably narrow range of FDH sequence space and missed out on large swaths of this space that contain functional enzymes, respectively. Family-wide activity analysis shows that similar fractions of FDHs from bacteria and fungi are soluble and active and that bromination is more commonly observed than chlorination. Broader sampling of this space has not only led to the identification of new enzymes with unique catalytic properties, but also highlighted regions of sequence space that are ripe for further exploration. As noted above, other regions might also be suitable for different types of substrates than those examined herein, but for the electron-rich aromatic compounds explored to date, these regions are clearly privileged. Moreover, the SSNs and substrate activity profiles developed in this study offer predictive ability for focusing biocatalyst selection or further genome mining efforts for particular applications. Extending this approach, involving SSN-guided selection of sequences from throughout an enzyme family, label-free high-throughput mass spectrometry screening using synthetic probe substrates, and activity profiling, to other enzymes has great potential for expediting biocatalyst identification.

Beyond these family-wide findings, a number of remarkably useful enzymes were identified in the representative set that was explored. Particularly notable in this regard are 1-F11, 1-F08, 2-C01, 1-F05, and 1-B12. Collectively, these enzymes enable C-H halogenation of previously inaccessible substrates, provide complementary site selectivity on complex biologically active substrates, and exhibit improved

thermostability relative to a commonly reported FDH. Their sequences also differ significantly from other FDHs that have been explored *in vitro*. With the exception of 1-B12, which is 64% identical to RebH, they are only 34-43% identical to FDHs that have been explored as biocatalysts. These novel and diverse halogenases therefore represent promising starting points for both directed evolution and additional genome mining aimed at identifying similarly-effective biocatalysts. The activity of these enzymes on complex natural products and pharmaceuticals also suggests that their native substrates could be similarly fascinating structures. It could therefore be interesting to examine the native function of these enzymes, reversing the enzymology-to-biocatalysis progression that has dominated biocatalyst development to date.[4] This approach could provide a unique means of identifying new halogenated natural products and other unique compounds when extended to other enzyme classes.

Acknowledgements

This study was supported by the NIH (1R01GM115665) and by the NSF under the CCI Center for Selective C-H Functionalization (CCHF, CHE-1700982). B.F.F. was supported by an NIH F32 postdoctoral fellowship (F32GM123693). Halogenase genes were provided by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, supported under Contract No. DE-AC02-05CH11231, and we thank Drs. Sam Deutch, Yasuo Yoshikuni, Igor Grigoriev, Asaf Salamov, Angela Tarver, and Sangeeta Nath for assistance with gene selection, transcriptomic analysis, and synthesis. We thank CCHF members for helpful discussions, Dr. Chris Welch for helpful comments regarding high-throughput screening, and Prof. Richmond Sarpong, Jose Roque, and Dr. David Stephens for providing samples of premalbrancheamide and keto-premalbrancheamide for halogenation studies.

Supporting Information

Full conversion data, full SSN data, ChemDraw files of all evaluated compounds and structures used in halenium affinity calculations, interactive .html heatmaps of conversion data; materials/instruments, sequence data and bioinformatics analysis, protein expression protocol, automated reaction setup protocol, LC-MS-based high-throughput reaction screening procedure and data processing/analysis, preparative-scale reaction procedures and compound data (^{1}H-NMR, ^{13}C-NMR, 2D-NMR, HR-MS), halenium affinity calculations, halide selectivity analysis, circular dichroism data.

Safety Statement

No uncommon safety risks were encountered while conducting the described research.

Present Author Addresses

[1]Department of Chemistry, Indiana University, Bloomington, Indiana 47405, United States
[2]Department of Chemistry, University of Chicago, Chicago, Illinois 60637, United States
[3]Department of Biology, Massachusetts Institute of Technology, Cambridge, Massachusetts 02139, United States

References

1. Devine, P. N.; Howard, R. M.; Kumar, R.; Thompson, M. P.; Truppo, M. D.; Turner, N. J., Extending the Application of Biocatalysis to Meet the Challenges of Drug Development. *Nat. Rev. Chem.* **2018,** *2*, 409.

2. Woodley, J. M., Accelerating the Implementation of Biocatalysis in Industry. *Appl. Microbiol. Biotechnol.* **2019,** *103*, 4733.

3. Truppo, M. D., Biocatalysis in the Pharmaceutical Industry – the Need for Speed. *ACS Med. Chem. Lett.* **2017,** *8*, 476.

4. Bornscheuer, U. T.; Huisman, G. W.; Kazlauskas, R. J.; Lutz, S.; Moore, J. C.; Robins, K., Engineering the Third Wave of Biocatalysis. *Nature* **2012,** *485*, 185.

5. Zaparucha, A.; Berardinis, V. d.; Vaxelaire-Vergne, C., Genome Mining for Enzyme Discovery. In *Modern Biocatalysis: Advances Towards Synthetic Biological Systems*, Williams, G.; Hall, M., Eds. The Royal Society of Chemistry: Cambridge, 2018.

6. Chen, C.; Huang, H.; Wu, C. H., Protein Bioinformatics Databases and Resources. In *Protein Bioinformatics: From Protein Modifications and Networks to Proteomics*, Wu, C. H.; Arighi, C. N.; Ross, K. E., Eds. Springer: New York, NY, 2017; pp 3.

7. Hughes, R. A.; Ellington, A. D., Synthetic DNA Synthesis and Assembly: Putting the Synthetic in Synthetic Biology. *CSH Perspect. Biol.* **2017,** *9*, a023812.

8. Kamble, A.; Srinivasan, S.; Singh, H., In-Silico Bioprospecting: Finding Better Enzymes. *Mol. Biotechnol.* **2019,** *61*, 53.

9. Baier, F.; Tokuriki, N., Connectivity between Catalytic Landscapes of the Metallo-Beta-Lactamase Superfamily. *J. Mol. Biol.* **2014,** *426*, 2442.

10. Colin, P.-Y.; Kintses, B.; Gielen, F.; Miton, C. M.; Fischer, G.; Mohamed, M. F.; Hyvönen, M.; Morgavi, D. P.; Janssen, D. B.; Hollfelder, F., Ultrahigh-Throughput Discovery of Promiscuous Enzymes by Picodroplet Functional Metagenomics. *Nat. Commun.* **2015,** *6*, 10008.

11. Huang, H.; Pandya, C.; Liu, C.; Al-Obaidi, N. F.; Wang, M.; Zheng, L.; Keating, S.; Aono, M.; Love, J. D.; Evans, B.; Seidel, R. D.; Hillerich, B. S.; Garforth, S. J.; Almo, S. C.; Mariano, P. S.; Dunaway-Mariano, D.; Allen, K. N.; Farelli, J. D., Panoramic View of a Superfamily of Phosphatases through Substrate Profiling. *Proc. Natl. Acad. Sci. USA* **2015,** *112*, E1974.

12. Baier, F.; Chen, J.; Solomonson, M.; Strynadka, N.; Tokuriki, N., Distinct Metal Isoforms Underlie Promiscuous Activity Profiles of Metalloenzymes. *ACS Chem. Biol.* **2015,** *10*, 1684.

13. Mashiyama, S. T.; Malabanan, M. M.; Akiva, E.; Bhosle, R.; Branch, M. C.; Hillerich, B.; Jagessar, K.; Kim, J.; Patskovsky, Y.; Seidel, R. D.; Stead, M.; Toro, R.; Vetting, M. W.; Almo, S. C.; Armstrong, R. N.; Babbitt, P. C., Large-Scale Determination of Sequence, Structure, and Function Relationships in Cytosolic Glutathione Transferases across the Biosphere. *PLOS Biol.* **2014,** *12*, e1001843.

14. Chan, W. Y.; Wong, M.; Guthrie, J.; Savchenko, A. V.; Yakunin, A. F.; Pai, E. F.; Edwards, E. A., Sequence‐ and Activity‐Based Screening of Microbial Genomes for Novel Dehalogenases. *Microb. Biotechnol.* **2010,** *3*, 107.

15. Martínez-Martínez, M.; Coscolín, C.; Santiago, G.; Chow, J.; Stogios, P. J.; Bargiela, R.; Gertler, C.; Navarro-Fernández, J.; Bollinger, A.; Thies, S.; Méndez-García, C.; Popovic, A.; Brown, G.; Chernikova, T. N.; García-Moyano, A.; Bjergah, G. E. K.; Pérez-García, P.; Hai, T.; Pozo, M. V.; Stokke, R.; Steen, I. H.; Cui, H.; Xu, X.; Nocek, B. P.; Alcaide, M.; Distaso, M.; Mesa, V.; Pelaez, A. I.; Sánchez, J.; Buchholz, P. C. F.; Pleiss, J.; Fernández-Guerra, A.; Glöckner, F. O.; Golyshina, O. V.; Yakimov, M. M.; Savchenko, A.; Jaeger, K.-E.; Yakunin, A. F.; Streit, W. R.; Golyshin, P. N.;

Guallar, V.; Ferrer, M., Determinants and Prediction of Esterase Substrate Promiscuity Patterns. *ACS Chem. Biol.* **2018,** *13*, 225.

16. Yang, M.; Fehl, C.; Lees, K. V.; Lim, E.-K.; Offen, W. A.; Davies, G. J.; Bowles, D. J.; Davidson, M. G.; Roberts, S. J.; Davis, B. G., Functional and Informatics Analysis Enables Glycosyltransferase Activity Prediction. *Nat. Chem. Biol.* **2018,** *14*, 1109.
17. Wetzl, D.; Berrera, M.; Sandon, N.; Fishlock, D.; Ebeling, M.; Müller, M.; Hanlon, S.; Wirz, B.; Iding, H., Expanding the Imine Reductase Toolbox by Exploring the Bacterial Protein ‐ Sequence Space. *ChemBioChem* **2015,** *16*, 1749.
18. Agarwal, V.; Miles, Z. D.; Winter, J. M.; Eustaquio, A. S.; El Gamal, A. A.; Moore, B. S., Enzymatic Halogenation and Dehalogenation Reactions: Pervasive and Mechanistically Diverse. *Chem. Rev.* **2017,** *117*, 5619.
19. Latham, J.; Brandenburger, E.; Shepherd, S. A.; Menon, B. R. K.; Micklefield, J., Development of Halogenase Enzymes for Use in Synthesis. *Chem. Rev.* **2018,** *118*, 232.
20. Payne, J. T.; Poor, C. B.; Lewis, J. C., Directed Evolution of Rebh for Site-Selective Halogenation of Large Biologically Active Molecules. *Angew. Chem. Int. Ed.* **2015,** *54*, 4226.
21. Durak, L. J.; Payne, J. T.; Lewis, J. C., Late-Stage Diversification of Biologically Active Molecules Via Chemoenzymatic C-H Functionalization. *ACS Catal.* **2016,** *6*, 1451.
22. Latham, J.; Henry, J. M.; Sharif, H. H.; Menon, B. R.; Shepherd, S. A.; Greaney, M. F.; Micklefield, J., Integrated Catalysis Opens New Arylation Pathways Via Regiodivergent Enzymatic C-H Activation. *Nat. Commun.* **2016,** *7*, 11873.
23. Frese, M.; Schnepel, C.; Minges, H.; Voß, H.; Feiner, R.; Sewald, N., Modular Combination of Enzymatic Halogenation of Tryptophan with Suzuki–Miyaura Cross-Coupling Reactions. *ChemCatChem* **2016,** *8*, 1799.
24. Frese, M.; Sewald, N., Enzymatic Halogenation of Tryptophan on a Gram Scale. *Angew. Chem. Int. Ed.* **2015,** *54*, 298.
25. Yeh, E.; Garneau, S.; Walsh, C. T., Robust in Vitro Activity of Rebf and Rebh, a Two-Component Reductase/Halogenase, Generating 7-Chlorotryptophan During Rebeccamycin Biosynthesis. *Proc. Natl. Acad. Sci. USA* **2005,** *102*, 3960.
26. Payne, J. T.; Andorfer, M. C.; Lewis, J. C., Regioselective Arene Halogenation Using the Fad-Dependent Halogenase Rebh. *Angew. Chem. Int. Ed.* **2013,** *52*, 5271.
27. Poor, C. B.; Andorfer, M. C.; Lewis, J. C., Improving the Stability and Catalyst Lifetime of the Halogenase Rebh by Directed Evolution. *ChemBiochem* **2014,** *15*, 1286.
28. Andorfer, M. C.; Park, H. J.; Vergara-Coll, J.; Lewis, J. C., Directed Evolution of Rebh for Catalyst-Controlled Halogenation of Indole C-H Bonds. *Chem. Sci.* **2016,** *7*, 3720.
29. Moritzer, A.-C.; Minges, H.; Prior, T.; Frese, M.; Sewald, N.; Niemann, H. H., Structure-Based Switch of Regioselectivity in the Flavin-Dependent Tryptophan 6-Halogenase Thal. *J. Biol. Chem.* **2019,** *294*, 2529.

30. Shepherd, S. A.; Menon, B. R. K.; Fisk, H.; Struck, A. W.; Levy, C.; Leys, D.; Micklefield, J., A Structure‐Guided Switch in the Regioselectivity of a Tryptophan Halogenase. *ChemBioChem* **2016,** *17*, 821.

31. Lang, A.; Polnick, S.; Nicke, T.; William, P.; Patallo, E. P.; Naismith, J. H.; van Pée, K.-H. H., Changing the Regioselectivity of the Tryptophan 7-Halogenase Prna by Site-Directed Mutagenesis. *Angew. Chem. Int. Ed.* **2011,** *50*, 2951.

32. Zeng, J.; Lytle, A. K.; Gage, D.; Johnson, S. J.; Zhan, J., Specific Chlorination of Isoquinolines by a Fungal Flavin-Dependent Halogenase. *Bioorg. Med. Chem. Lett.* **2013,** *23*, 1001.

33. Menon, B. R. K.; Brandenburger, E.; Sharif, H. H.; Klemstein, U.; Shepherd, S. A.; Greaney, M. F.; Micklefield, J., Radh: A Versatile Halogenase for Integration into Synthetic Pathways. *Angew. Chem. Int. Ed.* **2017,** *56*, 11841.

34. Andorfer, M. C.; Grob, J. E.; Hajdin, C. E.; Chael, J. R.; Siuti, P.; Lilly, J.; Tan, K. L.; Lewis, J. C., Understanding Flavin-Dependent Halogenase Reactivity Via Substrate Activity Profiling. *ACS Catal.* **2017,** *7*, 1897.

35. Gribble, G. W., *Naturally Occurring Organohalogen Compounds - a Comprehensive Update*. Springer: Vienna, 2010.

36. Gribble, G. W., A Recent Survey of Naturally Occurring Organohalogen Compounds. *Environ. Chem.* **2015,** *12*, 396.

37. Podzelinska, K.; Latimer, R.; Bhattacharya, A.; Vining, L. C.; Zechel, D. L.; Jia, Z., Chloramphenicol Biosynthesis: The Structure of Cmls, a Flavin-Dependent Halogenase Showing a Covalent Flavin–Aspartate Bond. *J. Mol. Biol.* **2010,** *397*, 316.

38. Fraley, A. E.; Garcia-Borràs, M.; Tripathi, A.; Khare, D.; Mercado-Marin, E. V.; Tran, H.; Dan, Q.; Webb, G. P.; Watts, K. R.; Crews, P.; Sarpong, R.; Williams, R. M.; Smith, J. L.; Houk, K. N.; Sherman, D. H., Function and Structure of Mala/Mala', Iterative Halogenases for Late-Stage C-H Functionalization of Indole Alkaloids. *J. Am. Chem. Soc.* **2017,** *139*, 12060.

39. Atkinson, H. J.; Morris, J. H.; Ferrin, T. E.; Babbitt, P. C., Using Sequence Similarity Networks for Visualization of Relationships across Diverse Protein Superfamilies. *PLOS ONE* **2009,** *4*, e4345.

40. Gerlt, J. A.; Bouvier, J. T.; Davidson, D. B.; Imker, H. J.; Sadkhin, B.; Slater, D. R.; Whalen, K. L., Enzyme Function Initiative-Enzyme Similarity Tool (Efi-Est): A Web Tool for Generating Protein Sequence Similarity Networks. *Biochim. Biophys. Acta, Proteins Proteomics* **2015,** *1854*, 1019.

41. Zallot, R.; Oberg, N.; Gerlt, J. A., The Efi Web Resource for Genomic Enzymology Tools: Leveraging Protein, Genome, and Metagenome Databases to Discover Novel Enzymes and Metabolic Pathways. *Biochemistry* **2019**, DOI: 10.1021/acs.biochem.9b00735.

42. Zehner, S.; Kotzsch, A.; Bister, B.; Süssmuth, R. D.; Méndez, C.; Salas, J. A.; van Pée, K. H., A Regioselective Tryptophan 5-Halogenase Is Involved in Pyrroindomycin Biosynthesis in Streptomyces Rugosporus Ll-42d005. *Chem. Biol.* **2005,** *12*, 445.

43. Zeng, J.; Zhan, J., Characterization of a Tryptophan 6-Halogenase from Streptomyces Toxytricini. *Biotechnol. Lett.* **2011,** *33*, 1607.

44. Sánchez, C.; Butovich, I. A.; Braña, A. F.; Rohr, J.; Méndez, C.; Salas, J. A., The Biosynthetic Gene Cluster for the Antitumor Rebeccamycin: Characterization and Generation of Indolocarbazole Derivatives. *Chem. Biol.* **2002,** *9*, 519.

45. Neubauer, P. R.; Widmann, C.; Wibberg, D.; Schröder, L.; Frese, M.; Kottke, T.; Kalinowski, J.; Niemann, H. H.; Sewald, N., A Flavin-Dependent Halogenase from Metagenomic Analysis Prefers Bromination over Chlorination. *PLOS ONE* **2018,** *13*, e0196797.

46. Ismail, M.; Frese, M.; Patschkowski, T.; Ortseifen, V.; Niehaus, K.; Sewald, N., Flavin‐Dependent Halogenases from Xanthomonas Campestris Pv. Campestris B100 Prefer Bromination over Chlorination. *Adv. Synth. Catal.* **2019,** *361*, 2475.

47. Pdb Id: 2pyx. Joint Center for Structural Genomics (JCSG). Crystal Structure of Tryptophan Halogenase (Yp_750003.1) from Shewanella Frigidimarina Ncimb 400 at 1.50 a Resolution.

48. Xiao, Y.; Li, S.; Niu, S.; Ma, L.; Zhang, G.; Zhang, H.; Zhang, G.; Ju, J.; Zhang, C., Characterization of Tiacumicin B Biosynthetic Gene Cluster Affording Diversified Tiacumicin Analogues and Revealing a Tailoring Dihalogenase. *J. Am. Chem. Soc.* **2011,** *133*, 1092.

49. Kittila, T.; Kittel, C.; Tailhades, J.; Butz, D.; Schoppet, M.; Buttner, A.; Goode, R. J. A.; Schittenhelm, R. B.; van Pee, K. H.; Sussmuth, R. D.; Wohlleben, W.; Cryle, M. J.; Stegmann, E., Halogenation of Glycopeptide Antibiotics Occurs at the Amino Acid Level During Non-Ribosomal Peptide Synthesis. *Chem. Sci.* **2017,** *8*, 5992.

50. Ahlert, J.; Shepard, E.; Lomovskaya, N.; Zazopoulos, E.; Staffa, A.; Bachmann, B. O.; Huang, K.; Fonstein, L.; Czisny, A.; Whitwam, R. E.; Farnet, C. M.; Thorson, J. S., The Calicheamicin Gene Cluster and Its Iterative Type I Enediyne Pks. *Science* **2002,** *297*, 1173.

51. Kirner, S.; Hammer, P. E.; Hill, D. S.; Altmann, A.; Fischer, I.; Weislo, L. J.; Lanahan, M.; van Pée, K.-H.; Ligon, J. M., Functions Encoded by Pyrrolnitrin Biosynthetic Genes from Pseudomonas Fluorescens. *J. Bacteriol.* **1998,** *180*, 1939.

52. Yan, Q.; Philmus, B.; Chang, J. H.; Loper, J. E., Novel Mechanism of Metabolic Co-Regulation Coordinates the Biosynthesis of Secondary Metabolites in Pseudomonas Protegens. *eLife* **2017,** *6*, e22835.

53. Zhang, X.; Parry, R. J., Cloning and Characterization of the Pyrrolomycin Biosynthetic Gene Clusters from Actinosporangium Vitaminophilum Atcc 31673 and Streptomyces Sp. Strain Uc 11065. *Antimicrob. Agents Chemother.* **2007,** *51*, 946.

54. Ortega, M. A.; Cogan, D. P.; Mukherjee, S.; Garg, N.; Li, B.; Thibodeaux, G. N.; Maffioli, S. I.; Donadio, S.; Sosio, M.; Escano, J.; Smith, L.; Nair, S. K.; van der Donk, W. A., Two Flavoenzymes Catalyze the Post-Translational Generation of 5-Chlorotryptophan and 2-Aminovinyl-Cysteine During Nai-107 Biosynthesis. *ACS Chem. Biol.* **2017,** *12*, 548.

55. Hoffmann, T.; Müller, S.; Nadmid, S.; Garcia, R.; Müller, R., Microsclerodermins from Terrestrial Myxobacteria: An Intriguing Biosynthesis Likely Connected to a Sponge Symbiont. *J. Am. Chem. Soc.* **2013,** *135*, 16904.

56. Smith, D. R. M.; Uria, A. R.; Helfrich, E. J. N.; Milbredt, D.; van Pée, K.-H.; Piel, J. r.; Goss, R. J. M., An Unusual Flavin-Dependent Halogenase from the Metagenome of the Marine Sponge Theonella Swinhoei Wa. *ACS Chem. Biol.* **2017,** *12*, 1281.

57. Neumann, C. S.; Walsh, C. T.; Kay, R. R., A Flavin-Dependent Halogenase Catalyzes the Chlorination Step in the Biosynthesis of Dictyostelium Differentiation-Inducing Factor 1. *Proc. Natl. Acad. Sci. USA* **2010,** *107*, 5798.

58. Binz, T. M.; Maffioli, S. I.; Sosio, M.; Donadio, S.; Müller, R., Insights into an Unusual Nonribosomal Peptide Synthetase Biosynthesis: Identification and Characterization of the Ge81112 Biosynthetic Gene Cluster. *J. Biol. Chem.* **2010,** *285*, 32710.

59. Dorrestein, P. C.; Yeh, E.; Garneau-Tsodikova, S.; Kelleher, N. L.; Walsh, C. T., Dichlorination of a Pyrrolyl-S-Carrier Protein by Fadh2-Dependent Halogenase Plta During Pyoluteorin Biosynthesis. *Proc. Natl. Acad. Sci. USA* **2005,** *102*, 13843.

60. Pang, A. H.; Garneau-Tsodikova, S.; Tsodikov, O. V., Crystal Structure of Halogenase Plta from the Pyoluteorin Biosynthetic Pathway. *J. Struct. Biol.* **2015,** *192*, 349.

61. El Gamal, A.; Agarwal, V.; Diethelm, S.; Rahman, I.; Schorn, M. A.; Sneed, J. M.; Louie, G. V.; Whalen, K. E.; Mincer, T. J.; Noel, J. P.; Paul, V. J.; Moore, B. S., Biosynthesis of Coral Settlement Cue Tetrabromopyrrole in Marine Bacteria by a Uniquely Adapted Brominase-Thioesterase Enzyme Pair. *Proc. Natl. Acad. Sci. USA* **2016,** *113*, 3797.

62. Nakai, T.; Deguchi, T.; Frébort, I.; Tanizawa, K.; Okajima, T., Identification of Genes Essential for the Biogenesis of Quinohemoprotein Amine Dehydrogenase. *Biochemistry* **2014,** *53*, 895.

63. Chacón-Verdú, M. D.; Gómez, D.; Solano, F.; Lucas-Elío, P.; Sánchez-Amat, A., Lodb Is Required for the Recombinant Synthesis of the Quinoprotein L-Lysine-E-Oxidase from Marinomonas Mediterranea. *Appl. Microbiol. Biotechnol.* **2014,** *98*, 2981.

64. See Supporting Information.

65. Fu, L.; Niu, B.; Zhu, Z.; Wu, S.; Li, W., Cd-Hit: Accelerated for Clustering the Next-Generation Sequencing Data. *Bioinformatics* **2012,** *28*, 3150.

66. Schmartz, P. C.; Zerbe, K.; Abou-Hadeed, K.; Robinson, J. A., Bis-Chlorination of a Hexapeptide–Pcp Conjugate by the Halogenase Involved in Vancomycin Biosynthesis. *Org. Biomol. Chem.* **2014,** *12*, 5574.

67. Ashtekar, K.; Marzijarani, N.; Jaganathan, A.; Holmes, D.; Jackson, J. E.; Borhan, B., A New Tool to Guide Halofunctionalization Reactions: The Halenium Affinity (Hala) Scale. *J. Am. Chem. Soc.* **2014,** *136*, 13355.

68. Wood, W. J. L.; Patterson, A. W.; Tsuruoka, H.; Jain, R. K.; Ellman, J. A., Substrate Activity Screening: A Fragment-Based Method for the Rapid Identification of Nonpeptidic Protease Inhibitors. *J. Am. Chem. Soc.* **2005,** *127*, 15521.

69. Welch, C. J.; Gong, X.; Schafer, W.; Pratt, E. C.; Brkovic, T.; Pirzada, Z.; Cuff, J. F.; Kosjek, B., Miser Chromatography (Multiple Injections in a Single Experimental Run): The Chromatogram Is the Graph. *Tetrahedron: Asymmetry* **2010,** *21*, 1674.

70. Zawatzky, K.; Barhate, C. L.; Regalado, E. L.; Mann, B. F.; Marshall, N.; Moore, J. C.; Welch, C. J., Overcoming "Speed Limits" in High Throughput Chromatographic Analysis. *J. Chromatogr. A* **2017,** *1499*, 211.

71. Santanilla, A. B.; Regalado, E. L.; Pereira, T.; Shevlin, M.; Bateman, K.; Campeau, L.-C.; Schneeweis, J.; Berritt, S.; Shi, Z.-C.; Nantermet, P.; Liu, Y.; Helmy, R.; Welch, C. J.; Vachal, P.; Davies, I. W.;

Cernak, T.; Dreher, S. D., Nanomole-Scale High-Throughput Chemistry for the Synthesis of Complex Molecules. *Science* **2015,** *347*, 49.

72. Ryan, K. S., Biosynthetic Gene Cluster for the Cladoniamides, Bis-Indoles with a Rearranged Scaffold. *PLOS ONE* **2011,** *6*, e23694.
73. Wick, J.; Heine, D.; Lackner, G.; Misiek, M.; Tauber, J.; Jagusch, H.; Hertweck, C.; Hoffmeister, D., A Fivefold Parallelized Biosynthetic Process Secures Chlorination of Armillaria Mellea (Honey Mushroom) Toxins. *Appl. Environ. Microb.* **2016,** *82*, 1196.
74. Mak, W. S.; Tran, S.; Marcheschi, R.; Bertolani, S.; Thompson, J.; Baker, D.; Liao, J. C.; Siegel, J. B., Integrative Genomic Mining for Enzyme Function to Enable Engineering of a Non-Natural Biosynthetic Pathway. *Nat. Commun.* **2015,** *6*, 10005.
75. Breitbart, M.; Bonnain, C.; Malki, K.; Sawaya, N. A., Phage Puppet Masters of the Marine Microbial Realm. *Nat. Microbiol.* **2018,** *3*, 754.
76. Agarwal, V.; El Gamal, A. A.; Yamanaka, K.; Poth, D.; Kersten, R. D.; Schorn, M.; Allen, E. E.; Moore, B. S., Biosynthesis of Polybrominated Aromatic Organic Compounds by Marine Bacteria. *Nat. Chem. Biol.* **2014,** *10*, 640.
77. Heeb, M. B.; Kristiana, I.; Trogolo, D.; Arey, J. S.; Gunten, U. v., Formation and Reactivity of Inorganic and Organic Chloramines and Bromamines During Oxidative Water Treatment. *Water Res.* **2017,** *110*, 91.
78. Payne, J. T.; Butkovich, P. H.; Gu, Y.; Kunze, K. N.; Park, H. J.; Wang, D.-S.; Lewis, J. C., Enantioselective Desymmetrization of Methylenedianilines Via Enzyme-Catalyzed Remote Halogenation. *J. Am. Chem. Soc.* **2018,** *140*, 546.
79. Schnepel, C.; Kemker, I.; Sewald, N., One-Pot Synthesis of D-Halotryptophans by Dynamic Stereoinversion Using a Specific L-Amino Acid Oxidase. *ACS Catal.* **2019,** *9*, 1149.
80. Sánchez, C.; Zhu, L.; Braña, A. F.; Salas, A. P.; Rohr, J.; Méndez, C.; Salas, J. A., Combinatorial Biosynthesis of Antitumor Indolocarbazole Compounds. *Proc. Natl. Acad. Sci. USA* **2005,** *102*, 461.
81. Seibold, C.; Schnerr, H.; Rumpf, J.; Kunzendorf, A.; Hatscher, C.; Wage, T.; Ernyei, A. J.; Dong, C.; Naismith, J. H.; van Pée, K.-H., A Flavin-Dependent Tryptophan 6-Halogenase and Its Use in Modification of Pyrrolnitrin Biosynthesis. *Biocatal. Biotransform.* **2009,** *24*, 401.
82. Runguphan, W.; Qu, X.; O'Connor, S. E., Integrating Carbon–Halogen Bond Formation into Medicinal Plant Metabolism. *Nature* **2010,** *468*, 461.
83. Ismail, M.; Schroeder, L.; Frese, M.; Kottke, T.; Hollmann, F.; Paul, C. E.; Sewald, N., Straightforward Regeneration of Reduced Flavin Adenine Dinucleotide Required for Enzymatic Tryptophan Halogenation. *ACS Catal.* **2019,** *9*, 1389.
84. Shepherd, S. A.; Karthikeyan, C.; Latham, J.; Struck, A.-W.; Thompson, M. L.; Menon, B. R. K.; Styles, M. Q.; Levy, C.; Leys, D.; Micklefield, J., Extending the Biocatalytic Scope of Regiocomplementary Flavin-Dependent Halogenase Enzymes. *Chem. Sci.* **2015,** *6*, 3454.
85. Niklasson, M.; Andresen, C.; Helander, S.; Roth, M.; Kahlin, A.; Appell, M.; Mårtensson, L. G.; Lundström, P., Robust and Convenient Analysis of Protein Thermal and Chemical Stability. *Protein Sci.* **2015,** *24*, 2055.

86. Bloom, J. D.; Labthavikul, S. T.; Otey, C. R.; Arnold, F. H., Protein Stability Promotes Evolvability. *Proc. Natl. Acad. Sci. USA* **2006,** *103*, 5869.

TOC Graphic