Evolutionarily Driven Domain Swap Alters Sigma Factor Dependence in Bacterial Signaling System

Megan E. Garber^{1, 2}, Vered Frank², Alexey E. Kazakov³, Hanqiao Zhang^{2, 4}, Matthew R. Incha^{2, 5}, Jay D. Keasling^{2, 4, 6}, Lara Rajeev², Aindrila Mukhopadhyay^{*1, 2, 3}

Author Affiliations:

1. University of California, Berkeley Department of Comparative Biochemistry
2. Lawrence Berkeley National Laboratory, Biological Systems and Engineering Division
3. Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division
4. University of California, Berkeley Department of Bioengineering
5. University of California, Berkeley Department of Plant and Microbial Biology
6. University of California, Berkeley Department of Chemical and Biomolecular Engineering

* Correspondence: amukhopadhyay@lbl.gov

Abstract

Functional diversity in microbes is driven by vertical or horizontal gene transfer. Using modular bacterial signaling systems, we experimentally validate domain swapping of modular genes. We computationally explore the domain architecture of bacterial two-component systems and find a transcriptional effector domain swap that reconstructed a sigma54-dependent system to become sigma70-dependent, all regulating short-chain carboxylic acid consumption. Verifying their ancestral relationship, we coerce the systems into a pseudo-ancestral state. The differences in

regulatory requirements for sigma factors in addition to the characterized system's involvement in carbon utilization imply that this domain swap reshaped the metabolic potential of proteobacteria. This work provides empirical evidence for a general domain swapping hypothesis, which enables bacteria to shuffle parts of genes to achieve neofunctionalization.

Keywords

Domain swap; Two-component system; Response regulator; Bacterial evolution; Coevolution; Specificity; Insulation; Protein-protein interaction; Transcriptional regulation; Sigma factors; Sigma70; Sigma54; Protein-DNA interaction; DAP-seq; Functional genomics

Introduction

New genes can arise in bacteria by vertical gene transfer (VGT) or horizontal gene transfer (HGT). VGT provides incremental heritable changes in the genome. Bacteria bypass slow genome modification by sharing genetic information via HGT. Computational studies interrogating the lineage of full-length genes or entire gene clusters, followed by empirical studies have cemented HGT as a mechanism of evolution in microbes. In this framework, VGT is responsible for developing new genes while HGT circulates new genes to foreign genomes. Disputing this model, computational studies on the lineage of parts of modular genes suggest that novel genes were developed by HGT mediated part sharing, or domain swapping. If true, rapid neofunctionalization of proteins is possible by HGT. While computational evidence for domain swapping is abundant, empirical evidence remains scarce. This study bolsters the domain swapping hypothesis with empirical evidence from a well-studied bacterial signaling system.

Bacteria use two-component systems (TCS) to sense and respond to signals. In a canonical TCS, a histidine kinase (HK) is auto-phosphorylated by ATP upon stimulation, then engages in

phosphotransfer to its cognate response regulator (RR). Although TCSs can be highly redundant in a single bacterial genome^{1,2}, biochemical and genetic evidence suggests that interactions between cognate HK and RR pairs are conserved. Coevolutionary events of duplication and divergence led to orthogonality of cognate pairs³⁻⁵. TCS are modular, input domains in HKs typically encode for sensing functionality, whereas effector domains in RRs confer output function (Figure 1a). The domains of HK-RR pair are typically carried through to descendants during VGT.

Figure 1: Domain architecture of TCS with transcriptional effector domains (TEDs): (A) Canonical TCS have input signals that activate histidine kinases (HK) and response regulators (RR) to coordinate some output response. An expanded view of HKs and RRs reveals their modularity. (B) Relevant RR families designated by their domain architecture of transcriptional effector domains (TEDs) are annotated by their TED, family, and sigma factor dependence.

Hypotheses of domain swapping have emerged as a potential mechanism for the diversification of TCSs. Domain swaps occur via HGT and homologous recombination⁶. Core parts of TCSs (Hpt, CA, REC domains) remain intact, while swaps of the input or sensing parts diversify the signals⁷ or swaps of output or effector domains modify the function of the entire cascade¹.

RR effector domains dictate the final effect of a signaling cascade. RRs can harbor effector domains that lead to chemotactic responses⁸, regulation of small molecules⁹, or transcriptional regulation^{1,2}. Transcriptional effector domains (TEDs) interact primarily with DNA in a sequence-specific manner to regulate transcription. TEDs are diverse in structure, protein-DNA interaction, protein-protein interaction, and function, and are binned into subclasses or families¹. In this study we predominantly observe TEDs associated with RRs that confer activation of genes with sigma54 factors (AAA⁺ tethered to HTH_8, NtrC family), and regulation of genes with sigma70 factors (GerE, NarL family or Trans_reg_C, OmpR family) (Figure 1b).

We take a computational approach to explore the domain architecture of RRs in a select set of pseudomonads to identify evolutionarily driven domain swaps. Computational predictions are then substantiated by original experimental results. We observe a domain swapping event that altered the sigma factor dependence of a cluster of carboxylic acid responsive sigma54-dependent RRs native to proteobacteria.

Results

We identified all RRs in five representative pseudomonads, subdividing them by their output domains, considering only RRs with TEDs (Supplementary Table 1). The phylogenetic tree of the REC domains agrees with a previously published tree of *P. aeruginosa* PAO1 REC domains¹⁰. Consistent with the current understanding how highly redundant proteins evolve in bacteria^{4,11}, our results demonstrate that the REC domains of RRs do not clade by species, but instead clade by TEDs (Figure 2). A domain swap, hypothesized when RRs of one family cluster with RRs of another, is observed within the AAA⁺-HTH_8/NtrC-like cluster (red and green) of the REC phylogenetic tree, where REC domains with an alternative TED, GerE/NarL-like (blue) are present (Figure 2).

Figure 2: Phylogenetic trees of REC (left) and HPT-CA (right) domains of RRs and HKs from select pseudomonads. Trees are annotated with species and transcriptional effector domain (TED) at the leaves. Hybrid HKs on the HPT-CA tree are highlighted in grey, and are notably associated with TCS that have GerE TEDs. The domain swap at the focus of this work is highlighted in black. A close-up of the domain swap cluster for REC (left) and HPT-CA (right) domains trees annotated with protein name and/or TED. Lines are drawn for predicted cognate HK-RR pairs.

To test a coevolutionary hypothesis, we applied the same phylogenetic strategy to sequences of HPT-CA (HisKA-HATPase_C) domains of all HKs in the pseudomonads. We observed that the leaves of the HPT-CA tree predominantly cluster by the domain architecture of

their cognate RRs (Figure 2). Interestingly, cognate HKs for the major clade of NarL-like RRs are hybrid HKs (grey bars), whereas the cognate HKs of the NarL-like RRs within the swapped cluster are not. These results suggest that our initial observation and hypothesis of a functional domain swap was probable.

We next determined the evolutionary distance of the domain swap, asking whether it was present in pseudomonads only, or all proteobacteria. Using representatives from α -, β -, γ -, and δ -proteobacteria, selected for their high proportions of NtrC-like and NarL-like RRs, we recapitulated the results of the *Pseudomonas* REC domain-based phylogenetic tree. The proteobacteria REC phylogenetic tree reveals that the swap is found in the select strains of α -, β -, and γ -proteobacteria, but not observed in the selected δ -proteobacteria (Supplementary Figure 1). This discrepancy could be the result of gene loss in the selected representatives, or could indicate that the swapped clade never existed in δ -proteobacteria. Notably, in the new view of the tree of life¹², δ -proteobacteria are more distantly related to the rest of the proteobacterial clade than previously thought. Together these results imply that a domain swap likely occurred at the cuspis of the proteobacteria (α -, β -, γ -) clade.

Functional similarity between the extra-familial representatives in the swapped cluster would indicate a genuine relationship. Closer observation of the swapped cluster reveals that there are four subclusters, three of which harbor AAA⁺ domains (NtrC-like/sigma54-dependent) and one which harbors a GerE domain (NarL-like/sigma70-dependent) (Figure 2). We focused our efforts on the functions of the TCSs in *P. putida* KT2440, because it is a well-studied model strain with a readily available genetic toolkit to enable validation studies. Because the domain swap occurs within the NtrC-like cluster, we hypothesized that the signaling cascades were involved in nutrient uptake of carbon or nitrogen sources¹³. Additionally, NtrC representatives within the

swapped cluster were previously characterized as amino acid or di-carboxylic acid sensors^{13–16}.
The functions of the NarL-like representatives were unknown.

We applied an automated pipeline of DNA-affinity purification with next generation sequencing (DAP-seq)^{17,18} (Supplementary Figure 2a), to the entire subset of RRs in the swapped cluster. We found highly enriched DNA binding targets for all of the NtrC-like RRs (Supplementary Figure 2b, Supplementary Data 1,2). With alignments of highly enriched targets from NtrC-like RRs, we manually assigned binding motifs to the homologs and queried each genome for additional DNA binding targets (Supplementary Figure 2b, Figure 3a,c, Supplementary Table 2). We were only able to find genomic targets for NarL-like RRs from *P. fluorescens* FW300-N2C3 (Supplementary Figure 2b, Supplementary Data 1,2), therefore we did not have suitable data to identify binding motifs or conserved genetic outputs for all NarL-like RRs.

To determine the signals for the cluster, we utilized randomly barcoded transposon libraries for next generation sequencing of *P. putida* KT2440 grown with single carbon sources. The genes with low fitness scores (fitness < -2) under a given growth condition have condition-specific essentiality. We observed that the gene clusters harboring the RRs of interest each had low fitness scores for growth with either glutamic, succinic, a-ketoglutaric, or butyric acids (Figure 3a). We speculated that genes and their corresponding TCSs with synonymous fitness values represented a signaling cascade where the signal was the carbon source. We proposed that the NtrC-like RRs, PP_1066, PP_0263, and PP_1401 are regulated by glutamic¹⁶, succinic, and a-ketoglutaric acids^{14,15} respectively. We also hypothesized that the NarL-like RR, PP_3551, is regulated by butyric acid. We were not able to identify binding sites or a DNA binding motif for PP_3551 and relied on the gene cluster and fitness data to hypothesize an output for the butyric acid responsive

TCS. The neighboring gene PP_3553 demonstrated a low fitness score consistent with the TCS, making it a good candidate for the signaling cascade's output.

Figure 3: Functional genomics for determining input signal and output genes: (A) Heatmap of fitness data for TCS gene clusters within the domain swapped cluster. A low fitness score (fitness < - 2) indicates a gene has condition-specific essentiality. Predicted promoters are represented by a rectangle below downstream genes, black fill indicates promoters with DAP-seq identified motifs. (B) GFP reporter plasmids in WT and Δ RR backgrounds were grown in minimal media with glucose without (-) or with a second carbon source: glutamic, a-ketoglutaric or butyric acids. Fluorescence was measured by flow cytometry and measurements are reported as GFP Mean $\times 10^3$ after gating. Statistical significance was determined between conditions by t-test (* = p-value < 0.05, n.s. = not significant). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points with black diamonds, outliers; n = 3. (C) Summary of findings from functional genomics studies. The relationship of the *P. putida* TCSs is shown by a pruned phylogenetic tree to the left, and results, RR family, validated signal, signal chemical structure, validated regulated genes, orthogonal binding motif identified by DAP-seq are drawn as a table to the right. If the findings were not validated by reporter assay, the results are displayed in grey.

To validate our predictions for signals and output genes, we generated GFP reporter strains. The *P. putida* KT2440 reporter strains responded to the proposed signals with increased fluorescence above background. In concert with our predictions, when the reporter strains were

stimulated in genetic backgrounds absent of the corresponding transcriptional RR, fluorescence response was ablated (Figure 3b).

Our experimental results show that all signals for RRs in the swapped cluster are short chain carboxylic acids (Figure 3c). We propose that these TCSs regulate either uptake genes for carboxylic acids or metabolic genes involved in the TCA cycle or beta-oxidation. The signal for HK-RR-promoter (PP_3552-PP_3551-pPP_3553) was unknown until this study. We therefore rename the REC paralogs in the extra-familial subcluster the Carboxylic Acid Responsive TCSs, CarSR I, II, III, IV (PP_0263, PP_1066, PP_1401, PP_3551).

We proposed an ancestral model in which common ancestors of CarSR were generalists, able to sense a broad range of carboxylic acids and interact promiscuously with cognate pairs (Figure 4a). To demonstrate the model we coerced the TCSs into a pseudo-ancestral state (Figure 4b,c). We applied co-variance analysis to HPt-CA-REC domains of cognate HK-RR pairs to identify co-varying residues (Supplementary Figure 3) between HPt-CA and REC domains. Consistent with previous studies¹⁹, the first alpha helix of the REC domain, previously shown to share points of contact with HPt domain, contained the highest scoring residues. We applied site-directed mutagenesis switching the highest scoring residues of the REC domain of PP_1066 (hereafter RR_g) and PP_3551 (hereafter RR_b) (Figure 4b).

Figure 4: *Covarying residue swaps alter REC specificity of butyric acid responsive TCS*: (A) Model where hypothetical generalist ancestors of domain swapped cluster evolved to become modern day specialists. (B) Sequence alignments of REC domains of native RRs PP_1066 (RR_g), PP_3551 (RR_b), and mutated RRs PP_1066* (RR_g*), PP_3551* (RR_b*). Residues with high co-variance scores > 1.1 are marked in a grey background and red text for RR_g-derived residues or blue text for RR_b-derived residues. Structural information for the REC domain is shown below the sequences, where arrows are β-barrels and ribbons are α-helices. (C) RR_g, driving glutamic acid responsive promoter (pRR_g) switches its specificity to butyric acid responsive when mutated to RR_g*. RR_b, driving butyric acid responsive promoter (pRR_b) partially switches its specificity to glutamic acid responsive when mutated to RR_b*. Statistical significance was determined between conditions by t-test (* = p-value < 0.05, n.s. = not significant). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points with black diamonds, outliers; n = 6.

Native RRs driving expression of their native promoters were characterized (Supplementary Figure 4) to determine optimal conditions for GFP expression in a heterologous system. RR_g-pRR_g demonstrated promiscuous and leaky behavior, sensing either glutamic or butyric acids while also expressing GFP in the absence of signal. RR_b-pRR_b, on the other hand, showed tight activation by butyric acid only (Figure 4c). The mutant RRs (indicated by *) exhibit behavior similar to their native counterparts. RR_b*-pRR_b, like RR_g-pRR_g, senses both glutamic and butyric acids and shows leaky GFP expression in the absence of signal. Strikingly, RR_g*-pRR_g, similar to RR_b-pRR_b, senses only butyric acid. Taken together, these results show, empirically, that gradual changes in specificity determining residues made over time could be responsible for the specialization of the CarSR TCSs. Furthermore they provide biologically

relevant evidence that the described extra-familial TCSs with different sigma factor dependencies, the CarR RRs, are closely related.

In an alternative hypothesis, the CarR subcluster could be derived from HGT alone, not domain swapping. We queried the configuration of RR families in bacterial phyla with high counts of NarL-like and NtrC-like RRs, and found representative strains from Firmicutes, Bacteroides, Acidobacteria, δ -Proteobacteria, and γ -Proteobacteria (*P. Putida* KT2440) (Supplementary Figure 5a). We identified all RRs and their corresponding REC domains from the representative species, and repeated phylogenetic analysis and domain mapping (Supplementary Figure 5b). We observed a similar configuration to pseudomonads and proteobacteria trees (Figure 2b, Supplementary Figure 1), indicating that the organization of TCS families observed in present day bacteria is as ancient as the bacterial domain of life, and the alternative hypothesis of HGT fails.

Close observation of cognate pairs in the carboxylic acid sensing subcluster, revealed that the cognate HKs of the NtrC-like RRs are transmembrane bound with a d_Cache domain and the cognate HKs to the NarL-like are cytosolic with a PAS domain. While both of these sensing domains directly interact with ligands, they likely did not evolve through VGT. To reconcile the discrepancy, we hypothesize that the HK domain architecture observed in the pseudomonads carboxylic acid sensing subcluster is the result of domain loss, where a common ancestor harbored both a periplasmic d_Cache domain and a cytosolic PAS domain. We anticipate that a closely related TCS might exist in nature without domain loss (Supplementary Figure 6).

Discussion

We provide empirical evidence for a carboxylic acid sensing NtrC-like TCS that underwent a domain swap, altering its sigma factor dependence (Supplementary Figure 7). Our proposed model assumes that an ancestral HK had promiscuous d_Cache and PAS carboxylic acid sensing

domains, and that when the domain swap occurred, the output gene was promiscuous within carboxylic acid metabolism and could moonlight a beneficial function²⁰. The *in vivo* mutational screen provides evidence that insulated, extra-familial RRs within the same species share sequence space and are closely related. In a previous study, an extra-familial specificity swap of an Omp family HK enabled interaction with an Ntr family RR in *C. crescentus*, leading to the conclusion that a duplicated Ntr family TCS overlapped in sequence space with an Omp family TCS²¹. Their results demonstrated how VGT of a duplicated TCS facilitated insulation from unwanted cross-talk, while our work highlights how domain swapping can alter the output TED of an already well-insulated TCS.

NtrC-like RRs require sigma54 transcription factors to activate genes in their regulons, whereas NarL-like RRs interact with sigma70. It is well documented that cellular levels of small nucleotide, (p)ppGpp, in combination with the protein DksA regulate RNA polymerase's (RNAP) preference for sigma70 (extracytoplasmic function family)²² or sigma54^{23,24}. While the environmental conditions needed for pseudomonads to reach the appropriate levels of (p)ppGpp and DksA for sigma54 occupation of RNAP are unknown²⁵, it can be reasoned that the switch is lifestyle dependent. TCSs implicated in this study were found to be metabolite responsive systems, and could play a major role in survival and fitness. We speculate that the capability to consume carboxylic acids without regulatory constraints dictated by sigma factor dependence, may have given a fitness advantage to the ancestral strain.

Domain swapping in general is an important biological phenomenon that can impact all domains of life⁶. HGT is known to mobilize genes or entire gene clusters that can be subsequently integrated into foreign genomes²⁶. Including domain swapping in HGT models implies that bacteria have another mechanism to develop new proteins. While we focused on a single domain

swap, our results (Figure 2, Supplementary Figures 1,5) and results of previous studies suggest that domain swapping in TCS and other modular systems is prevalent. Given the potential ubiquity of domain swapping, we propose that our strategies can be applied to other complex modular proteins and systems.

Native enzymes and proteins provide attractive parts for practical applications^{27,28} and potential answers to plug-and-play biology via synthetic domain swapping. One critical component of synthetic domain swapping is identifying domain boundaries for a seamless swap. A previous study identified domain boundaries between REC domains and TEDs for *E. coli* RRs, engineering only inter-family TED swaps²⁹. Our study demonstrates that evolution synthesized an extra-familial swap in a functional signaling cascade. Evolutionarily driven domain swaps could be an untapped source for identifying domain boundaries for synthetic biology.

This work establishes and validates domain swapping as a mechanism for neofunctionalization of modular genes. Our results confirm the close relationship between systems that would otherwise be described as distantly related. Ultimately, our work provides evidence for an extension to the current model of neofunctionalization to include functional part sharing via domain swapping.

Methods

Phylogenetic analysis of REC and HPT-CA trees

RRs and HKs were identified with hmmsearch from HMMER v3.1b2³⁰ to search for all proteins with Response_reg (PF00072) and HisKA (PF00512) domains. Alternatively, HKs and RRs were identified with the Microbial Signal Transduction database (MISTDB)³¹. Using HMMER again, we queried the domain architecture of the signaling proteins and selected for RRs with HTH_8,

Trans_reg_C, AAA⁺, GerE, HTH_18 (AraC) domains. REC and HPt-CA domain sequences were extracted from whole sequences based on coordinates determined by the hmmsearch. Domains were aligned using the MAFFT-LINSI algorithm from MAFFT v7.310³². Phylogenetic trees were constructed using FastTree 2³³. Trees were visualized and annotated using python based ETE3³⁴.

Automated DNA affinity purification - seq

DNA preparation for NGS

Pseudomonas isolates were cultured in either LB or minimal media (see Supplementary Table 3 for strain specific minimal media 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 Bioanalyzer high sensitivity DNA kit (Agilent, Santa Clara). Sheared DNA was then adapter-ligated (AL) with NEBnext Ultra ii Library Preparation kit (New England Biolabs, Ipswich, MA). AL-DNA quality was again confirmed by Bioanalyzer high sensitivity DNA kit (Agilent, Santa Clara). AL-DNA was stored at -20°C until required for downstream use.

Expression Strain Design

Pet28 expression vectors with N-terminal 6x-His-tagged RRs were cloned by Gibson assembly³⁵. Plasmid design was facilitated by j5 DNA assembly design³⁶ (diva.jbei.org), see Supplementary Table 4 for primers.

Automated DNA affinity purification

Quadruplicates of expression strains were grown in autoinduction media (Zyp-5052³⁷) 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 (1X TBS, 100 µM PMSF (Millipore Sigma, Burlington MA), 2.5 units/mL Benzonase nuclease (Millipore Sigma, Burlington MA), 1 mg/mL Lysozyme (Millipore Sigma, Burlington MA)). Lysed cells were then clarified by centrifugation at 3214 x g and further filtered in 96-well filter plates by centrifugation at 1800 x g. To enable high-throughput processing, protein-DNA purification steps were performed with 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 (1X TBS, 10 mM Imidazole, 0.1% Tween 20). The bead bound RRs were then mixed with 60 µL of DNA binding buffer (1X TBS, 10 mM magnesium chloride, 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 (1X TBS, 10 mM Imidazole, 0.1% Tween 20) and eluted in an elution buffer (1X TBS, 180 mM Imidazole). The elution was stored at -20°C for a minimum of one day and up to a week before proceeding to the NGS library generation. See supplementary methods for detailed protocol.

NGS Library Generation

3.2 µL of the elution from the previous step was added to 3.5 µL SYBR green ssoAdvanced (Biorad, Hercules, CA) and 0.15 µL of each dual indexed NGS primers. NGS libraries were

prepared by following the protocols for fluorescent amplification of NGS libraries³⁸. 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 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.0.36³⁹. The resulting reads were checked for contamination using FOCUS⁴⁰. Then the reads were aligned to the corresponding *Pseudomonas* spp. genome using Bowtie v1.1.2⁴¹ with `-m 1` parameter (report reads with single alignment only). Resulting SAM files were converted to BAM format and sorted using samtools v 0.1.19⁴². Peak calling was performed using SPP 1.16.0⁴³ with false discovery rate threshold of 0.01 and MLE enrichment ratio threshold of 4.0. Enriched motifs were discovered in genome fragments corresponding to the peaks using MEME⁴⁴ with parameters `-mod anr -minw 12 -maxw 30 -revcomp -pal -nmotifs 1`. 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 (1-2 per genome), binding motifs were predicted manually by comparative genomics approach. Orthologous RRs were identified by OrthoFinder²⁴⁵. For each of 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 GenomeExplorer package⁴⁶ with the “inverted repeat” option.

Fitness Experiments

The *P. putida* KT2440 RB-TnSeq library, JBEI-1⁴⁷, was thawed, inoculated into 25 mL of LB medium supplemented with kanamycin, and grown to OD₆₀₀ of 0.5. Three 1 mL samples were taken after this step to serve as t₀ records of barcode abundance. The library was then washed via centrifugation and resuspension in an equal volume of MOPS minimal medium. The washed cells were then diluted 1:50 in MOPS minimal medium 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 hrs. Duplicate 600 µL samples were then combined and BarSeq analysis was conducted as described previously^{48–50}. Single carbon source fitness data is available at <http://fit.genomics.lbl.gov>⁵¹.

Co-variance analysis

Cognate RRs and HKs from *Pseudomonas* and *E. coli* strains were identified as pairs if they were found neighboring each other in their respective genomes. HPt (HisKA), CA (HATPase_C), and REC (Response_reg) domain boundaries were determined with hmmsearch from HMMER v3.1b2³⁰. Fasta files of concatenated HPt-CA-REC domains from cognate and randomized HK-RR pairs were aligned with the MAFFT-LINSI algorithm from MAFFT v7.310³². Alignment files were then queried for coevolution with the ProDy *Evol* suite^{52,53} in python and were plotted in a heatmap. The highest scoring residues > 1.1 were used to inform hypotheses for specificity switch strains.

GFP reporter strain generation and assays

Knockout strain generation

1000bp 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 on LB agar plates supplemented with 30 mg/ml kanamycin and 30 mg/ml chloramphenicol. Transconjugants were then grown overnight on LB media 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, p3553 were identified 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 GFP on a broad host range plasmid with BBR1 origin and Kanamycin resistance with Gibson cloning³⁵, primers in Supplementary Table 4. 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 leveraging Golden Gate cloning⁵⁴ and j5 DNA assembly design³⁶ (diva.jbei.org), primers in Supplementary Table 4. The plasmids were transformed into knockout strains of *P. putida* KT2440 by electroporation. 3-6

biological replicates of each strain were cultured in LB and stored with 25 % (v/v) glycerol at -80 °C.

RRs with switched specificity

Gene blocks (TWIST Biosciences, San Francisco, CA) of REC domains (Supplementary Table 3) with co-varying mutations (co-variation score > 1.1) were cloned into the complementation plasmids with Gibson assembly³⁵. 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 minimal media (MM) (see supplemental methods for strain specific minimal media recipes) supplemented with 0.5 % (wt/vol) glucose as the sole carbon source in 3 overnight passages, and stored in MM at -80 °C in 25 % (vol/vol) glycerol. Adapted strains were cultured in MM + 0.5 % (wt/vol) glucose and passaged to MM + 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 1X PBS, and fluorescence was measured by flow cytometry on the BD Accuri C6 (BD Biosciences, San Jose, CA). Autofluorescence was gated out with FlowJo (BD Biosciences, San Jose, CA), using a non-fluorescent strain of *P. putida* KT2440 carrying an empty vector plasmid for reference (Supplementary Figure 8). To remove noise, the GFP mean for samples with less than 150 events after gating was set to 0 (Supplementary Figure 9). Otherwise, the GFP mean of the remaining events after gating was reported. Statistical significance was determined by T-test between replicates.

Resource Availability

Data and Code Availability

Source code of the DAP-seq analysis pipeline is available at <https://github.com/novichkov-lab/dap-seq-utils>

DAP-seq data have been deposited in NCBI's Gene Expression Omnibus⁵⁵ and are accessible through GEO Series accession number GSE157075 (<https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157075>).

Single carbon source fitness data is available at <http://fit.genomics.lbl.gov>

References

1. Galperin, M. Y. Diversity of structure and function of response regulator output domains. *Curr. Opin. Microbiol.* **13**, 150–159 (2010).
2. Wuichet, K., Cantwell, B. J. & Zhulin, I. B. Evolution and phyletic distribution of two-component signal transduction systems. *Curr. Opin. Microbiol.* **13**, 219–225 (2010).
3. Capra, E. J. & Laub, M. T. Evolution of two-component signal transduction systems. *Annu. Rev. Microbiol.* **66**, 325–347 (2012).
4. McClune, C. J. & Laub, M. T. Constraints on the expansion of paralogous protein families. *Curr. Biol.* **30**, R460–R464 (2020).
5. Choi, K. & Kim, S. Building interacting partner predictors using co-varying residue pairs between histidine kinase and response regulator pairs of 48 bacterial two-component systems. *Proteins* **79**, 1118–1131 (2011).
6. Forslund, S. K., Kaduk, M. & Sonnhammer, E. L. L. Evolution of protein domain

- architectures. *Methods Mol. Biol.* **1910**, 469–504 (2019).
7. Ortega, Á., Zhulin, I. B. & Krell, T. Sensory repertoire of bacterial chemoreceptors. *Microbiol. Mol. Biol. Rev.* **81**, (2017).
8. Briegel, A. *et al.* Universal architecture of bacterial chemoreceptor arrays. *Proc Natl Acad Sci USA* **106**, 17181–17186 (2009).
9. Ryjenkov, D. A., Tarutina, M., Moskvina, O. V. & Gomelsky, M. Cyclic diguanylate is a ubiquitous signaling molecule in bacteria: insights into biochemistry of the GGDEF protein domain. *J. Bacteriol.* **187**, 1792–1798 (2005).
10. Chen, Y.-T., Chang, H. Y., Lu, C. L. & Peng, H.-L. Evolutionary analysis of the two-component systems in *Pseudomonas aeruginosa* PAO1. *J. Mol. Evol.* **59**, 725–737 (2004).
11. Voordeckers, K., Pougach, K. & Verstrepen, K. J. How do regulatory networks evolve and expand throughout evolution? *Curr. Opin. Biotechnol.* **34**, 180–188 (2015).
12. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1**, 16048 (2016).
13. Cases, I., Ussery, D. W. & De Lorenzo, V. The σ^{54} regulon (sigmulon) of *Pseudomonas putida*. *Environ. Microbiol.* **5**, 1281–1293 (2003).
14. Tatke, G., Kumari, H., Silva-Herzog, E., Ramirez, L. & Mathee, K. *Pseudomonas aeruginosa* MifS-MifR Two-Component System Is Specific for α -Ketoglutarate Utilization. *PLoS ONE* **10**, e0129629 (2015).
15. Lundgren, B. R. *et al.* Genetic analysis of the assimilation of C5-dicarboxylic acids in *Pseudomonas aeruginosa* PAO1. *J. Bacteriol.* **196**, 2543–2551 (2014).
16. Sonawane, A. M., Singh, B. & Röhm, K.-H. The AauR-AauS two-component system regulates uptake and metabolism of acidic amino acids in *Pseudomonas putida*. *Appl. Environ. Microbiol.* **72**, 6569–6577 (2006).

17. Rajeev, L., Garber, M. E. & Mukhopadhyay, A. Tools to map target genes of bacterial two-component system response regulators. *Environ. Microbiol. Rep.* **12**, 267–276 (2020).
18. Garber, M. E. *et al.* Multiple signaling systems target a core set of transition metal homeostasis genes using similar binding motifs. *Mol. Microbiol.* **107**, 704–717 (2018).
19. Skerker, J. M. *et al.* Rewiring the specificity of two-component signal transduction systems. *Cell* **133**, 1043–1054 (2008).
20. Pougach, K. *et al.* Duplication of a promiscuous transcription factor drives the emergence of a new regulatory network. *Nat. Commun.* **5**, 4868 (2014).
21. Capra, E. J., Perchuk, B. S., Skerker, J. M. & Laub, M. T. Adaptive mutations that prevent crosstalk enable the expansion of paralogous signaling protein families. *Cell* **150**, 222–232 (2012).
22. Casas-Pastor, D. *et al.* Expansion and re-classification of the extracytoplasmic function (ECF) σ factor family. *BioRxiv* (2019). doi:10.1101/2019.12.11.873521
23. Ronneau, S. & Hallez, R. Make and break the alarmone: regulation of (p)ppGpp synthetase/hydrolase enzymes in bacteria. *FEMS Microbiol. Rev.* **43**, 389–400 (2019).
24. Dalebroux, Z. D. & Swanson, M. S. ppGpp: magic beyond RNA polymerase. *Nat. Rev. Microbiol.* **10**, 203–212 (2012).
25. Potvin, E., Sanschagrin, F. & Levesque, R. C. Sigma factors in *Pseudomonas aeruginosa*. *FEMS Microbiol. Rev.* **32**, 38–55 (2008).
26. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life. *Nat. Rev. Genet.* **16**, 472–482 (2015).
27. Smanski, M. J. *et al.* Synthetic biology to access and expand nature’s chemical diversity. *Nat. Rev. Microbiol.* **14**, 135–149 (2016).

28. Way, J. C., Collins, J. J., Keasling, J. D. & Silver, P. A. Integrating biological redesign: where synthetic biology came from and where it needs to go. *Cell* **157**, 151–161 (2014).
29. Schmidl, S. R. *et al.* Rewiring bacterial two-component systems by modular DNA-binding domain swapping. *Nat. Chem. Biol.* **15**, 690–698 (2019).
30. Mistry, J., Finn, R. D., Eddy, S. R., Bateman, A. & Punta, M. Challenges in homology search: HMMER3 and convergent evolution of coiled-coil regions. *Nucleic Acids Res.* **41**, e121 (2013).
31. Gumerov, V. M., Ortega, D. R., Adebali, O., Ulrich, L. E. & Zhulin, I. B. MiST 3.0: an updated microbial signal transduction database with an emphasis on chemosensory systems. *Nucleic Acids Res.* **48**, D459–D464 (2020).
32. Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Mol. Biol. Evol.* **30**, 772–780 (2013).
33. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 — approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (2010).
34. Huerta-Cepas, J., Serra, F. & Bork, P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. *Mol. Biol. Evol.* **33**, 1635–1638 (2016).
35. Gibson, D. G. *et al.* Enzymatic assembly of DNA molecules up to several hundred kilobases. *Nat. Methods* **6**, 343–345 (2009).
36. Hillson, N. J., Rosengarten, R. D. & Keasling, J. D. j5 DNA assembly design automation software. *ACS Synth. Biol.* **1**, 14–21 (2012).
37. Studier, F. W. Protein production by auto-induction in high density shaking cultures. *Protein Expr. Purif.* **41**, 207–234 (2005).
38. Chiniquy, J., Garber, M. E., Mukhopadhyay, A. & Hillson, N. J. Fluorescent amplification

for next generation sequencing (FA-NGS) library preparation. *BMC Genomics* **21**, 85 (2020).

39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**, 2114–2120 (2014).

40. Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E. & Edwards, R. A. FOCUS: an alignment-free model to identify organisms in metagenomes using non-negative least squares. *PeerJ* **2**, e425 (2014).

41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient alignment of short DNA sequences to the human genome. *Genome Biol.* **10**, R25 (2009).

42. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25**, 2078–2079 (2009).

43. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq experiments for DNA-binding proteins. *Nat. Biotechnol.* **26**, 1351–1359 (2008).

44. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids Res.* **37**, W202–8 (2009).

45. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol.* **20**, 238 (2019).

46. Mironov, A. A., Vinokurova, N. P. & Gelfand, M. S. Software for analysis of bacterial genomes. *Mol Biol (NY)* **34**, 222–231 (2000).

47. Thompson, M. G. *et al.* Massively Parallel Fitness Profiling Reveals Multiple Novel Enzymes in *Pseudomonas putida* Lysine Metabolism. *MBio* **10**, (2019).

48. Rand, J. M. *et al.* A metabolic pathway for catabolizing levulinic acid in bacteria. *Nat. Microbiol.* **2**, 1624–1634 (2017).

49. Incha, M. R. *et al.* Leveraging host metabolism for bisdemethoxycurcumin production in *Pseudomonas putida*. *Metab. Eng. Commun.* **10**, e00119 (2020).
50. Wetmore, K. M. *et al.* Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. *MBio* **6**, e00306-15 (2015).
51. Thompson, M. G. *et al.* Fatty Acid and Alcohol Metabolism in *Pseudomonas putida*: Functional Analysis Using Random Barcode Transposon Sequencing. *Appl. Environ. Microbiol.* **86**, (2020).
52. Bakan, A., Meireles, L. M. & Bahar, I. ProDy: protein dynamics inferred from theory and experiments. *Bioinformatics* **27**, 1575–1577 (2011).
53. Bakan, A. *et al.* Evol and ProDy for bridging protein sequence evolution and structural dynamics. *Bioinformatics* **30**, 2681–2683 (2014).
54. Engler, C., Kandzia, R. & Marillonnet, S. A one pot, one step, precision cloning method with high throughput capability. *PLoS ONE* **3**, e3647 (2008).
55. Edgar, R., Domrachev, M. & Lash, A. E. Gene Expression Omnibus: NCBI gene expression and hybridization array data repository. *Nucleic Acids Res.* **30**, 207–210 (2002).

Acknowledgements

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

capabilities to perform high throughput fitness assays, Mitchell G. Thompson, Morgan Price, Adam Deutschbauer.

Funding source:

This work was part of the ENIGMA- Ecosystems and Networks Integrated with Genes and Molecular Assemblies (<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 & Environmental Research under contract number DE-AC02-05CH11231 between LBNL and the U. S. Department of Energy. The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication. The United States Government retains and the publisher, by accepting the article for publication, acknowledges that the United States Government retains a non-exclusive, paid-up, irrevocable, world-wide license to publish or reproduce the published form of this manuscript, or allow others to do so, for United States Government purposes.

Author Contributions

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

554 Declarations of Interest

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