Title

REC protein family expansion by the emergence of a new signaling pathway.

Permalink

https://escholarship.org/uc/item/65m1q830

Journal

mBio, 14(6)

Authors

Garber, Megan

Frank, Vered

Kazakov, Alexey

et al.

Publication Date

2023-11-22

DOI

10.1128/mbio.02622-23

Peer reviewed

AMERICAN SOCIETY FOR MICROBIOLOGY | mBio

| Evolution | Research Article

REC protein family expansion by the emergence of a new signaling pathway

Megan E. Garber,[1,2] Vered Frank,[2] Alexey E. Kazakov,[3] Matthew R. Incha,[2,4] Alberto A. Nava,[2,5] Hanqiao Zhang,[2,6] Luis E. Valencia,[2,6] Jay D. Keasling,[2,4,5,6,7,8] Lara Rajeev,[2] Aindrila Mukhopadhyay[1,2,3]

AUTHOR AFFILIATIONS See affiliation list on p. 17.

ABSTRACT This report presents multi-genome evidence that REC protein family expansion occurs when the emergence of new pathways gives rise to functional discordance. Specificity between residues in REC domain containing response regulators with paired histidine kinases is under negative purifying selection, constrained by the presence of other bacterial two-component systems signaling cascades that share sequence and structural identity. Presuming that the two-component systems can evolve by neutral amino acid changes (neutral drift) when purifying evolutionary constraints are relaxed, how might the REC protein family expand by amino acid changes when these constraints remain intact? Using an unsupervised machine learning approach to observe the sequence landscape of REC domains across long phylogenetic distances, we find that within-gene recombination, a subcategory of gene conversion, switched the effector domain and, consequently, the regulatory context of a duplicated response regulator from transcriptional regulation by σ54 to that by σ70. We determined that the recombined response regulator diverged from its parent by episodic diversifying selection and neutral drift. Functional experiments of the parent of recombined response regulators in a model *Pseudomonas putida* KT2440 model system revealed that the parent and recombined response regulators sense and respond to different carboxylic acids. Finally, a residue-switching experiment using structural predictions and functional characterization suggests that the new residues in the recombined regulator could form a new interaction interface and mediate condition-specific phosphotransfer. Overall, our study finds that genetic perturbations can create conditions of functional discordance, whereby the REC protein family can evolve by episodic diversifying selection.

IMPORTANCE We explore when and why large classes of proteins expand into new sequence space. We used an unsupervised machine learning approach to observe the sequence landscape of REC domains of bacterial response regulator proteins. We find that within-gene recombination can switch effector domains and, consequently, change the regulatory context of the duplicated protein.

KEYWORDS large protein families, two-component regulatory systems, evolution, gene regulation, genetic recombination, domain swapping, signal transduction, response regulator

Protein families are a group of proteins that share a common ancestor, showing sequence conservation across long phylogenetic distances (1–4). Amino acid consensus motifs are used to understand protein family function and specificity; active-site residues are fixed and tell us about the chemistry the protein family performs or undergoes, while variable residues determine the protein's specificity for different substrates or ligands, binding affinity, reaction kinetics, protein stability, etc. —this report will focus on specificity for different substrates or ligands. Variable,

Editor Michael T. Laub, Massachusetts Institute of Technology, Cambridge, Massachusetts, USA

Address correspondence to Aindrila Mukhopadhyay, amukhopadhyay@lbl.gov.

J.D.K. has financial interests in Amyris, Lygos, Demetrix, Napigen, and Maple Bio.

See the funding table on p. 17.

Received 28 September 2023
Accepted 20 October 2023
Published 22 November 2023

specificity-determining, residues form the molecular basis of tightly regulated signal transduction (5–7), such as those found in phosphotransfer receiver (REC) domains of response regulators, allowing for specific, context-dependent, protein-protein interactions.

Members of the REC protein family (PF00072) function as phosphotransfer receivers in bacterial signaling cascades called two-component systems. Most bacterial genomes encode dozens to over a hundred of these systems (8–10). Two-component systems modulate important functions in bacterial pathogenesis, antibiotic resistance, nutrient use (e.g., carbon, nitrogen, or phosphate utilization), fitness in a microbiome context, and a vast number of other functions (5, 10), making them attractive targets for addressing infection (5) and pathogenesis (11, 12) and for applications in synthetic biology and biotechnology (13–15). Despite large numbers of highly analogous systems encoded in a single genome, when a signal is recognized by a histidine kinase, phosphotransfer (phosphorylation or dephosphorylation) to its cognate response regulator occurs with precise biochemical specificity (16–18). The phosphorylated cognate response regulator then regulates cellular functions, mainly via transcription, chemotaxis, or modulation of second messengers (Fig. 1A). In response regulators, N-terminal REC domains are fused to a broad range of C-terminal effector domains and predominantly function as transcription regulators (Fig. 1B). Previous structural and biochemical studies of the REC protein family showed that variable amino acid sequences in active-site adjacent alpha-helices that form the interaction interface determine interaction with cognate histidine kinases to mediate highly specific, context-dependent, phosphotransfer to the active-site aspartate residue (16, 18–20).

How the REC protein family evolves is of interest due to its ubiquitous role in cellular signaling and regulation of cellular function. For the purposes of this report, we define three key mechanisms in the evolution of the REC protein family: (i) neutral drift: synonymous or nonsynonymous amino acid changes that occur through the natural processes of mutation (21, 22); (ii) negative purifying selection: selection against nonsynonymous amino acid changes; and (iii) episodic diversifying selection: selection for nonsynonymous amino acid changes. In the evolution of bacterial two-component systems, the interactions between cognate histidine kinases and response regulators are under negative purifying selection to prevent deleterious interactions between non-cognate two-component systems. Presumably, loss of a single two-component system—due to gene loss or growth in conditions that suppress two-component system activity—can relax these purifying constraints, such that new two-component systems can sample the unused combinations of amino acids sequences that code for the interaction interfaces (23). It has been shown that while two components can sample new sequences, the phosphotransfer functions between encoded cognate kinases and regulators do not break, and phosphotransfer levels can be maintained during the evolutionary process of nonsynonymous amino acid changes (24). This implies that REC protein family expansion likely occurs by neutral drift when purifying constraints are relaxed, but we still do not understand if new combinations of sequences forming new interaction specificities are sampled when these constraints remain intact (Fig. 1C).

One challenge in understanding the evolution of large protein families is parsing relationships between members of the protein family from amino acid sequences alone. Previous studies (7, 25, 26) using phylogenetic and clustering methods to parse molecular information from REC domains from species that span short evolutionary distances (population of a species or species within a genus) showed that members of the REC protein family resulted from independent gene duplications, sharing a common ancestor in bacterial history. Studies comparing domain architectures of REC domains from extant species (8, 27, 28) suggest that REC domains evolve vertically, passing the fused effector domains to daughter REC domains (Fig. 1C). To overcome the challenge of understanding how REC protein family expansion occurs over long phylogenetic distances, we build upon insights from these previous studies, making use of the

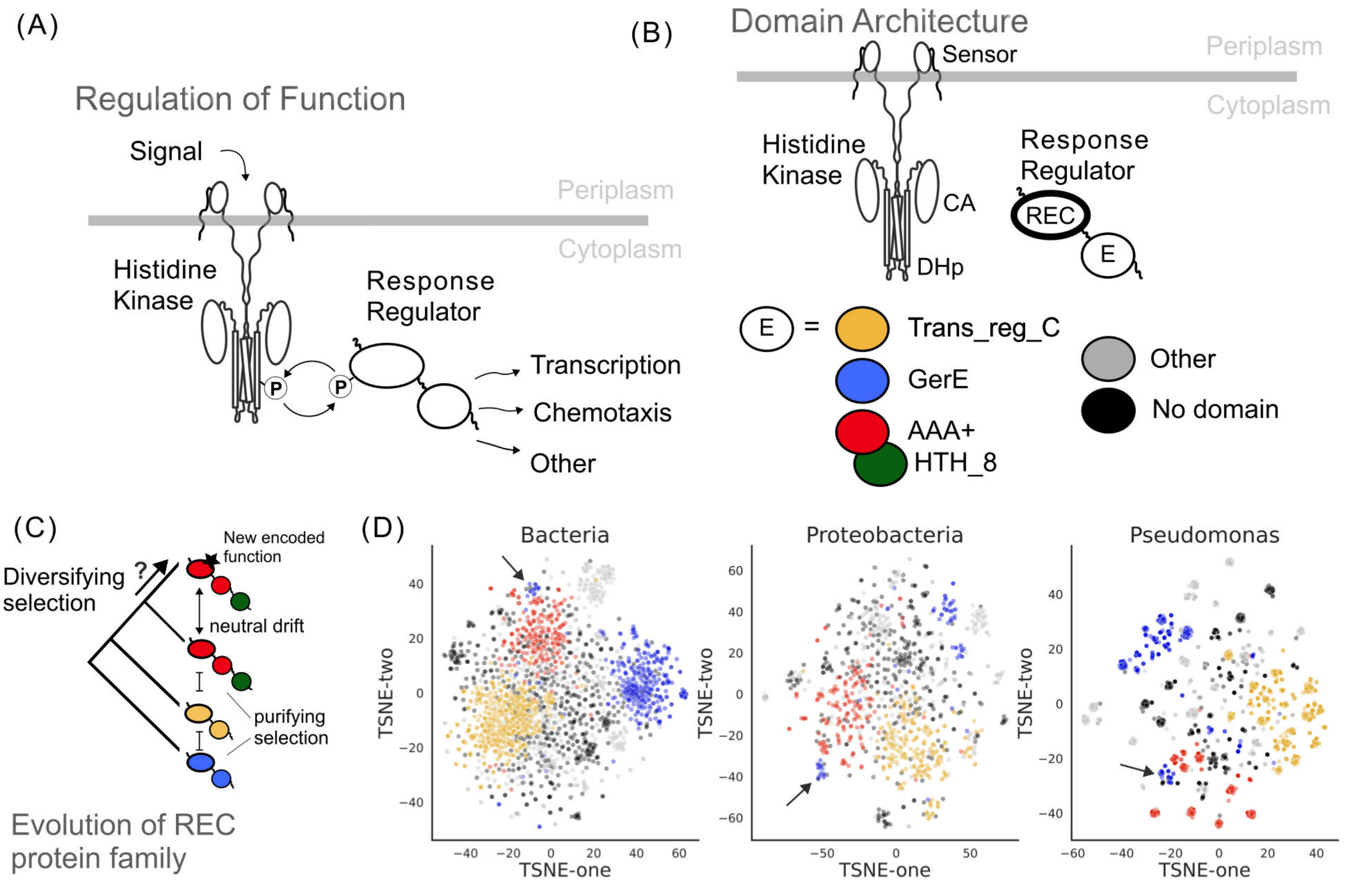

FIG 1 Functional context for insight into the evolution of the REC protein family. (A) REC protein family in the context of bacterial two-component systems: membrane-bound histidine kinases sense a signal and undergo phosphotransfer to a cognate response regulator, which regulates cellular functions. (B) Domain architecture of the two-component system. The histidine kinases' sensor, dimerization and histidine phosphotransfer (DHp), and catalytic (CA) domains coordinate the initial autophosphorylation. The signal is transmitted by phosphotransfer to the cognate response regulators' receiver (REC) domain (bolded oval). The effector domain (oval labeled "E") completes the signaling cascade. Effector domains "E" are mostly DNA-binding, and the most abundant are Trans_reg_C (yellow), GerE (blue), AAA+ (red), and HTH_8 (green). Because HTH_8 domains often co-occur with AAA+, they are shown as overlapping red and green ovals—for simplicity, AAA+-HTH_8 domains are called AAA+ domains throughout the text. Effector domains with non-DNA-binding function are shown as other (gray) and no domains (black). The effector domain color coding shown in this figure is used for all subsequent figures. (C) Hypothetical tree showing model REC protein family evolution. Assuming that REC domains share a common ancestor and that REC domains fused to the same effector domains are more closely related than those fused to different effector domains, showing regulators with distinct domain architectures and colors (refer to B) at the tips of the tree with coalescent branches. Neutral drift occurs when purifying constraints are relaxed; however, what causes REC domains to diversify and sample new sequence combinations that encode new specificities when constraints are intact remains unknown. (D) Similarity between members of the REC protein families showing the relationship between REC domain sequence alignments with gap replacement and explaining 95% of the variation between the sequences. REC domains were sampled from species found within the taxonomic rank (kingdom, Bacteria; phylum, Proteobacteria; or genera, *Pseudomonas*) labeled at the top of each plot and sampled as described in Materials and Methods. After species sampling, we identified all proteins with REC domains in each of the sampled species' genomes and aligned their sequences. Points represent the sequences of unique REC domains from the sampled genomes; points that are close together share sequence identity with each other. Each REC domain is annotated by the identity of its effector domain Trans_reg_C (yellow), GerE (blue), AAA+ (red), other (gray), and no domains (black). Note that the location of REC domains is not fixed in the plot between independent runs of the algorithm; however, the relative distribution between the points is visually consistent between runs. Black arrow indicates the REC domains linked to GerE effector domains as a result of within-gene recombination of parent REC domains fused to AAA+ effector domains. TSNE: t-Distributed Stochastic Neighbor Embedding.

over 200,000 sequenced bacterial genomes, to develop a method that can parse the evolution of the sequenced REC domains in extant, sequenced species (10).

In this study, we ask whether neutral drift alone or another evolutionary mechanism can explain how new members of protein families emerge. Using an unsupervised machine learning (ML) strategy, we explore the amino acid sequence landscape of the REC protein family. We identify a subfamily of the REC protein family that emerged and expanded after a within-gene recombination event changed the effector domain of this REC subfamily. We show that functional discordance between parent and daughter signaling pathways, which occurred after the within-gene recombination event, gave rise to a new structural interface that, we propose, facilitates phosphotransfer.

RESULTS

ML approach leads to (re)-discovery of recombination events

To understand how large protein families expand and evolve, we applied an unsupervised ML algorithm that uses dimensional reduction to group similar domains based entirely on the identities and positions of amino acid sequences in an alignment. We established this strategy for large protein families, using the REC protein family (PF00072). To ensure diverse sampling of REC domains across the bacterial clade, we randomly sampled bacterial genomes from three taxonomic ranks, kingdom (Bacteria), phylum (Proteobacteria), and genera (*Pseudomonas*), generating independent data sets of REC sequences. To control for misgrouping due to gaps in the sequence alignments, we applied the algorithm to REC sequence alignments with and without gap replacement by randomly generated amino acids. Using the t-distributed stochastic neighbor embedding algorithm (29) with N principal components to explain 95% of the variation in the sequences, we projected the dimensionally reduced output of the ML algorithm in two dimensions (Fig. 1D and 2; Fig. S1 to S3). Instead of assigning cluster boundaries to the mapped data—which would be an inappropriate use of the dimensional reduction algorithm (29)—we color-coded each REC domain (single point on the plot) with independent information about the effector domain it is fused to in nature. The data structure of the REC protein family becomes evident by highlighting the predominant effector domains that have transcriptional function, GerE, Trans_reg_C, and AAA+. Interestingly, few differences in data structure were apparent between REC sequences with and without gap replacement (Fig. 1D; Fig. S1A). Furthermore, when we applied the ML algorithm to scrambled amino acid sequence alignments with gap replacement, the REC domains appeared randomly distributed; however, REC domains with scrambled sequences without gap replacement appeared to be arranged non-randomly (Fig. S2). We propose that replacing gaps with randomly generated amino acids does not affect the results of the unscrambled sequences, because the data structure is driven by the similarity among the variable residues. In the case of random gap replacement, the gaps are treated like noise, whereas if they are not replaced, the gaps are treated like the highly conserved residues, hence the appearance of a non-random arrangement in the scrambled sequence without gap replacement (Fig. S2). The non-random arrangement in the scrambled sequence (Fig. S2) also suggests that sequence ambiguity assigned by the hidden Markov model alignment strategy (30) results in more ambiguous alignments for REC amino acid sequences that share less sequence identity with the response regulators used to build the REC consensus model (7). Based on these collective results, sequences with gap replacement provide a data structure independent of the biases introduced by the alignment strategy. To demonstrate the consistency of the method, we repeated the random sampling of REC sequences two more times and applied the strategy with gap replacement to each sampling, each showing similar data structures to the first sampling (Fig. S1B and C). Together, these results validate this strategy, allowing us to infer from it the evolutionary history of the REC protein family.

The REC protein family is divided into distinct groups, which we will call subclusters; each REC sequence is represented by a point in the two-dimensional sequence landscape (Fig. 1D and 2; Fig. S1 to S3) and shares sequence identity with neighboring points. Subclusters that are qualitatively closer together tend to share the identity of their fused effector domains (Fig. 1B to D); this result confirms our expectation that the REC protein family evolved vertically and carried effector domain architectures through to the next generations (Fig. 1C). Interestingly, the subclusters become more pronounced as we sample REC domains from lower taxonomic ranks (kingdom, phylum, and genera); this signature of single points appearing more attracted to similar sequences and, at the same time, more repellent to dissimilar sequences as we sample sequences from shorter and shorter evolutionary distances suggests that paralogous REC domains are indeed under negative purifying selection at the organismal level. Yet, despite purifying selection, the REC protein family has diversified over the course of bacterial evolution, which can be observed by the multiplication of subclusters in the REC domain landscape at each evolutionary distance.

Although we understand that protein families can diversify neutrally by sampling new combinations of amino acid sequences when purifying constraints are relaxed, we do not know whether they can diversify when these constraints remain intact. Selection to eliminate crosstalk (communication between distinct pathways) does not completely explain REC protein family expansion, as crosstalk can be tolerated and does not always result in a loss in fitness to the organism and, in cases of cross-regulation (communication between intertwined pathways), can also provide some fitness benefit to the organism (17, 31). For diversifying selection to occur, changes to the signaling system would need to affect cellular function, such as changes that can result from within-gene recombination (32–37) (a subcategory of gene conversion, also known as domain swapping or domain shuffling).

Within-gene recombination has occurred many times throughout the evolution of the REC protein family and has been previously documented (7, 25, 38, 39). We chose to focus on the largest, most pronounced instance of within-gene recombination in our data set. This event can be visualized in the two-dimensional sequence landscape by the presence of two major clusters of REC sequences with GerE effector domains (colored in blue); the GerE cluster that lies near REC sequences with AAA+ effectors (colored in red) is the product of a within-gene recombination event that changed the effector domain of a parent response regulator with a AAA+ effector to GerE (7, 25). We focus on this event for three reasons: (i) it has been previously documented (7); (ii) it is a clear example of protein family expansion, given that the two distinct clusters of GerE-fused response regulators are highly visible in the two-dimensional sequence landscape; and (iii) the recombination event is by default (and as discussed in Pao and Saier (7)) linked to functional discordance, because AAA+ and GerE domains regulate transcription with distinct sigma (σ) factors, σ70 and σ54, which are active in distinct cellular contexts (40–43).

When did this recombination event occur?

To determine whether within-gene recombination can lead to protein family expansion, we needed to determine whether these proteins were under diversifying selection as a result of the within-gene recombination event. Using our unsupervised ML strategy, we were able to find that the recombined response regulator likely emerged sometime in the Proteobacteria lineage, based on the presence of a second GerE subcluster in Proteobacteria and its absence in other phyla (Fig. S3A and B). More specifically, the GerE subcluster was expanded in the subclades of Alphaproteobacteria, Betaproteobacteria, and Gammaproteobacteria (Fig. 2; Fig. S3C and S4) but was not observable in Deltaproteobacteria. In the species tree of life (44), Alphaproteobacteria predates Betaproteobacteria and Gammaproteobacteria; we therefore reasoned that the within-gene recombination event had occurred sometime in the Alphaproteobacteria lineage. This insight enabled us to repurpose the REC sequence landscapes to home in on the REC protein family in Alphaproteobacteria (Fig. 2; Fig. S4) and search for the REC sequences of interest (Fig. 3A) for detection of differential rates of evolution (Fig. 3B; Fig. S5; Table S2). We found that the within-gene recombination event caused episodic diversifying selection on the branch of response regulators that have GerE domains in extant Alphaproteobacteria species (Fig. S5A). The consensus between the REC sequences of the parent and recombined response regulators highlights the shared sequence identity between the parent REC domains fused to AAA+ domains and recombined REC domains fused to GerE domains, apart from several highly conserved residues, in the recombined REC domains fused to GerE domains (Fig. 3C and D). Using a method

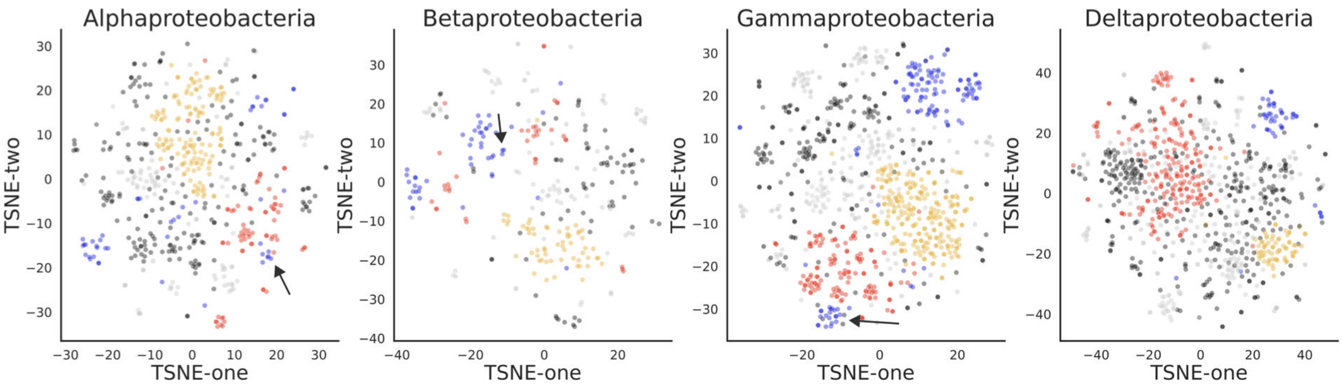

FIG 2 Within-gene recombination that changed the parent REC domain fused to AAA+ domain to REC domain fused to GerE domain occurred during the Alphaproteobacteria lineage. Assessment of similarity between members of the REC protein families showing the relationship between REC domain sequence alignments with gap replacement and explaining 95% of the variation between the sequences. As in Fig. 1D, REC domains are annotated by the identity of its effector domain Trans_reg_C (yellow), GerE (blue), AAA+ (red), other (gray), and no domains (black). Black arrow indicates the REC domains fused to GerE effector domains as a result of within-gene recombination of parent REC domains fused to AAA+ effector domains.

that identifies individual sites subject to episodic diversifying selection determined by the differential rates of nonsynonymous to synonymous amino acid changes (45) (Table S2; see Supplemental Data A at https://doi.org/10.6084/m9.figshare.24205968.v2), we found that among the highly conserved residues, two residues in the recombined REC domains fused to GerE domains were also under positive diversifying selection. These results show that after the within-gene recombination, episodic diversifying selection occurred in the evolution of the recombined regulators; however, it does not explain the roles of the conserved and/or selected residues in their functional context of regulating gene expression in response to signals.

Functional characterization of the recombined regulators: understanding the roles of conserved residues

To understand how the conserved and/or selected residues shaped the expansion of the recombined response regulators, we needed to understand the function of these regulators—the input signals they respond to and the genes they regulate—which were not completely known before this study. We set out to determine those functions, by pursuing functional experiments in *P. putida* KT2440, which proved advantageous for several reasons: (i) *Pseudomonas putida* is a model organism in bioremediation and synthetic biology applications (46, 47); therefore, many genetic and functional genomic experiments have been applied to understand this strain's broad metabolism, allowing us to rapidly phenotype regulators to determine their functions; (ii) *P. putida* KT2440 is a Gammaproteobacteria and is among the species where we have identified the parent AAA+ response regulators and the recombined GerE domain response regulator (note that these regulators are not present in *Escherichia coli*). Two functional genomic screens were used to determine: (i) the media conditions that cause randomly barcoded transposon insertion (RB-Tn) mutants at positions of the response regulators to grow more slowly in a population of other RB-Tn mutants (Fig. S6A) and (ii) the genomic location of the transcription factor-binding sites the response regulators use to regulate transcription (Fig. S5B; Table S3; Data S1 and S2). RB-Tn mutants of the response regulators showed growth phenotypes in defined media with carboxylic acids as the sole carbon source and transcription factor-binding sites were found upstream of genes related to carboxylic acid assimilation; we, therefore, proposed that the regulators might respond to carboxylic acids to transcriptionally regulate a set of genes involved in carboxylic acid assimilation (Fig. 4A). We validated these hypotheses with green fluorescent protein (GFP) reporter assays (Fig. 4B; Fig. S7), finding that the recombined regulator in *P. putida* responds to butyrate and regulates beta-oxidation (48), while

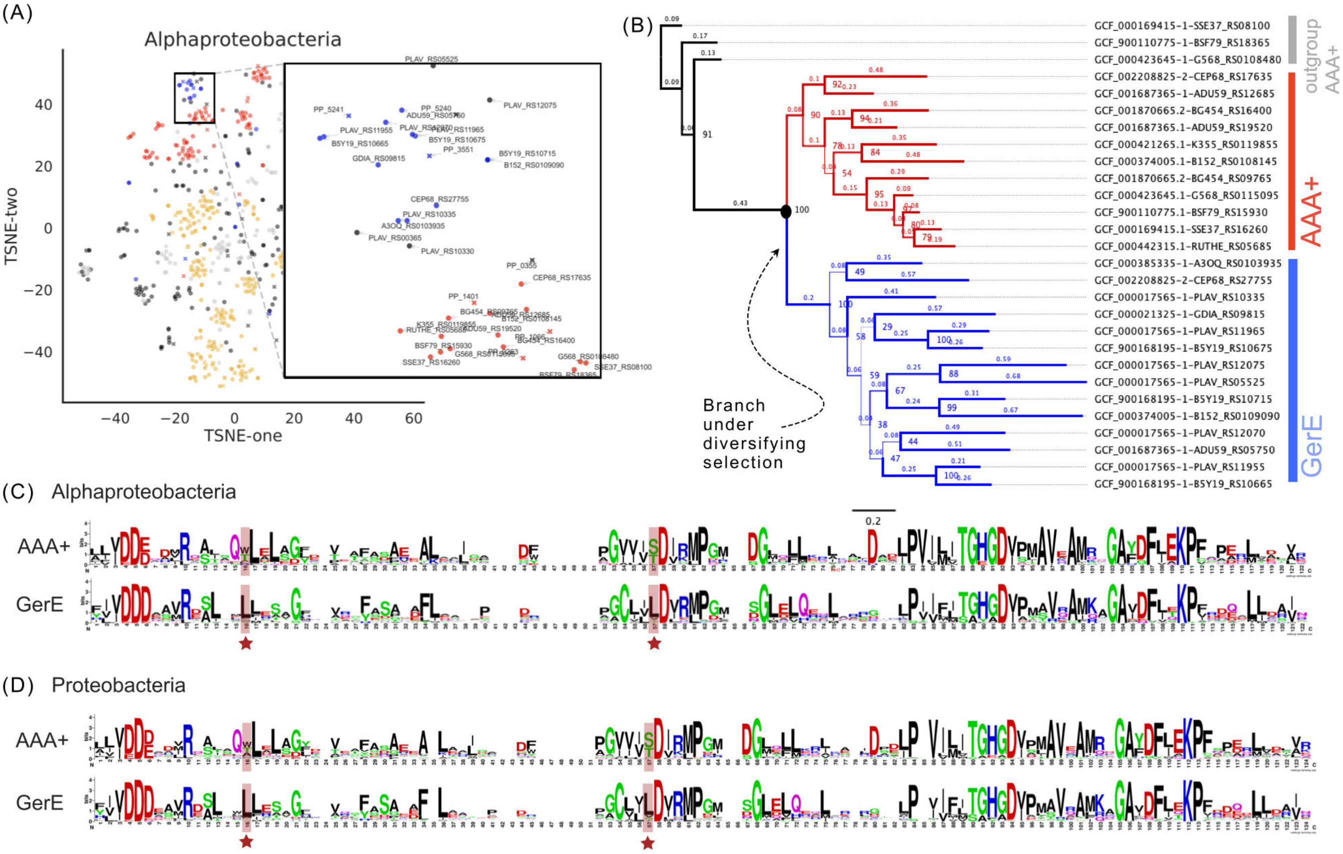

FIG 3 Episodic diversifying selection in the REC domain occurred after within-gene recombination that changed the parent REC domain fused to AAA+ domain to REC domain fused to GerE domain. (A) Fig. 2 data (Alphaproteobacteria panel) displayed as a dimensionally reduced map showing REC domains from *P. putida* (x's) or Alphaproteobacteria (circles). The recombined cluster (black box) becomes apparent using two-component systems from *P. putida* that were previously known to be among the recombined clusters, PP_1066 and PP_3551. The zoom panel shows the protein names for REC domains in the recombined cluster. (B) The REC domains in the Alphaproteobacteria recombined cluster (Fig. 3A) (circles) are used to construct a domain tree using GCF_000169415-SSE37_RS08100 as an outgroup. Branches are red for AAA+ or blue for GerE (or no domain). Bootstrap supports are labeled at the nodes and shown as branch thickness. Nodes under episodic diversifying selection are indicated by an arrow and black circle. (C and D) Amino acid sequences of the REC domains were aligned and separated into two distinct groups (AAA+ or GerE) based on the identity of their effector domains and were used to generate WebLogo consensus motifs. Residues under positive diversifying selection (red background and a red star).

the parent regulator responds to glutamate and regulates amino acid assimilation. The overlap in chemical structure between carboxylic acids (Fig. 4A) adds context of missignaling by chemical sensing to our proposed model for diversifying selection of the highly conserved residues in the recombined regulator, but we needed to investigate whether the evolved residues could have a role in breaking interaction between the parent and recombined regulators' signaling pathways.

The domain-based tree of the parent and recombined regulators in Alphaproteobacteria not only revealed that episodic diversifying selection occurred after the emergence of the recombined regulator but also revealed that the parent regulator underwent a second and third duplication event after the initial within-gene recombination (Fig. 3; Fig. S4). We did not detect instances of episodic diversifying selection on these post-recombination, duplicated REC domain branches and thus propose that they diverged from their parent REC domain by neutral drift. From functional genomic experiments, we determined that these regulators are also regulated by carboxylic acids and regulate carboxylic acid assimilation (Fig. S5, S7A andB). This result raises an important distinction between the regulators that duplicated after the recombination event and the recombined regulators. Despite the same possibility of crosstalk in response to carboxylic

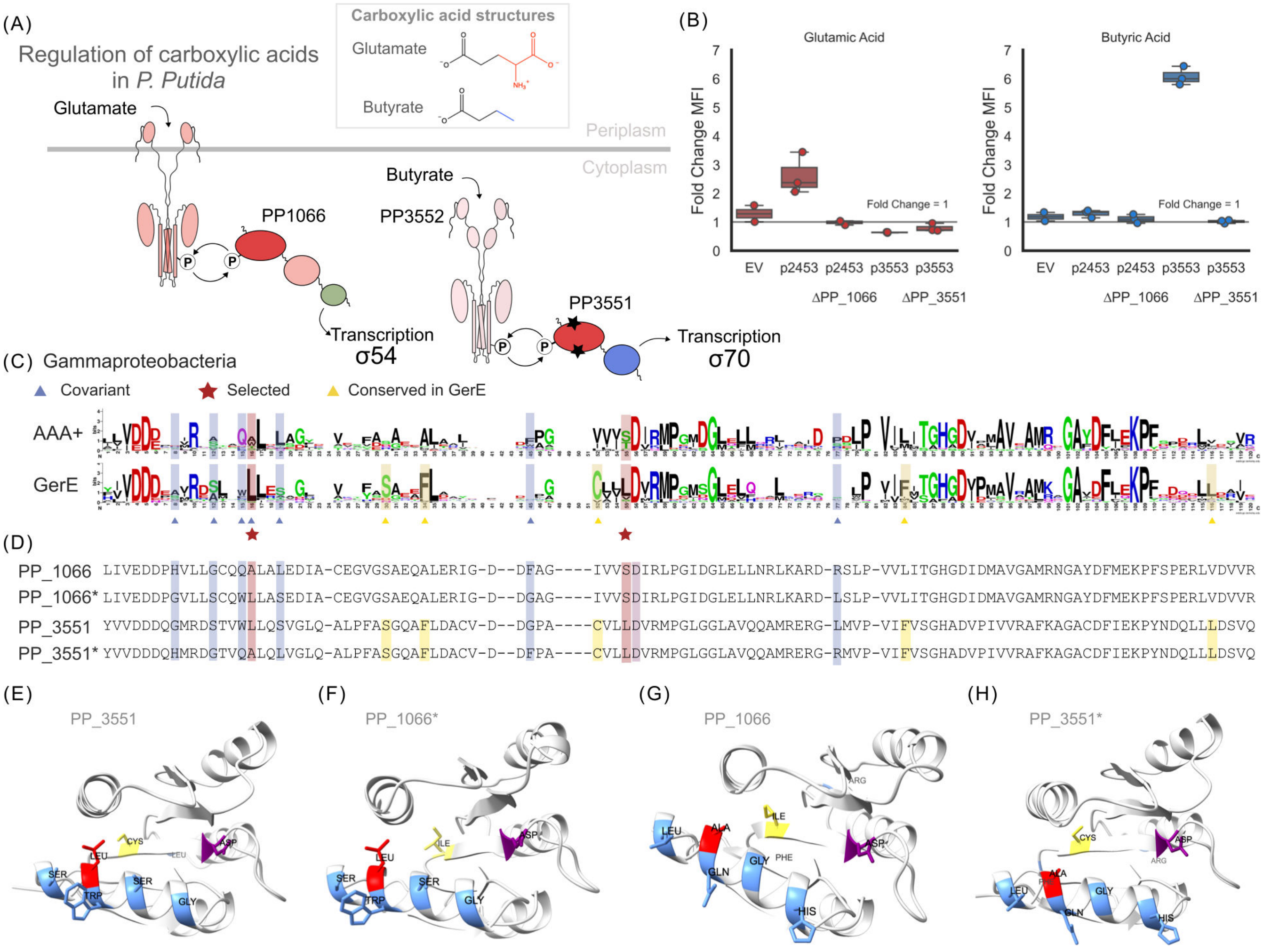
(A)
Regulation of carboxylic acids in P. Putida
Carboxylic acid structures
Glutamate
Butyrate
Glutamate
Periplasm
Cytoplasm
Butyrate
PP1066
PP3552
PP3551
P
P
Transcription σ54
Transcription σ70
(B)
Glutamic Acid
Butyric Acid
Fold Change MFI
Fold Change = 1
0
1
2
3
4
5
6
7
EV
p2453
p2453
p3553
p3553
ΔPP_1066
ΔPP_3551
(C) Gammaproteobacteria
Covariant
Selected
Conserved in GerE
AAA+
GerE
(D)
PP_1066 LIVEDDPHVLLGCQQALALEDIA-CEGVGSAEQALERIG-D--DFAG----IVVSDIRLPGIDGLELLNRLKARD-RSLP-VVLITGHGDIDMAVGAMRNGAYDFMEKPFSPERLVDVVR
PP_1066* LIVEDDPGVLLSCQWLLASEDIA-CEGVGSAEQALERIG-D--DGAG----IVVSDIRLPGIDGLELLNRLKARD-LSLP-VVLITGHGDIDMAVGAMRNGAYDFMEKPFSPERLVDVVR
PP_3551 YVVDDDQGMRDSTVWLLQSVGLQ-ALPFASGQAFLDACV-D--DGPA----CVLLDVRMPGLGGLAVQQAMRERG-LMVP-VIFVSGHADVPIVVRAFKAGACDFIEKPYNDQLLLDSVQ
PP_3551* YVVDDDQHMRDGTVQALQLVGLQ-ALPFASGQAFLDACV-D--DFPA----CVLLDVRMPGLGGLAVQQAMRERG-RMVP-VIFVSGHADVPIVVRAFKAGACDFIEKPYNDQLLLDSVQ
(E)
PP_3551
(F)
PP_1066*
(G)
PP_1066
(H)
PP_3551*
SER
LEU
TRP
SER
GLY
CYS
ASP
ILE
ALA
GLN
HIS
ARG
PHE

FIG 4 Parent and recombined response regulators in *P. putida* KT2440 are regulators of carbon assimilation and evolved by episodic diversifying selection to break an overlap in specificity. (A) Model of regulation of carboxylic acids in *P. putida* KT2440 by the parent (PP1066) and recombined (PP3551) response regulators. The parent system (PP1067/PP1066, red) is activated by glutamate and uses AAA+-dependent σ54 transcription machinery. The recombined system (PP3552/PP3551, also red) is activated by butyrate and uses GerE-dependent σ70 transcription machinery (the recombined GerE domain, blue). Based on transmembrane or sensing domains in the histidine kinases, PP1067 is modeled as membrane bound with a dCache domain for sensing glutamate and PP3351 as cytosolic with two PAS domains for sensing butyrate. The residues in PP3551 REC domain under diversifying selection are represented by black stars on the PP3551 REC domain. Chemical similarity between glutamate and butyrate is shown in the box. (B) Fold change of the median fluorescence intensity (MFI) of control strains bearing plasmids of the indicated upstream promoter region [empty vector (EV), p2435, p3553] driving expression of the GFP reporter. Strains grown with glutamic acid or butyric were compared to strains grown without an inducer. Centerline, median; box limits, upper and lower quartiles; whiskers, 1.5× interquartile range; points with black diamonds, outliers; $n = 3$. (C) Amino acid sequences of the REC domains of Gammaproteobacteria species were aligned and separated into two distinct groups (AAA+ or GerE) based on the identity of their effector domains and were used to generate WebLogo consensus motifs. Residues of interest: covarying residues (Fig. S9) (blue background and blue triangle); residues under positive diversifying selection (red background and red star), and residues that show conservation in the GerE group but are not under positive diversifying selection (yellow background and yellow triangle). (D) Alignment of PP1066 and PP3551 REC domains. Covariant (blue), selected (red), conserved but not selected (yellow), and active aspartate (purple) residues are highlighted by the indicated background color. PP1066* and PP3551* match the sequence of PP1066 and PP3551, respectively, but are switched at the covariant positions (blue). (E to H) Structural impact of switching residues: modeled structures of PP3551 (E), PP1066* (F), PP1066 (G), and PP3551* (H) of the full-length protein using AlphaFold are shown and color-coded as in D.

acids in an ancestral organism, the post-recombination duplicated regulators diverged neutrally, while the recombined regulators did not.

To understand if functional discordance following within-gene recombination plays a role in evolution in the REC protein family, we asked whether the identified amino acid changes in the recombined regulator enable it to interact specifically with its cognate kinase. To answer this question, we applied a residue-switching thought experiment, in which we computationally switched the six residues that tend to covary with cognate kinases across *Pseudomonas* species (Fig. S8); the list of six residues included a conserved leucine residue, which is also under episodic diversifying selection. The list, however, excluded a second conserved and selected leucine residue, as well the other conserved residues that were not found to be under selection (Fig. 4C and D). For the purposes of comparing residues that were under selection and those that were just found to be conserved, we will focus on the first leucine, which falls among the covariant residues, and compare it to a conserved cysteine that was not found to be under episodic diversifying selection. To understand the physical role of these residues in PP_3551, we turned to the structural prediction of PP_3551, PP_1066, and their respective mutants determined by AlphaFold (49). In PP_3551, the conserved/selected leucine residue (Fig. 4E and F; Fig. S8) sits inside of the alpha-helix that canonically interacts directly with its cognate kinase. It is therefore plausible that the new interface is responsible for the interaction of PP_3551 with its cognate kinase. The conserved but not selected cysteine residue lies within the same beta-sheet as the active site for phosphotransfer (conserved aspartate residue) (Fig. 4G and H) also making it a candidate for mediating communications with its cognate kinase.

Do the conserved and/or selected residues make or break the condition-specific responses of the parent regulator?

We speculated that the conserved residues could mediate contact with the cognate kinase of the recombined regulator to facilitate phosphotransfer. So we asked whether changes to the glutamic acid responsive parent regulator, PP_1066, affect the regulator's native behavior under activating conditions. To test this question, we complimented GFP reporter strains (using the p2453 promoter) of *P. putida* KT2440 Δ*PP1066*Δ*PP3551* with four variants of PP_1066: (i) the native PP_1066, (ii) PP_1066* (* indicates covariant mutations as described above), (iii) PP_1066I52C, and (iv) PP_1066*I52C (* indicates covariant mutations as described above) (Fig. S10). High basal expression of GFP without an inducer for both the mutant regulators and the native regulator suggested that the amino acid changes we introduced did not impede the regulator's ability to turn on the expression of GFP through the reporter's promoter (Fig. S10C). To determine whether the tested residues play a role in phosphotransfer during condition-specific activation, we compared activation of GFP expression with two signals, glutamic acid and butyric acid. The covariant mutations in PP_1066* appear to break the regulator's native response to both signals, whereas the I52C mutation only breaks the regulator's native response in glutamic acid. These data support the hypothesis that the conserved residues mediate phosphotransfer under activating conditions; however, we still do not know which histidine kinases are responsible for phosphotransfer under either inducing condition. Collectively, these results show that both episodic diversifying selection and neutral drift occurred after the within-gene recombination event, forging new interactions that mediate activation by phosphotransfer.

DISCUSSION

The REC protein family undergoes three mechanisms of evolution, neutral drift, negative purifying selection, and, as determined by this study, episodic diversifying selection (Fig. 5A and B). Prior to this study, it was understood that neutral drift occurs when crosstalk/cross-regulation between communicating signaling systems does not affect an organism's fitness (17, 24, 50–52). It was also established that negative purifying selection occurs when crosstalk between signaling systems results in dysregulation of cellular function affecting an organism's fitness (23). This study demonstrates that REC protein family expansion can occur by episodic diversifying selection and has occurred in a test case, in which a new signaling system emerged by within-gene recombination and communication between the old and new signaling systems affected cellular fitness.

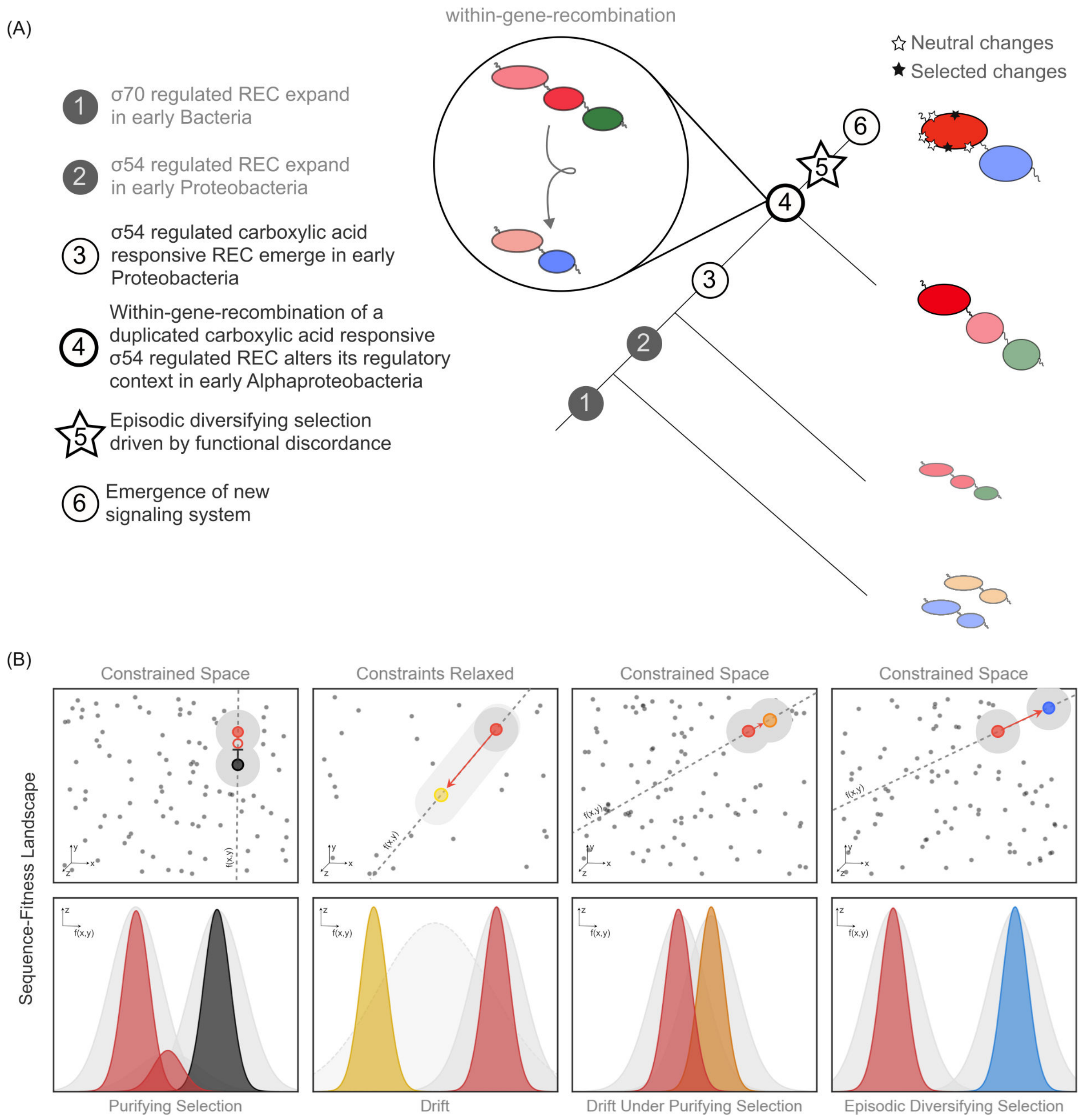
(A)
within-gene-recombination
☆ Neutral changes
★ Selected changes
1 σ70 regulated REC expand in early Bacteria
2 σ54 regulated REC expand in early Proteobacteria
3 σ54 regulated carboxylic acid responsive REC emerge in early Proteobacteria
4 Within-gene-recombination of a duplicated carboxylic acid responsive σ54 regulated REC alters its regulatory context in early Alphaproteobacteria
5 Episodic diversifying selection driven by functional discordance
6 Emergence of new signaling system
1
2
3
4
5
6
(B)
Sequence-Fitness Landscape
Constrained Space
Constraints Relaxed
Constrained Space
Constrained Space
f(x,y)
y
x
z
Purifying Selection
Drift
Drift Under Purifying Selection
Episodic Diversifying Selection

FIG 5 Models of the evolution of the REC protein family. (A) Model of the evolutionary tree of the REC protein family, updated by this study. We assume that all REC domains share a common ancestor and that REC domains fused to the same effector evolved vertically, showing regulators with distinct domain architectures and colors (refer to Fig. 1B) at the tips of the tree with coalescent branches. Note that the order of events in Steps 1–3 is uncertain. Selected changes (black stars) and neutral changes (white stars) occurred after the within-gene recombination that changed a parent REC domain fused to AAA+ domain to a recombined REC domain fused to a GerE domain. (B) Models of the evolution of the REC protein family in a sequence-fitness landscape. Top and bottom panels show the sequence-fitness landscape as projections of sequence and fitness. The top panes show the *x*,*y* plane representing the available combinations of amino acid sequences to the REC protein family, where *z* is the fitness (or distribution of REC domains with amino acid sequences defined by the *x*,*y* plane). If the fitness of a domain is high, it is shown as a black point on the plot; the boundaries of the fitness (or possible combinations of amino acid sequences that allow the domain to maintain its specificity) for the domains that are colored in red, yellow, orange, or blue are shown as a gray radius. Pointed arrows show the direction of positive selection or neutral drift; blunted arrows show the direction of negative selection. The bottom panes show the fitness as a function of the cross-section in the *x*,*y* plane [*f*(*x*,*y*)] (dashed line in the top panel) of the domains that are colored red, yellow, orange, or blue. Each column shows a different mode of evolution discussed in this report (from left to right: purifying selection, drift, drift under purifying selection, and episodic diversifying selection).

Specifically, our results show that a within-gene recombination in a duplicated response regulator of carbon assimilation changed the context in which this new regulator performed its transcriptional function (changing the effector domain from its parent regulator's σ54-dependent transcription to σ70-dependent transcription). We show that specificity and functional overlap between the recombined response regulator and its parent gave rise to new amino acid changes in the recombined response regulator that may form a new interface for phosphotransfer (Fig. 5A).

In this study, we applied an unsupervised ML strategy to study the protein sequence landscape of the REC protein family. One of the benefits of this strategy, over previously applied strategies (7, 25, 26, 32, 34), is that it enabled us to characterize the relationship between REC protein family members over long evolutionary distances. The alternative approach of inferring the evolution of a large protein family across long phylogenetic distances from a domain tree typically results in long branch distances and uncertainty of branch positions (53, 54). For this reason, it is challenging to build reliable domain trees for large protein families, specifically the REC protein family, which we estimate has between 10 and 100 million members in sequenced bacteria (10). We overcame this challenge by repurposing the ML strategy to find closely related REC domains over a defined evolutionary distance (Fig. 3A), after which we could build a reliable domain tree (Fig. 3B) and apply methods to detect differential rates of evolution (Fig. S5A). One pitfall of the ML strategy is that the points on resultant maps do not maintain their positions for the same data sets between runs; however, the relative position of the points for the same data sets is maintained between runs (29). Further development of this strategy, specifically implementing a bootstrapping method, will enhance the overall reliability of its application to the REC and other protein families. Another limitation of this method is that the interpretation of the results is qualitative—as clustering methods do not perform well against neighbor embedding methods (29). To overcome this limitation, we applied an alternative to clustering using a fitting approach to find closely related sequences of interest (Fig. S4). This fitting approach could be used in the future to distinguish the relationships between orthologous and paralogous response regulators of interest across species of interest. This method may also be applicable to other large protein families but may have additional limitations depending on the specific application.

On the basis of this study, we propose that episodic diversifying selection in the REC protein family occurs when two conditions are met: (i) communication between distinct signaling systems (due to detection of the same signals or due to overlap in specificity in downstream interactions) and (ii) the outcome of this communication that negatively impacts cellular fitness. Results from this study and others (15, 24, 50–52, 55) suggest that communication between distinct two-component systems that arise after gene duplication may not negatively impact cellular fitness (Fig. 5B). In this study, we found that the parent response regulator duplicated and diverged a second and third time after the selection of the recombined regulator (Fig. 3B; Fig. S5). Although these new regulators presumably communicated with their parent's signaling system as they sampled sequences in their trajectory to new specificity, we determined that these changes were not under episodic diversifying selection (Fig. 3B). As the recombined regulator probably had comparable overlapping communication with its parent regulator to these other duplicated response regulators, why did the recombined regulators and regulators that duplicated and diverged after the recombination show different rates of evolution? We propose that after the within-gene recombination changed the effector domain from its parent regulator's σ54-dependent transcription to σ70-dependent transcription, the recombined and parent response regulators responded to the same signals under the distinct regulatory contexts of σ54 and σ70 (40–43). Although we cannot rule out the possibility that episodic diversifying selection occurred for a different reason, a simple explanation for the evolution of the recombined response regulator is that differences in regulation by the sigma factors negatively impacted cellular fitness, giving rise to the episodic amino acid changes in the recombined regulator (Fig. 5A).

In our model for episodic diversifying selection of a test case in the REC protein family, duplication followed by recombination led to an overlap in signals and outputs under competing regulatory contexts, giving rise to new residues that may facilitate a new interaction interface. We propose that if conditions of functional discordance are met, episodic diversifying selection can become a mechanism of evolution for large protein family expansion. We further propose that a within-gene recombination can facilitate such conditions. In comparison to expansion in protein families where evolution is directly influenced by an immediate requirement—for example, dynamic environmental factors that need response in immune systems proteins (e.g., toxin-antitoxin [56] and Ig [57]) or small molecule sensing protein families (e.g., G protein-coupled receptors [GPCRs]/olfactory receptors [58, 59] and efflux pumps [60])—the REC protein family is a midpoint in signaling cascades, and immediate fitness or “arms race” factors may not influence its expansion. However, our study indicates that even for this class of large protein families, positive selection occurs, not by changing environmental conditions—as we might have predicted—but instead due to genetic perturbations like recombination events. It should be noted that although recombination can be a source of functional discordance, as it is in this study’s test case, recombination can be neutral and will not always result in functional discordance. Further exploration of the REC sequence landscape or the sequence landscape of other protein families may elucidate whether other genetic perturbations create conditions for diversifying selection when purifying constraints remain intact. It will also be interesting to understand whether there are limitations on protein family expansion that occurs by neutral drift, even when purifying constraints are lifted. Overall, this study shows that episodic diversifying selection is a mechanism in the expansion of the REC protein family and that it occurs by functional discordance. We believe this result is significant in understanding the evolution of the REC protein family, which directly modulates a vast range of cellular functions. This knowledge provides important context for assessing their role in microbial function and microbial communities and in engineering efforts in biosystem design and synthetic biology applications.

MATERIALS AND METHODS

Two-dimensional visualization of protein family sequences

Sampling genomes and creating REC domain databases

Databases of REC domains (PF00072) that were used to make all REC sequence landscapes were sampled from the Microbial Signal Transduction Database (MISTDB) (10)—a curated database of the signal transduction genes from over 200,000 bacterial genomes—using their API and custom functions written in Python. To control for bias driven by genome availability for species of various taxonomic ranks (e.g., Gammaproteobacteria more represented than any other class in the Proteobacteria phylum), we took a subset of the available data by randomly sampling species based on taxonomic rank (e.g., the taxonomic rank below kingdom is phylum; the maximum of species from each bacterial phylum was randomly sampled to generate the bacteria data set; Table S1). The following data sets were forced to include *P. putida* KT2440, which we used to track the parent and recombined regulator in the two-dimensional projections of the REC domains: Bacteria, Proteobacteria, *Pseudomonas*, and Alphaproteobacteria in Fig. 3A. Again, using the MISTDB API, we used a custom function to generate a second database of all of the amino acid sequences of two-component system genes from the sampled species, while keeping a record of each protein’s metadata (e.g., fused domains). We then sampled this database to find all response regulators and their respective sequences to generate Fasta files that include entries for each response regulator in the database.

REC domain alignments

We aligned the REC domain entries in the generated Fasta files from the REC domain sampling functions using hmmalign (30, 61) to the REC pfam consensus motif, PF00072 (7). Using a custom Python function, we linked the aligned domain back to its metadata in the REC domain database. When gap replacement was applied, gaps were replaced by a randomly generated amino acid. To generate scrambled data sets, we scrambled the individual sequences using a custom function that randomly samples residues and rejoins them in a new order while maintaining the original length of the alignment.

Projection of REC domain alignment variability in two-dimensional space

We then encoded the sequences into a discrete variable vector representation (62), by encoding each amino acid into a 21-dimensional zero vector (one dimension for each possible amino acid and a gap character), except for the index representing the amino acid, which is set to 1. Each Lx21-dimensional matrix, where L is the length of the aligned protein sequence, is flattened to a Lx21-length vector. Each vector representing the alignment of a single protein sequence is stacked into a Nx(Lx21) matrix, where N is the number of independent protein sequences, and is then dimensionally reduced to n dimensions (or n-components) to explain 95% of the variation by principal component analysis before being reduced to Nx2 dimensions using the t-distributed stochastic neighbor embedding (29). We approximated the perplexity or the expected size of the clusters, as the number of clades in the taxonomic rank below the highest taxonomic rank queried (e.g., if we were looking at a kingdom rank, the next clade down is the phyla). The perplexity is then equal to the number of unique phyla in the data set—see Table S1 (unique taxa in rank below). Note that we found that a binary encoding was sufficient for our use case, but another encoding may be more suitable for other large protein families. Source code and data are available at https://github.com/mgarber21/Large_Protein_Families.git.

Isolating the parent and recombined regulators from randomly sampled REC domains in Alphaproteobacteria, Gammaproteobacteria, or Proteobacteria species

Using *P. putida* KT2440 to track the parent and recombined regulator in the two-dimensional projections of the REC domains in Alphaproteobacteria, Gammaproteobacteria, or Proteobacteria species, we defined boundaries in the projection and to isolate the candidate proteins. Using the amino acid sequences and reverse translated DNA sequences (63), we aligned the candidate domains (isolating the REC domain from the protein or protein coding sequence) using the MAFFT-LINSI algorithm from MAFFT v7.310 (64). We then used IQTREE (65) to build the Alphaproteobacteria REC domain tree, using the aligned reverse translated DNA sequences of the REC domains with a transversion model, empirical base frequencies, and discrete Gamma model (TVM + F + I + G4), also using outgroup species, GCF_000169415-1-SSE37_RS08100 and 1,000 bootstraps. Using these same sequences, we applied an adaptive branch-site REL test for episodic diversification and detected individual sites subject to diversifying episodic selection using the online aBSREL (66, 67) and MEME (45) tools from HyPhy in Datamonkey (68–70) with default parameters and according to the user instructions. Logs and results can be found in Supplemental Data A online (https://doi.org/10.6084/m9.figshare.c.6854898.v2). To generate amino acid consensus motif images for each subgroup, REC domains fused to either AAA+ or GerE, excluding outgroup AAA+ domains, we binned the aligned domains into their respective subgroups and generated a consensus motif using WebLogo (71).

Fitness experiments using a library of RB-Tn mutants

As described in prior reports (72), the *P. putida* KT2440 RB-TnSeq library, JBEI-1, was thawed, inoculated into 25 mL of Luria broth (LB) supplemented with 50-µg/mL kanamycin, and grown to OD_{600} of 0.5. Three 1-mL samples were taken after this step to serve as t_0 records of barcode abundance. The library was then washed via centrifugation and resuspension in an equal volume of MOPS [3-(N-morpholino)propanesulfonic acid] minimal medium (MM) (Table S4). The washed cells were then diluted 1:50 in MOPS MM with 10 mM L-glutamate serving as the sole carbon source. The library was cultured in a 96-well deep well plate sealed with a gas-permeable membrane (VWR, USA). The plate was shaken (700 rpm) in an INFORS HT Multitron (Infors USA Inc.) at 30°C for 24 hours. Duplicate 600-µL samples were then combined, and BarSeq analysis was conducted as described previously (73–75). Single carbon source fitness data are available at http://fit.genomics.lbl.gov (48).

Automated DNA affinity purification with NGS

DNA preparation for NGS

Pseudomonas isolates were cultured in either LB or MMs (see Table S4 for strain-specific MM recipes). Genomic DNA was purified with a Promega Wizard Genomic Preparation Kit (Promega, Madison, WI). DNA was sheared with Covaris miniTUBE (Covaris, Woburn, MA) to an average size of 200 bp. The DNA quality was confirmed by the Bioanalyzer High-Sensitivity DNA Kit (Agilent, Santa Clara, CA). Sheared DNA was then adapter-ligated (AL) with the NEBnext Ultra ii Library Preparation Kit (New England Biolabs, Ipswich, MA). AL-DNA quality was again confirmed by the Bioanalyzer High-Sensitivity DNA Kit (Agilent, Santa Clara, CA). AL-DNA was stored at −20°C until required for downstream use.

Expression strain design

pet28 expression vectors with N-terminal 6×-His-tagged response regulators (RRs) were cloned by Gibson assembly (76). Plasmid design was facilitated by j5 DNA assembly design (77); see Table S5 for primers.

Automated DNA affinity purification

Quadruplicates of expression strains were grown in autoinduction media (Zyp-5052 [78]) at 37°C, 250 rpm, for 5–6 hours, and then transferred to grow at 17°C, 250 rpm, overnight. Cell pellets were harvested and lysed at 37°C for 1 hour in a lysis buffer—1× Tris-buffered saline (TBS) (diluted from a 10× TBS stock [0.2 M Tris, 1.5 M sodium chloride, pH 7.6]), 100 µM phenylmethylsulfonyl fluoride (PMSF) (Millipore Sigma, Burlington, MA), 2.5 units/mL of benzonase nuclease (Millipore Sigma, Burlington, MA), and 1-mg/mL lysozyme (Millipore Sigma, Burlington, MA). Lysed cells were then clarified by centrifugation at 3,214 × *g* and further filtered in 96-well filter plates by centrifugation at 1,800 × *g*. To enable high-throughput processing, protein-DNA purification steps were performed with immobilized metal affinity chromatography (IMAC) resin pipette tips (PhyNexus, San Jose, CA) using a custom automated platform with the Biomek FX liquid handler (Beckman Coulter, Indianapolis, IN). The expressed RRs were individually bound to metal affinity resin embedded within the IMAC resin pipette tips and washed in a wash buffer (1× TBS, 10 mM imidazole, and 0.1% [vol/vol] Tween 20). The bead-bound RRs were then mixed with 60 µL of DNA binding buffer (1× TBS, 10 mM magnesium chloride, and 0.4-ng/µL AL-DNA, with or without 50 mM acetyl phosphate [split into duplicates]). The protein bound to its target DNA was then enriched in an enrichment buffer (1× TBS, 10 mM imidazole, and 0.1% [vol/vol] Tween 20) and eluted in an elution buffer (1× TBS and 180 mM imidazole). The elution was stored at −20°C for a minimum of 1 day and up to a week before proceeding to the next-generation sequencing (NGS) library generation. See the Supplemental Methods for detailed protocol.

NGS library generation

A 3.2-µL elution from the previous step was added to 3.5-µL SYBR green SsoAdvanced (Bio-Rad, Hercules, CA) and 0.15 µL of each dual-indexed NGS primer. NGS libraries were prepared by following the protocols for fluorescent amplification of NGS libraries (79). Pooled libraries were sequenced by Illumina NovaSeq 6000 SP (100 cycles) (Illumina, San Diego, CA).

DAP-seq data analysis

Sequenced reads were processed by a computational DNA affinity purification-sequencing (DAP-seq) analysis pipeline as follows. Adapters and low-quality bases were trimmed, and reads shorter than 30 bp were filtered out using Trimmomatic v.0.36 (80). The resulting reads were checked for contamination using FOCUS (81). Then, the reads were aligned to the corresponding *Pseudomonas* spp. genome using Bowtie v1.1.2 (82) with –m 1 parameter (report reads with single alignment only). The resulting SAM files were converted to BAM format and sorted using SAMtools v 0.1.19 (83). Peak calling was performed using SPP 1.16.0 (84) with a false discovery rate threshold of 0.01 and a maximum likelihood enrichment (MLE) ratio threshold of 4.0. Enriched motifs were discovered in genome fragments corresponding to the peaks using MEME (85) with parameters –mod anr –minw 12 –maxw 30 –revcomp –pal –nmotifs 1. The source code of the DAP-seq analysis pipeline is available at https://github.com/novichkov-lab/dap-seq-utils.

For conserved RRs with small numbers of high-confidence peaks (one to two per genome), binding motifs were predicted manually by a comparative genomic approach. Orthologous RRs were identified by OrthoFinder2 (86). For each of the orthologous RRs, one genome fragment corresponding to the peak with the highest enrichment value was selected for motif search. Conserved motifs were discovered using the SignalX tool from the GenomeExplorer package (87) with the “inverted repeat” option.

Covariance analysis

Cognate RRs and histidine kinases (HKs) from *Pseudomonas* and *E. coli* strains were identified as pairs if they were found neighboring each other in their respective genomes. Dimerization and histidine phosphotransfer (DHp) (HisKA), catalytic (CA) (HATPase_C), and REC (Response_reg) domain boundaries were determined with hmmsearch from HMMER v3.1b2 (88). Fasta files of concatenated DHp-CA-REC domains from cognate and randomized HK-RR pairs were aligned with the MAFFT-LINSI algorithm from MAFFT v7.310 (64). Alignment files were then queried for coevolution with the ProDy Evol suite (89, 90) in Python and were plotted in a heatmap. The highest scoring residues > 1.1 were used to inform hypotheses for specificity switch in AlphaFold structural predictions.

GFP reporter strain generation and assays

Knockout strain generation

A total of 1,000-bp homology fragments upstream and downstream of the target gene were cloned into plasmid pKS18. Plasmids were then transformed into *E. coli* S17 and then mated into *P. putida* via conjugation. Transconjugants were selected for LB agar plates supplemented with 30-mg/mL kanamycin and 30-mg/mL chloramphenicol. Transconjugants were then grown overnight on LB medium and were then plated on LB agar with no NaCl that was supplemented with 10% (wt/vol) sucrose. Putative deletions were screened on LB agar with no NaCl supplemented with 10% (wt/vol) sucrose and LB agar plate with kanamycin. Colonies that grew in the presence of sucrose but had no resistance to kanamycin were further tested via PCR with primers flanking the target gene to confirm gene deletion.

GFP reporter strains

Promoter boundaries for p2453, p1400, and p3553 were selected as the region just upstream of the gene’s start codon up until the start or stop codon of the next nearest gene. The promoters were cloned upstream of the gene-encoding sfGFP on a broad host range plasmid with BBR1 origin and kanamycin resistance, with Gibson cloning (76). Primers in Table S5. The plasmids were transformed into *P. putida* KT2440 or *P. putida* KT2440 mutant strains by electroporation. Three biological replicates of each strain were cultured in LB and stored with 25% (vol/vol) glycerol at −80°C. Complementation plasmids (GFP reporter plasmids with full-length RR driven by pBAD promoter and constitutively expressed AraC) were combinatorially built using Golden Gate cloning (91) and j5 DNA assembly design (77) (diva.jbei.org), primers in Table S5. The plasmids were transformed into gene-knockout strains of *P. putida* KT2440 by electroporation. Three to six biological replicates of each strain were cultured in LB and stored with 25% (vol/vol) glycerol at −80°C.

Covariant and point mutants

Gene blocks (TWIST Biosciences, San Francisco, CA) of REC domains (Table S5) with covarying mutations (co-variation score > 1.1; see Fig. S9) were cloned into the complementation plasmids with Gibson assembly (76). Point mutants were made using the Q5 Site-Directed Mutagenesis Kit (New England Biolabs, Ipswich, MA). The plasmids were transformed into *P. putida* KT2440 or knockout strains of *P. putida* KT2440 by electroporation. Six biological replicates of each strain were cultured in LB and stored with 25% (vol/vol) glycerol at −80°C.

GFP reporter assays

Reporter strains were adapted to M9 MMs (see the Supplemental Methods for strain-specific MM recipes) supplemented with 0.5% (wt/vol) glucose as the sole carbon source in three overnight passages and stored in MM at −80°C in 25% (vol/vol) glycerol. Adapted strains were cultured in MM in +0.5% (wt/vol) glucose and passaged to MM in +0.5% (wt/vol) glucose with or without a second carbon source (40 mM glutamic acid, 40 mM α-ketoglutaric acid, or 20 mM butyric acid, unless otherwise specified). After 24 hours of growth, samples were diluted 1:100 in 1× PBS, and fluorescence was measured by flow cytometry on the BD Accuri C6 (BD Biosciences, San Jose, CA) configured to detect GFP fluorescence with either fluorescence channel 1 (FL1-A) or fluorescence channel 3 (FL3-A) channels (as indicated in the figures). To remove noise, single-cell measurements with a forward scatter area (FSC-A) greater than a corresponding forward scatter height (FSC-H) by more than 1,000 and an FL1-A or FL3-A less than 10 were removed for further analysis. Median fluorescence intensity (MFI) (median FL1-A or FL3-A) was calculated by determining the median fluorescence of the single-cell measurements. Fold change MFI was calculated by dividing the MFI of the treated sample by the average MFI of the untreated replicates.

AlphaFold predictions of wild-type and mutant response regulators

Full-length amino acid sequences of wild-type PP_1066 and the respective specificity switching mutants of PP_3551 and PP_1066 were queried using ColabFold by AlphaFold (49) using the default parameters. PDB files and prediction logs can be found in Supplemental Data B online (https://doi.org/10.6084/m9.figshare.24205953.v2). The structural prediction of PP_3551 was retrieved from the AlphaFold DeepMind Database (92). The PDB file from each protein's highest-ranking structure from AlphaFold was then visualized and annotated with Chimera X (93).

ACKNOWLEDGMENTS

We would like to thank the following individuals for their contribution to this body of work: for valuable comments on the manuscript, Adam P. Arkin; for helping with lab tasks, Andrew Lau, Julie Lake, Rodrigo Frogeso, and Joyce Luk; for providing valuable insight through discussion about evolution, fitness, and bacterial transcriptional regulation, Ankita Kothari, Pablo Cruz-Morales, and Mitchell G. Thomson; for providing expertise toward the development of high-throughput cloning, protein purification, and NGS techniques, Nurgul Kaplan, Jennifer Chiniquy, Joel M. Guenther, and Brett Garabedian; and for providing us with capabilities to perform high-throughput fitness assays, Mitchell G. Thompson, Morgan Price, and Adam Deutschbauer.

This work was part of the Ecosystems and Networks Integrated with Genes and Molecular Assemblies (ENIGMA) (http://enigma.lbl.gov), a Science Focus Area Program at Lawrence Berkeley National Laboratory (LBNL), and is supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, under contract number DE-AC02-05CH11231 between LBNL and the U.S. Department of Energy. A.A.N. was supported by a National Science Foundation Graduate Research Fellowship, fellow ID 2018253421. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. The U.S. Government retains and the publisher, by accepting the article for publication, acknowledges that the U.S. Government retains a non-exclusive, paid-up, irrevocable, worldwide license to publish or reproduce the published form of this manuscript or allow others to do so, for U.S. Government purposes.

M.E.G., A.E.K., A.A.N., and H.Z. performed the phylogenetic and computational analyses; M.E.G. and L.R. designed and built the expression strains for DAP-seq; M.E.G. developed and conducted the automated pipeline for the DAP-seq experiments; A.E.K. analyzed the DAP-seq data; M.R.I. performed the single-carbon source fitness assays; M.E.G., V.F., and L.E.V. conducted and analyzed in vivo GFP reporter experiments; J.D.K provided support and resources; M.E.G. and A.M. designed the experiments and composed the initial drafts; and A.M. provided resources supervision and support. All authors reviewed and edited the final draft.

AUTHOR AFFILIATIONS

[1]Department of Comparative Biochemistry, University of California, Berkeley, California, USA
[2]Biological Systems and Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[3]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA
[4]Department of Plant and Microbial Biology, University of California, Berkeley, California, USA
[5]Department of Chemical and Biomolecular Engineering, University of California, Berkeley, California, USA
[6]Department of Bioengineering, University of California, Berkeley, California, USA
[7]Center for Biosustainability, Danish Technical University, Lyngby, Denmark
[8]Center for Synthetic Biochemistry, Shenzhen Institutes for Advanced Technologies, Shenzhen, China

AUTHOR ORCIDs

Megan E. Garber http://orcid.org/0000-0003-2886-8808
Hanqiao Zhang http://orcid.org/0000-0001-7394-8781
Aindrila Mukhopadhyay http://orcid.org/0000-0002-6513-7425

FUNDING

Funder	Grant(s)	Author(s)
U.S. Department of Energy (DOE)	2018253421	Megan E. Garber Megan E. Garber Megan E. Garber Megan E. Garber
National Science Foundation (NSF)	2018253421	Alberto A. Nava

AUTHOR CONTRIBUTIONS

Megan E. Garber, Conceptualization, Data curation, Formal analysis, Investigation, Methodology, Software, Validation, Visualization, Writing – original draft, Writing – review and editing | Vered Frank, Data curation, Formal analysis, Methodology, Validation, Writing – review and editing | Alexey E. Kazakov, Data curation, Formal analysis, Methodology, Software, Writing – review and editing | Matthew R. Incha, Data curation, Formal analysis, Methodology, Writing – review and editing | Alberto A. Nava, Data curation, Software, Validation, Writing – review and editing | Hanqiao Zhang, Data curation, Methodology, Software, Writing – review and editing | Luis E. Valencia, Validation, Writing – review and editing | Jay D. Keasling, Resources, Supervision, Writing – review and editing | Lara Rajeev, Conceptualization, Resources, Supervision, Writing – review and editing | Aindrila Mukhopadhyay, Conceptualization, Funding acquisition, Project administration, Resources, Supervision, Writing – review and editing

DATA AVAILABILITY

The source code of the DAP-seq analysis pipeline is available online (https://github.com/novichkov-lab/dap-seq-utils). DAP-seq data have been deposited in NCBI's Gene Expression Omnibus and are accessible through GEO Series accession number GSE157075. Single-carbon source fitness data is available online (http://fit.genomics.lbl.gov). Source code data for REC sequence analysis and ML strategy are available online (https://github.com/mgarber21/Large_Protein_Families). Supplemental Data A (https://doi.org/10.6084/m9.figshare.24205968.v2) and Supplemental Data B are available online (https://doi.org/10.6084/m9.figshare.24205953.v2).

ADDITIONAL FILES

The following material is available online.

Supplemental Material

Data S1 (mBio02622-23-s0001.xlsx). Results output from motif calling with orthologous RRs. Each tab contains data for each ortholog.

Data S2 (mBio02622-23-s0002.xlsx). Raw output from DAP-seq peak calling.

Supplemental Materials (mBio02622-23-s0003.pdf). Supplemental figures, tables, and methods.

Table S5 (mBio02622-23-s0004.xlsx). List of primers used in this study.

REFERENCES

1. Buljan M, Bateman A. 2009. The evolution of protein domain families. Biochem Soc Trans 37:751–755. https://doi.org/10.1042/BST0370751
2. El-Gebali S, Mistry J, Bateman A, Eddy SR, Luciani A, Potter SC, Qureshi M, Richardson LJ, Salazar GA, Smart A, Sonnhammer ELL, Hirsh L, Paladin L, Piovesan D, Tosatto SCE, Finn RD. 2019. The Pfam protein families database in 2019. Nucleic Acids Res 47:D427–D432. https://doi.org/10.1093/nar/gky995
3. Heger A, Holm L. 2003. Exhaustive enumeration of protein domain families. J Mol Biol 328:749–767. https://doi.org/10.1016/s0022-2836(03)00269-9
4. Finn RD, Attwood TK, Babbitt PC, Bateman A, Bork P, Bridge AJ, Chang H-Y, Dosztányi Z, El-Gebali S, Fraser M, et al. 2017. InterPro in 2017-beyond protein family and domain annotations. Nucleic Acids Res 45:D190–D199. https://doi.org/10.1093/nar/gkw1107
5. Hirakawa H, Kurushima J, Hashimoto Y, Tomita H. 2020. Progress overview of bacterial two-component regulatory systems as potential targets for antimicrobial chemotherapy. Antibiotics (Basel) 9:635. https://doi.org/10.3390/antibiotics9100635
6. Capra EJ, Laub MT. 2012. Evolution of two-component signal transduction systems. Annu Rev Microbiol 66:325–347. https://doi.org/10.1146/annurev-micro-092611-150039
7. Pao GM, Saier MH. 1995. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J Mol Evol 40:136–154. https://doi.org/10.1007/BF00167109
8. Galperin MY. 2018. What bacteria want. Environ Microbiol 20:4221–4229. https://doi.org/10.1111/1462-2920.14398
9. Galperin MY. 2006. Structural classification of bacterial response regulators: diversity of output domains and domain combinations. J Bacteriol 188:4169–4182. https://doi.org/10.1128/JB.01887-05
10. Gumerov VM, Ortega DR, Adebali O, Ulrich LE, Zhulin IB. 2020. MiST 3.0: an updated microbial signal transduction database with an emphasis on chemosensory systems. Nucleic Acids Res 48:D459–D464. https://doi.org/10.1093/nar/gkz988
11. Tiwari S, Jamal SB, Hassan SS, Carvalho P, Almeida S, Barh D, Ghosh P, Silva A, Castro TLP, Azevedo V. 2017. Two-component signal transduction systems of pathogenic bacteria as targets for antimicrobial therapy: an overview. Front Microbiol 8:1878. https://doi.org/10.3389/fmicb.2017.01878
12. Wang BX, Cady KC, Oyarce GC, Ribbeck K, Laub MT, Elkins CA. 2021. Two-component signaling systems regulate diverse virulence-associated traits in *Pseudomonas aeruginosa*. Appl Environ Microbiol 87:e03089-20. https://doi.org/10.1128/AEM.03089-20

13. Volke DC, Turlin J, Mol V, Nikel PI. 2020. Physical decoupling of XylS/Pm regulatory elements and conditional proteolysis enable precise control of gene expression in *Pseudomonas Putida*. Microb Biotechnol 13:222–232. https://doi.org/10.1111/1751-7915.13383
14. Schmidl SR, Ekness F, Sofjan K, Daeffler KN-M, Brink KR, Landry BP, Gerhardt KP, Dyulgyarov N, Sheth RU, Tabor JJ. 2019. Rewiring bacterial two-component systems by modular DNA-binding domain swapping. Nat Chem Biol 15:690–698. https://doi.org/10.1038/s41589-019-0286-6
15. McClune CJ, Alvarez-Buylla A, Voigt CA, Laub MT. 2019. Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. Nature 574:702–706. https://doi.org/10.1038/s41586-019-1639-8
16. Stock AM, Robinson VL, Goudreau PN. 2000. Two-component signal transduction. Annu Rev Biochem 69:183–215. https://doi.org/10.1146/annurev.biochem.69.1.183
17. Laub MT, Goulian M. 2007. Specificity in two-component signal transduction pathways. Annu Rev Genet 41:121–145. https://doi.org/10.1146/annurev.genet.41.042007.170548
18. Papon N, Stock AM. 2019. Two-component systems. Curr Biol 29:R724–R725. https://doi.org/10.1016/j.cub.2019.06.010
19. Varughese KI, Zhou XZ, Whiteley JM, Hoch JA. 1998. Formation of a novel four-helix bundle and molecular recognition sites by dimerization of a response regulator phosphotransferase. Mol Cell 2:485–493. https://doi.org/10.1016/s1097-2765(00)80148-3
20. Casino P, Rubio V, Marina A. 2009. Structural insight into partner specificity and phosphoryl transfer in two-component signal transduction. Cell 139:325–336. https://doi.org/10.1016/j.cell.2009.08.032
21. Rocha EPC, Kumar S. 2018. Neutral theory, microbial practice: challenges in bacterial population genetics. Mol Biol Evol 35:1338–1347. https://doi.org/10.1093/molbev/msy078
22. Tomoko O. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. J Mol Evol 40:56–63. https://doi.org/10.1007/BF00166595
23. Capra EJ, Perchuk BS, Skerker JM, Laub MT. 2012. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. Cell 150:222–232. https://doi.org/10.1016/j.cell.2012.05.033
24. Capra EJ, Perchuk BS, Lubin EA, Ashenberg O, Skerker JM, Laub MT. 2010. Systematic dissection and trajectory-scanning mutagenesis of the molecular interface that ensures specificity of two-component signaling pathways. PLoS Genet 6:e1001220. https://doi.org/10.1371/journal.pgen.1001220
25. Chen Y-T, Chang HY, Lu CL, Peng H-L. 2004. Evolutionary analysis of the two-component systems in *Pseudomonas aeruginosa* PAO1. J Mol Evol 59:725–737. https://doi.org/10.1007/s00239-004-2663-2
26. Ashby MK, Houmard J. 2006. Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution. Microbiol Mol Biol Rev 70:472–509. https://doi.org/10.1128/MMBR.00046-05
27. Galperin MY. 2005. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol 5:35. https://doi.org/10.1186/1471-2180-5-35
28. Galperin MY. 2010. Diversity of structure and function of response regulator output domains. Curr Opin Microbiol 13:150–159. https://doi.org/10.1016/j.mib.2010.01.005
29. van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. J Mach Learn Res 8:2579–2605.
30. Zhang Z, Wood WI. 2003. A profile hidden Markov model for signal peptides generated by HMMER. Bioinformatics 19:307–308. https://doi.org/10.1093/bioinformatics/19.2.307
31. Garber ME, Rajeev L, Kazakov AE, Trinh J, Masuno D, Thompson MG, Kaplan N, Luk J, Novichkov PS, Mukhopadhyay A. 2018. Multiple signaling systems target a core set of transition metal homeostasis genes using similar binding motifs. Mol Microbiol 107:704–717. https://doi.org/10.1111/mmi.13909
32. Forslund SK, Kaduk M, Sonnhammer ELL. 2019. Evolution of protein domain architectures. Methods Mol Biol 1910:469–504. https://doi.org/10.1007/978-1-4939-9074-0_15
33. Björklund AK, Ekman D, Light S, Frey-Skött J, Elofsson A. 2005. Domain rearrangements in protein evolution. J Mol Biol 353:911–923. https://doi.org/10.1016/j.jmb.2005.08.067
34. Forslund K, Henricson A, Hollich V, Sonnhammer ELL. 2008. Domain tree-based analysis of protein architecture evolution. Mol Biol Evol 25:254–264. https://doi.org/10.1093/molbev/msm254
35. Ohta T. 1991. Role of diversifying selection and gene conversion in evolution of major histocompatibility complex loci. Proc Natl Acad Sci U S A 88:6716–6720. https://doi.org/10.1073/pnas.88.15.6716
36. Chan CX, Darling AE, Beiko RG, Ragan MA. 2009. Are protein domains modules of lateral genetic transfer? PLoS One 4:e4524. https://doi.org/10.1371/journal.pone.0004524
37. Chan CX, Beiko RG, Darling AE, Ragan MA. 2009. Lateral transfer of genes and gene fragments in prokaryotes. Genome Biol Evol 1:429–438. https://doi.org/10.1093/gbe/evp044
38. Qian W, Han Z-J, He C. 2008. Two-component signal transduction systems of *Xanthomonas* spp.: a lesson from genomics. Mol Plant Microbe Interact 21:151–161. https://doi.org/10.1094/MPMI-21-2-0151
39. Stephenson K, Hoch JA. 2002. Evolution of signalling in the sporulation phosphorelay. Mol Microbiol 46:297–304. https://doi.org/10.1046/j.1365-2958.2002.03186.x
40. Casas-Pastor D, Müller RR, Jaenicke S, Brinkrolf K, Becker A, Buttner MJ, Gross CA, Mascher T, Goesmann A, Fritz G. 2021. Expansion and re-classification of the extracytoplasmic function (ECF) σ factor family. Nucleic Acids Res 49:986–1005. https://doi.org/10.1093/nar/gkaa1229
41. Cases I, Ussery DW, de Lorenzo V. 2003. The σ^{54} regulon (sigmulon) of *Pseudomonas putida*. Environ Microbiol 5:1281–1293. https://doi.org/10.1111/j.1462-2920.2003.00528.x
42. Potvin E, Sanschagrin F, Levesque RC. 2008. Sigma factors in *Pseudomonas aeruginosa*. FEMS Microbiol Rev 32:38–55. https://doi.org/10.1111/j.1574-6976.2007.00092.x
43. Ronneau S, Hallez R. 2019. Make and break the alarmone: regulation of (p)ppGpp synthetase/hydrolase enzymes in bacteria. FEMS Microbiol Rev 43:389–400. https://doi.org/10.1093/femsre/fuz009
44. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. Nat Microbiol 1:16048. https://doi.org/10.1038/nmicrobiol.2016.48
45. Murrell B, Wertheim JO, Moola S, Weighill T, Scheffler K, Kosakovsky Pond SL. 2012. Detecting individual sites subject to episodic diversifying selection. PLoS Genet 8:e1002764. https://doi.org/10.1371/journal.pgen.1002764
46. Loeschcke A, Thies S. 2015. *Pseudomonas putida*—a versatile host for the production of natural products. Appl Microbiol Biotechnol 99:6197–6214. https://doi.org/10.1007/s00253-015-6745-4
47. Sharma P, Bano A, Singh SP, Sharma S, Xia C, Nadda AK, Lam SS, Tong YW. 2022. Engineered microbes as effective tools for the remediation of polyaromatic aromatic hydrocarbons and heavy metals. Chemosphere 306:135538. https://doi.org/10.1016/j.chemosphere.2022.135538
48. Thompson MG, Incha MR, Pearson AN, Schmidt M, Sharpless WA, Eiben CB, Cruz-Morales P, Blake-Hedges JM, Liu Y, Adams CA, Haushalter RW, Krishna RN, Lichtner P, Blank LM, Mukhopadhyay A, Deutschbauer AM, Shih PM, Keasling JD, Zhou N-Y. 2020. Fatty acid and alcohol metabolism in *Pseudomonas putida*: functional analysis using random barcode transposon sequencing. Appl Environ Microbiol 86:e01665-20. https://doi.org/10.1128/AEM.01665-20
49. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. 2021. Highly accurate protein structure prediction with AlphaFold. Nature 596:583–589. https://doi.org/10.1038/s41586-021-03819-2
50. Siryaporn A, Perchuk BS, Laub MT, Goulian M. 2010. Evolving a robust signal transduction pathway from weak cross-talk. Mol Syst Biol 6:452. https://doi.org/10.1038/msb.2010.105
51. Siryaporn A, Goulian M. 2008. Cross-talk suppression between the CpxA-CpxR and EnvZ-OmpR two-component systems in *E. coli*. Mol Microbiol 70:494–506. https://doi.org/10.1111/j.1365-2958.2008.06426.x
52. Groban ES, Clarke EJ, Salis HM, Miller SM, Voigt CA. 2009. Kinetic buffering of cross talk between bacterial two-component sensors. J Mol Biol 390:380–393. https://doi.org/10.1016/j.jmb.2009.05.007
53. Efron B, Halloran E, Holmes S. 1996. Bootstrap confidence levels for phylogenetic trees. Proc Natl Acad Sci U S A 93:13429–13434. https://doi.org/10.1073/pnas.93.23.13429

54. Simon C. 2022. An evolving view of phylogenetic support. Syst Biol 71:921–928. https://doi.org/10.1093/sysbio/syaa068
55. Noriega CE, Lin H-Y, Chen L-L, Williams SB, Stewart V. 2010. Asymmetric cross-regulation between the nitrate-responsive NarX-NarL and NarQ-NarP two-component regulatory systems from *Escherichia coli* K-12. Mol Microbiol 75:394–412. https://doi.org/10.1111/j.1365-2958.2009.06987.x
56. Jurėnas D, Fraikin N, Goormaghtigh F, Van Melderen L. 2022. Biology and evolution of bacterial toxin-antitoxin systems. Nat Rev Microbiol 20:335–350. https://doi.org/10.1038/s41579-021-00661-1
57. Gaebler C, Wang Z, Lorenzi JCC, Muecksch F, Finkin S, Tokuyama M, Cho A, Jankovic M, Schaefer-Babajew D, Oliveira TY, et al. 2021. Evolution of antibody immunity to SARS-CoV-2. Nature 591:639–644. https://doi.org/10.1038/s41586-021-03207-w
58. Vandewege MW, Mangum SF, Gabaldón T, Castoe TA, Ray DA, Hoffmann FG. 2016. Contrasting patterns of evolutionary diversification in the olfactory repertoires of reptile and bird genomes. Genome Biol Evol 8:evw013. https://doi.org/10.1093/gbe/evw013
59. Bargmann CI. 2006. Comparative chemosensation from receptors to ecology. Nature 444:295–301. https://doi.org/10.1038/nature05402
60. Henderson PJF, Maher C, Elbourne LDH, Eijkelkamp BA, Paulsen IT, Hassan KA. 2021. Physiological functions of bacterial "multidrug" efflux pumps. Chem Rev 121:5417–5478. https://doi.org/10.1021/acs.chemrev.0c01226
61. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:W29–37. https://doi.org/10.1093/nar/gkr367
62. Jing X, Dong Q, Hong D, Lu R. 2020. Amino acid encoding methods for protein sequences: a comprehensive review and assessment. IEEE/ACM Trans Comput Biol Bioinform 17:1918–1931. https://doi.org/10.1109/TCBB.2019.2911677
63. Suyama M, Torrents D, Bork P. 2006. PAL2NAL: robust conversion of protein sequence alignments into the corresponding codon alignments. Nucleic Acids Res 34:W609–12. https://doi.org/10.1093/nar/gkl315
64. Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol Biol Evol 30:772–780. https://doi.org/10.1093/molbev/mst010
65. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol Biol Evol 32:268–274. https://doi.org/10.1093/molbev/msu300
66. Kosakovsky Pond SL, Murrell B, Fourment M, Frost SDW, Delport W, Scheffler K. 2011. A random effects branch-site model for detecting episodic diversifying selection. Mol Biol Evol 28:3033–3043. https://doi.org/10.1093/molbev/msr125
67. Smith MD, Wertheim JO, Weaver S, Murrell B, Scheffler K, Kosakovsky Pond SL. 2015. Less is more: an adaptive branch-site random effects model for efficient detection of episodic diversifying selection. Mol Biol Evol 32:1342–1353. https://doi.org/10.1093/molbev/msv022
68. Weaver S, Shank SD, Spielman SJ, Li M, Muse SV, Kosakovsky Pond SL. 2018. Datamonkey 2.0: a modern web application for characterizing selective and other evolutionary processes. Mol Biol Evol 35:773–777. https://doi.org/10.1093/molbev/msx335
69. Pond SLK, Frost SDW. 2005. Datamonkey: rapid detection of selective pressure on individual sites of codon alignments. Bioinformatics 21:2531–2533. https://doi.org/10.1093/bioinformatics/bti320
70. Delport W, Poon AFY, Frost SDW, Kosakovsky Pond SL. 2010. Datamonkey 2010: a suite of phylogenetic analysis tools for evolutionary biology. Bioinformatics 26:2455–2457. https://doi.org/10.1093/bioinformatics/btq429
71. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence logo generator. Genome Res 14:1188–1190. https://doi.org/10.1101/gr.849004
72. Thompson MG, Blake-Hedges JM, Cruz-Morales P, Barajas JF, Curran SC, Eiben CB, Harris NC, Benites VT, Gin JW, Sharpless WA, Twigg FF, Skyrud W, Krishna RN, Pereira JH, Baidoo EEK, Petzold CJ, Adams PD, Arkin AP, Deutschbauer AM, Keasling JD. 2019. Massively parallel fitness profiling reveals multiple novel enzymes in *Pseudomonas putida* lysine metabolism. mBio 10:e02577-18. https://doi.org/10.1128/mBio.02577-18
73. Rand JM, Pisithkul T, Clark RL, Thiede JM, Mehrer CR, Agnew DE, Campbell CE, Markley AL, Price MN, Ray J, Wetmore KM, Suh Y, Arkin AP, Deutschbauer AM, Amador-Noguez D, Pfleger BF. 2017. A metabolic pathway for catabolizing levulinic acid in bacteria. Nat Microbiol 2:1624–1634. https://doi.org/10.1038/s41564-017-0028-z
74. Incha MR, Thompson MG, Blake-Hedges JM, Liu Y, Pearson AN, Schmidt M, Gin JW, Petzold CJ, Deutschbauer AM, Keasling JD. 2020. Leveraging host metabolism for bisdemethoxycurcumin production in *Pseudomonas putida*. Metab Eng Commun 10:e00119. https://doi.org/10.1016/j.mec.2019.e00119
75. Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, Blow MJ, Bristow J, Butland G, Arkin AP, Deutschbauer A. 2015. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. mBio 6:e00306–15. https://doi.org/10.1128/mBio.00306-15
76. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. 2009. Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat Methods 6:343–345. https://doi.org/10.1038/nmeth.1318
77. Hillson NJ, Rosengarten RD, Keasling JD. 2012. j5 DNA assembly design automation software. ACS Synth Biol 1:14–21. https://doi.org/10.1021/sb2000116
78. Studier FW. 2005. Protein production by auto-induction in high density shaking cultures. Protein Expr Purif 41:207–234. https://doi.org/10.1016/j.pep.2005.01.016
79. Chiniquy J, Garber ME, Mukhopadhyay A, Hillson NJ. 2020. Fluorescent amplification for next generation sequencing (FA-NGS) library preparation. BMC Genomics 21:85. https://doi.org/10.1186/s12864-020-6481-8
80. Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120. https://doi.org/10.1093/bioinformatics/btu170
81. Silva GGZ, Cuevas DA, Dutilh BE, Edwards RA. 2014. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. PeerJ 2:e425. https://doi.org/10.7717/peerj.425
82. Langmead B, Trapnell C, Pop M, Salzberg SL. 2009. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. Genome Biol 10:R25. https://doi.org/10.1186/gb-2009-10-3-r25
83. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, Marth G, Abecasis G, Durbin R, 1000 Genome Project Data Processing Subgroup. 2009. The sequence alignment/map format and SAMtools. Bioinformatics 25:2078–2079. https://doi.org/10.1093/bioinformatics/btp352
84. Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat Biotechnol 26:1351–1359. https://doi.org/10.1038/nbt.1508
85. Bailey TL, Boden M, Buske FA, Frith M, Grant CE, Clementi L, Ren J, Li WW, Noble WS. 2009. MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res 37:W202–8. https://doi.org/10.1093/nar/gkp335
86. Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol 20:238. https://doi.org/10.1186/s13059-019-1832-y
87. Mironov AA, Vinokurova NP, Gelfand MS. 2000. Software for analysis of bacterial genomes. Mol Biol 34:222–231. https://doi.org/10.1007/BF02759643
88. Mistry J, Finn RD, Eddy SR, Bateman A, Punta M. 2013. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res 41:e121. https://doi.org/10.1093/nar/gkt263
89. Bakan A, Meireles LM, Bahar I. 2011. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 27:1575–1577. https://doi.org/10.1093/bioinformatics/btr168
90. Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon TR, Bahar I. 2014. Evol and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics 30:2681–2683. https://doi.org/10.1093/bioinformatics/btu336
91. Engler C, Kandzia R, Marillonnet S. 2008. A one pot, one step, precision cloning method with high throughput capability. PLoS One 3:e3647. https://doi.org/10.1371/journal.pone.0003647
92. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, Yuan D, Stroe O, Wood G, Laydon A, et al. 2022. AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Res 50:D439–D444. https://doi.org/10.1093/nar/gkab1061
93. Pettersen EF, Goddard TD, Huang CC, Meng EC, Couch GS, Croll TI, Morris JH, Ferrin TE. 2021. UCSF ChimeraX: structure visualization for

researchers, educators, and developers. Protein Sci 30:70–82. https://doi.org/10.1002/pro.3943