# Sequence-space expansion is driven by the emergence of new signaling pathways

Megan E. Garber[1, 2], Vered Frank[2], Alexey E. Kazakov[6], Matthew R. Incha[2,4], Alberto A. Nava[2,5], Hanqiao Zhang[2,6], Jay D. Keasling[2,4,5,7,8], Lara Rajeev[2], 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. University of California, Berkeley Department of Bioengineering
4. University of California, Berkeley Department of Plant and Microbial Biology
5. University of California, Berkeley Department of Chemical and Biomolecular Engineering
6. Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division
7. Center for Biosustainability, Danish Technical University, Lyngby, Denmark
8. Center for Synthetic Biochemistry, Shenzhen Institutes for Advanced Technologies, Shenzhen, China

* Correspondence: amukhopadhyay@lbl.gov

## Abstract

This report presents multi-genomes evidence that sequence-space expansion occurs when pathways with incompatible functions emerge. It is speculated that as pathways evolve within the sequence-space of highly paralogous signaling systems, they insulate against unwanted interactions, and can expand into new sequence-space. To prove that sequence-space expansion is driven by pathway insulation against emergent pathways, we develop a comprehensive, holistic (whole genomes) new pathway discovery pipeline. The holistic pipeline, which uses t-distributed stochastic neighbor embedding algorithms and phylogeny to map the extant sequence-space of thousands of bacterial two-component systems, reveals a new signaling pathway, resultant from domain swapping. By employing functional genomics screens in our model system *Pseudomonas putida* KT2440, we provide empirical evidence that descendents of the functionally discordant domain swapped pathway and its nearest neighbors in sequence-space are paralogous. We conclude that the emergence of this new pathway necessitated pathway insulation and drove sequence-space expansion.


## Introduction

Sequence-space is a theoretical construct in which similar proteins segregate into clusters based on their amino acid sequences(Starr et al. 2017; McClune and Laub 2020). The concept of sequence-space has been previously utilized to explain the evolution of highly paralogous signaling pathways such as two-component systems (TCS)(Capra and Laub 2012). Previous empirical studies have demonstrated that histidine kinases (HK) and response regulators (RR) have co-evolved specificity between cognate pairs(Capra et al. 2010; Capra et al. 2012; McClune et al. 2019; McClune and Laub 2020). From these studies, we can speculate that ancient TCSs competed

within sequence-space to insulate their interactions from unwanted cross-talk. However, sequence-space expansion as a vital mode of insulation for new pathways has never been shown. In order to test whether the emergence of new signaling pathways can drive sequence-space expansion, we needed to build out the actual sequence-space of extant TCSs. This required comprehensive, holistic evaluation of the phylogenetic relationships between extant TCSs across the bacterial clade.

Evaluating phylogenetic relationships between extant TCSs is challenging due to their extensive redundancy within genomes(Capra and Laub 2012). Varying counts of TCSs in genomes reflect the niche adaptation of a species(Alm et al. 2006; Ashby and Houmard 2006), rather than the species' taxonomic classification(Koretke et al. 2000). TCS families(Galperin 2010), which are defined by the domain architecture of RRs (Figure 1a,b), are currently the best tool available for classifying TCSs. This classification, however, does not take TCS phylogeny or sequence-space into account. Our holistic pipeline pairs phylogenetic and sequence-space analyses with classification by TCS families (Fig. S1). We utilize this pipeline to elucidate anomalous clusters, which represent domain swapping events.

Domain swapping provides microbes with a rapid tool to neofunctionalize their existing repertoire of signaling systems(Pao and Saier 1995; Alm et al. 2006; Forslund et al. 2019). Within a TCS domain swap, a genomic recombination event can replace the transcriptional effector domain (TED) of a RR (Pao and Saier 1995; Alm et al. 2006; Forslund et al. 2019), effectively rewiring the downstream regulation. A domain swap creates paralogous sets of TCSs that do not fall within the same TCS family(Galperin 2010). TED domain swaps are important in a physiological context, because they have the potential to rewire the cellular response for the

upstream signal being detected. When a TED domain swap occurs, a new pathway with a new function emerges within TCS sequence-space (Figure 1c,d).

To identify newly emerged pathways, we utilized our holistic pipeline to discover anomalous clustering events in TCS sequence-space. The holistic pipeline searches a multi-dimensional landscape of TCS evolution, in which the sequence-space and phylogeny represent the clustering of over 2000 TCSs found in diverse species across the entire bacterial tree of life. Known TCS families (e.g. OmpR, NarL, NtrC) segregate into clusters (Figure 1d), allowing us to define anomalous clustering events, which are the result of domain swapping, as recently emerged pathways. However, the presence of an anomalous cluster in sequence-space is not sufficient to support the hypothesis that expansion of sequence-space, required for insulation, has also occurred. Empirical evidence is needed to connect these anomalous clusters to their closest relatives, but such experimental evidence has hitherto been absent. Using functional characterization experiments, guided by evolutionary hypotheses from our holistic pipeline, we demonstrate that a predicted new pathway is functionally and biochemically tied to its parent pathway. Specifically, we determine that a set of paralogous TCSs identified within an anomalous cluster are involved in regulating the consumption of critical carbon sources with similar chemical structure and are also biochemically linked. Our results strongly support a long-standing theory on the evolution for TCSs, in which sequence-space expansion is driven by newly emerged pathways evolving towards insulation and against cross-talk.

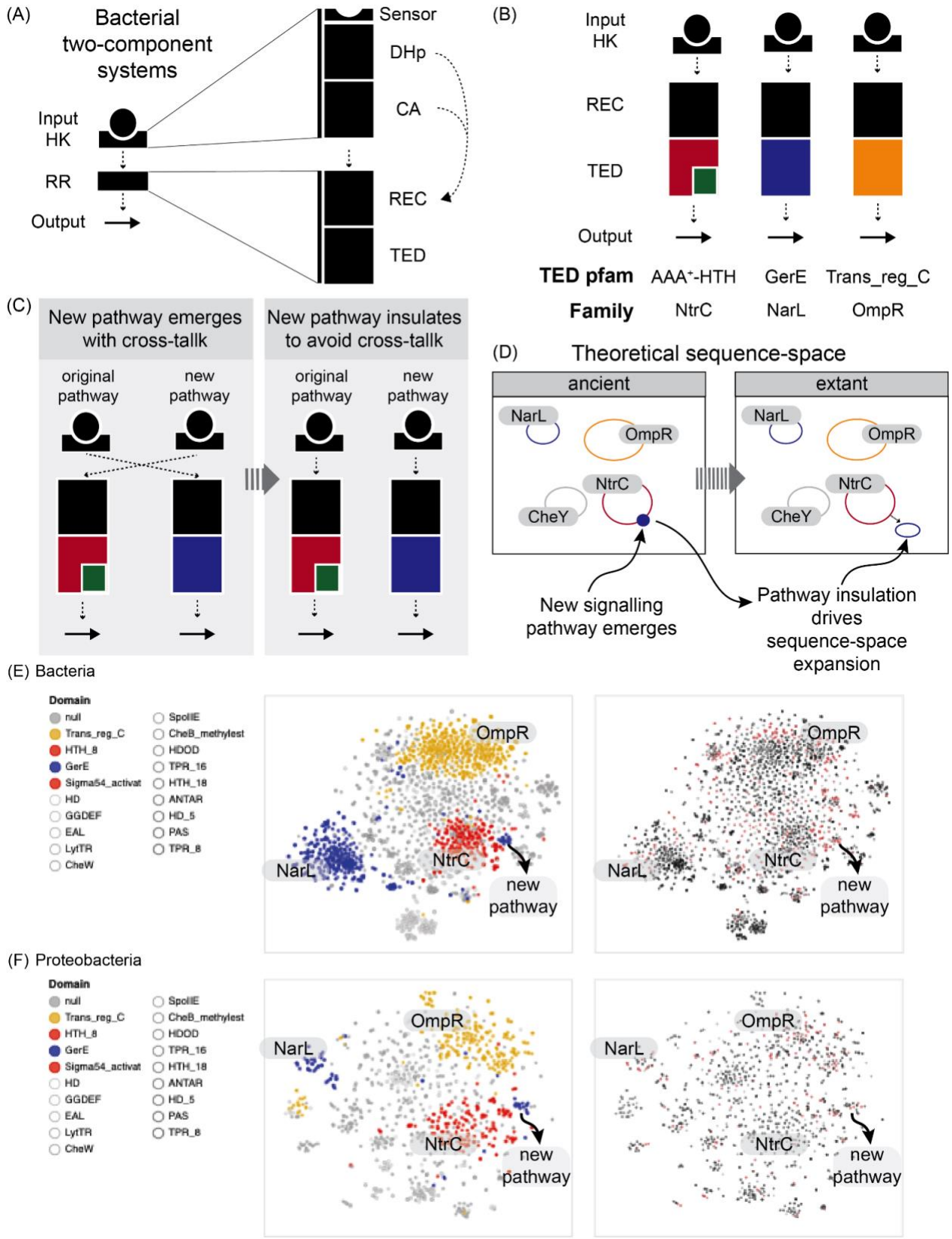


Figure 1: ***Discovery and impact of domain swaps in two-component systems sequence-space:*** (A) Canonical TCS have input signals that activate histidine kinases (HK) and response regulators (RR) to

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 and family. (C) New pathways that emerge from domain swapping cross-talk with their closest relatives. Over time, these pathways will become insulated from each other to avoid cross-talk. (D) Theoretical model for sequence-space expansion driven by the emergence of a new signaling pathway. (E, F) Sequence-space of two-component systems by t-Distributed Stochastic Neighbor Embedding (t-SNE) of the receiver domains for thousands of response regulators identified from bacteria and proteobacteria. The amino acid sequences are aligned, numerically encoded (one-hot-encode), and clustered by tSNE. The resultant two-dimensional visualization of protein sequences is annotated by domain architecture (right) or taxonomy (left). The red cluster which represents the NtrC family of TCSs always harbors a small blue (NarL family) cluster. This anomalous cluster is the domain swap cluster. The left hand panels highlight the genomes relevant to this study, proteobacteria and gammaproteobacteria, in pink. We note the domain swap cluster is dominated by proteobacteria although other phyla are present.

## Results

The holistic, whole genomes pipeline takes into account the alignment and evolution of conserved domains of RRs (REC) from divergent species (Fig. S1). It differs from previous phylogenetic descriptions of TCSs(Pao and Saier 1995; Chen et al. 2004), because we consider TCSs from genomes representing the TCS landscape from the entire bacterial tree of life. For this, we collected and aligned the amino acid sequences of 2,178 RRs (REC domains) from a diverse set of 64 species across the entire bacterial clade (See External databases 1, 2) and implemented phylogenetic or machine-learning clustering analysis (Figure 1e,f, Fig. S2). Each point in the calculated sequence-space represents a unique RR. An RR's orientation in the sequence-space indicates the sequence-based relationship to the other RRs in sequence-space. The clustering in our calculated sequence-space reaffirms that two-component systems from diverse taxa do not segregate by clades of species, but instead cluster by domain architecture(Pao and Saier 1995). A domain swap, indicated by anomalous clustering, is observed within the NtrC-like cluster (red) of the REC sequence-space and tree. Here, NarL-like RRs (blue) are found. Interestingly, the anomalous cluster in the bacterial sequence-space consisted primarily of TCSs from Proteobacteria. When we repeated our holistic pipeline with 1,065 RRs from 22 representative

species of Proteobacteria (See External database 3, 4), we observed a result consistent with the bacterial TCS phylogeny and sequence-space.

The orientation of the anomalous cluster with TCS sequence-space for both bacteria and proteobacteria implied sequence-space expansion was driven by the emergence of the new signaling pathway. To confirm this hypothesis, we needed to verify the relationship between the newly emerged pathway and its closest relatives. We proposed that the relationships observed in the phylogenetic and sequence-space of RRs could enable us to narrow the experimental test space for functional characterization. To do so, we focused on the TCSs from pseudomonads (Fig S3a, see external databases 5, 6), which harbor the domain swap in extant species.

To identify the newly emerged pathway's closest relatives within pseudomonads, we applied a second computational pipeline (Fig. S3b). We identified all RRs in five representative pseudomonads, subdividing them by their effector domains, considering only RRs with transcriptional effector domains (TEDs) (Table S1). The observed clustering within the domain swap cluster indicated that the NarL-like RRs are not related to their expected TCS family, NarL/FixJ RRs, but instead evolved from the NtrC-like RRs. We observed a similar trend with the co-evolving DHp-CA (HisKA-HATPase_C) tree. The leaves of the DHp-CA tree cluster by the domain architecture of their cognate RRs (Figure 2) including the domain swapped cluster. This result indicates that the cognate HKs to the NarL-like RRs in the domain swapped cluster are not related to the cognate HKs for NarL/FixJ regulators, such as NarP (NarQ)/FixL.

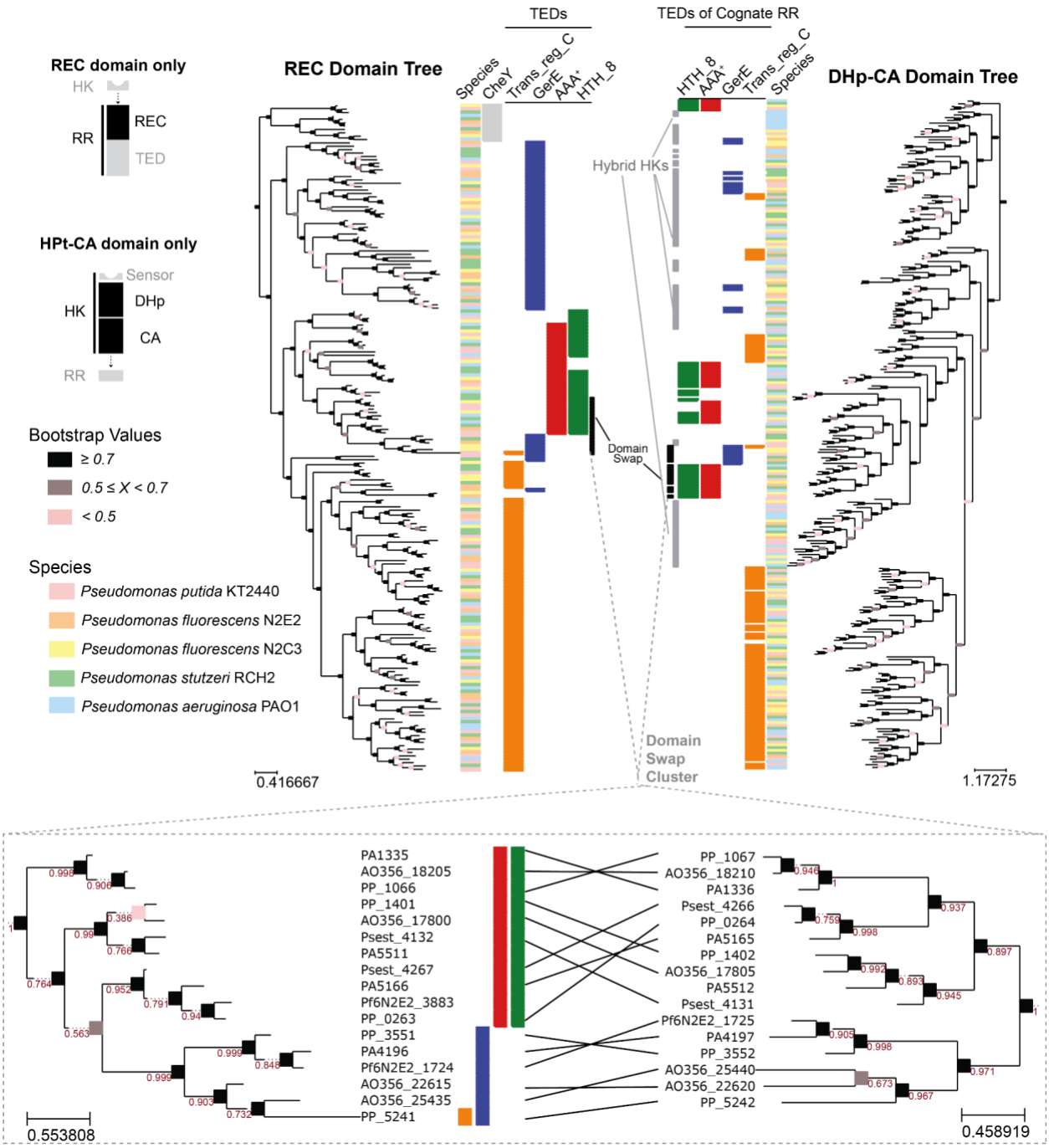


Figure 2: **Phylogenetic trees of REC (left) and DHp-CA (right) domains of RRs and HKs from select pseudomonads**. Trees are annotated with species and transcriptional effector domain (TED) at the leaves. Hybrid HKs on the DHp-CA tree are highlighted in grey, and are notably associated with TCS that 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 DHp-CA (right) domains trees annotated with protein name and/or TED. Lines are drawn for predicted cognate HK-RR pairs.

We were interested in understanding the functional relationship between the members of the domain swapped cluster. Within the swapped cluster (Figure 2) 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). 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 sources(Cases et al. 2003). Additionally, NtrC representatives within the swapped cluster were previously characterized as regulators of amino acid or di-carboxylic acid metabolism(Cases et al. 2003; Sonawane et al. 2006; Lundgren et al. 2014; Tatke et al. 2015). The functions of the NarL-like representatives were unknown. We hypothesized that their phylogenetic relationship indicated a functional (signal sensing) similarity, and could enable us to strategically target the functions of the unknown TCSs within the cluster.

To determine the cis-regulatory binding sites, we applied an automated pipeline of DNA-affinity purification with next generation sequencing (DAP-seq)(Garber et al. 2018; Rajeev et al. 2020) (Fig. S4a), to the entire subset of RRs in the swapped cluster. We found highly enriched DNA binding targets for all of the NtrC-like RRs (Fig. S4b, External Databases 7,8). 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 (Figure 3a, c, Fig. S4b, Table S2). We were only able to find genomic targets for NarL-like RRs from *P. fluorescens* FW300-N2C3 (Fig. S4b, External Databases 7,8), 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 fitness libraries, RB-TnSeq (Wetmore et al. 2015) of *P. putida* KT2440 grown in single carbon source selection conditions. We observed that the gene clusters harboring the TCSs 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 glutamic(Sonawane et al. 2006), succinic, and a-ketoglutaric acids(Lundgren et al. 2014; Tatke et al. 2015) 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. 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. When the reporter strains were stimulated in genetic backgrounds absent of the corresponding transcriptional RR, fluorescence response was ablated (Figure 3b).

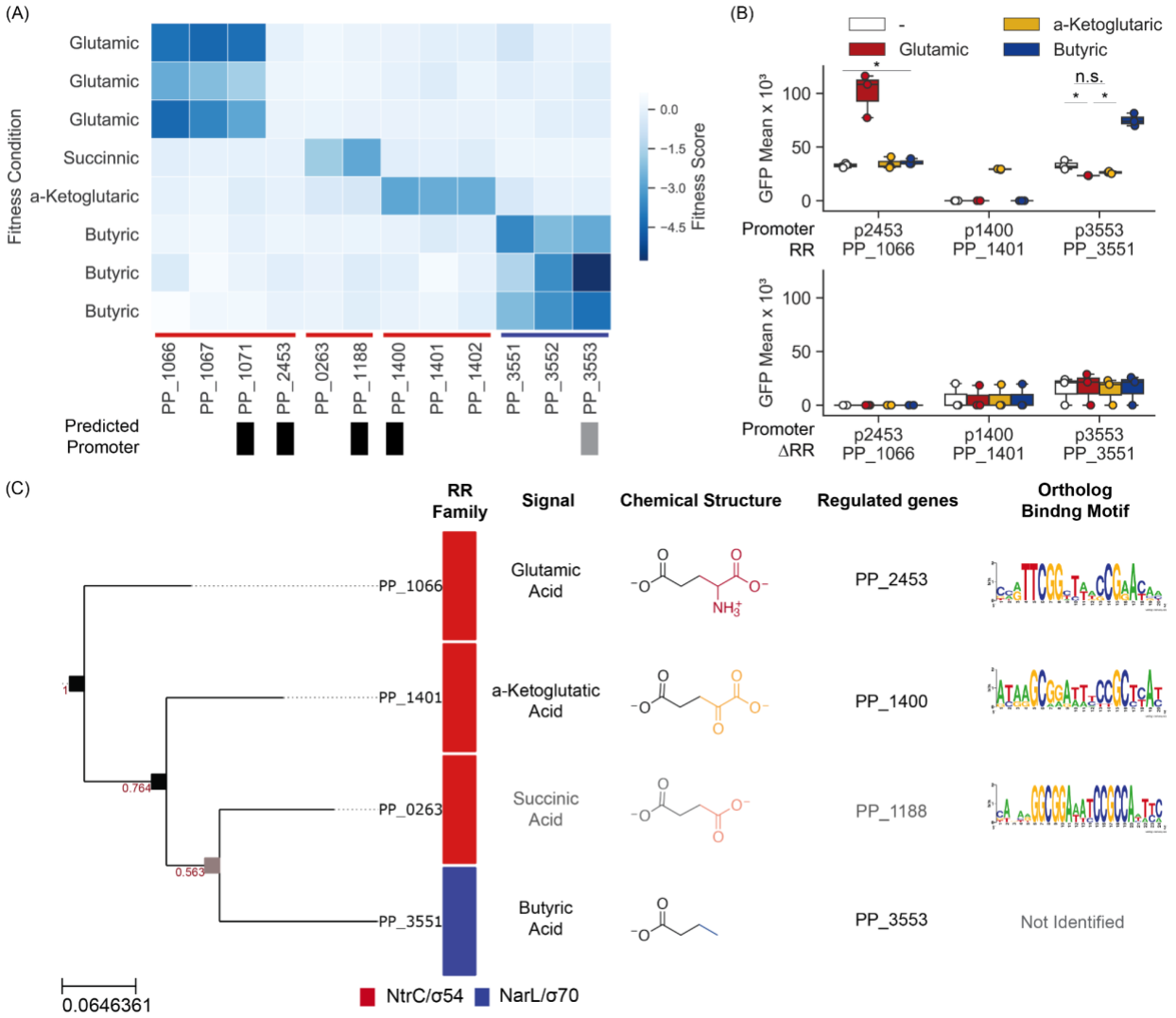


Figure 3: ***Functional genomics for determining input signal and output genes*:** (A) Heatmap of fitness data for TCS gene clusters within the domain swapped cluster. Replicated rows indicate unique fitness experiments applied with the same selection condition. A low fitness score (fitness< - 2) indicates a gene may have condition-specific essentiality. Predicted target 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 ΔRR 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 x $10^3$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.

Our experimental results validate the hypothesis that the members of domain swapped cluster are functionally similar and sense organic compounds with similar structures. All signals for TCSs in the 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 the HK-RR-promoter (PP_3552-PP_3551-pPP_3553) was unknown until this study. We therefore rename the related TCSs with REC paralogs, the Carboxylic Acid Responsive TCSs, CarSR I, II, III, IV (PP_0263, PP_1066, PP_1401, PP_3551).

Despite their ability to sense molecules with chemical similarity, modern CarSRs sense a broad range of signals. We therefore propose an ancestral model to explain this observation. In our model (Figure 4a), common ancestors of CarSR (i) had promiscuous sensors, detecting varying carboxylic acids; (ii) readily engaged in cross-talk with each other; and (iii) regulated promiscuous enzymes(Huberts and van der Klei 2010) that could act on a spectrum of carboxylic acid substrates.

To demonstrate the ancestral model we engineered the TCSs into a pseudo-ancestral state (Figure 4b,c). We employed co-variance analysis to identify co-varying residues between DHp-CA and REC domains (Fig. S4). The first alpha helix of the REC domain, previously shown to share points of contact with the DHp domain(Skerker et al. 2008), contained the highest scoring residues. We applied site-directed mutagenesis switching the highest scoring residues of the REC domain of CarR II ($RR_{II}$) and CarR IV ($RR_{IV}$) and tested activity with a complementation assay (Figure 4b, Fig. S6-7). Mutant RRs (indicated by *) exhibited behavior similar to their native counterparts. $RR_{II}$*-$pRR_{II}$, similar to $RR_{IV}$-$pRR_{IV}$, responded significantly to butyric acid only when the cognate, histidine, $HK_{IV}$, was present. These results show how a promiscuity hypothesis is viable and how gradual changes in specificity determining residues could be responsible for the specialization observed in modern CarSR homologs.

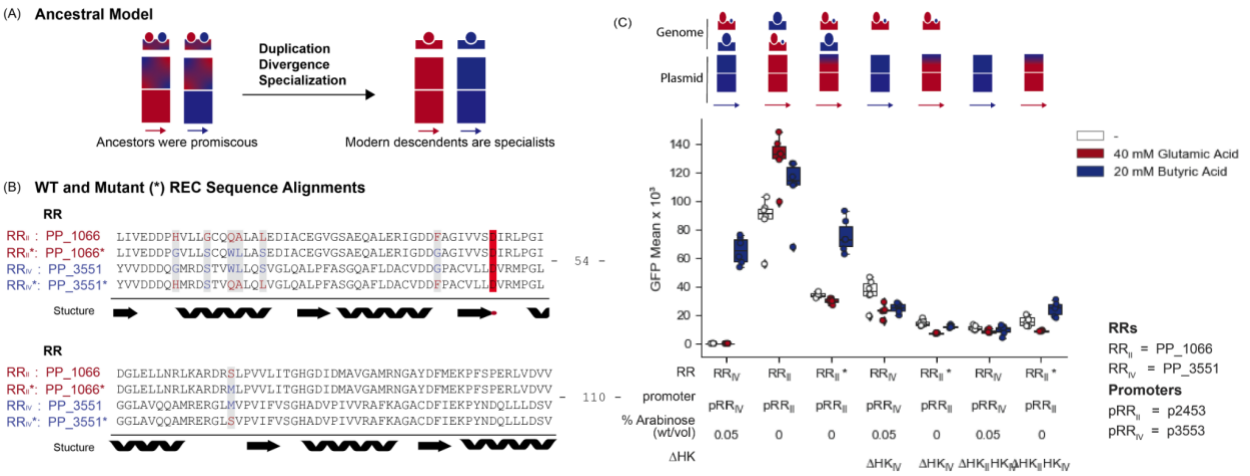


Figure 4: ***Covarying residue swaps alter REC specificity of butyric acid responsive TCS:*** (A) Model where hypothetical promiscuous ancestors of the domain swapped cluster evolved to become modern day specialists. (B) Sequence alignments of REC domains of native RRs PP_1066 (CarR II: $RR_{II}$), PP_3551 (CarR IV: $RR_{II}$), and mutated $RR_{II}$*, $RR_{IV}$*. 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 $RR_{IV}$-derived residues. Structural information for the REC domain is shown below the sequences, where arrows are β-sheets and ribbons are α-helices. (C) $RR_{II}$, driving glutamic acid responsive promoter ($pRR_{II}$) switches its specificity to butyric acid responsive when mutated to $RR_{II}$*. The response is diminished in the absence of $HK_{IV}$. Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points with black diamonds, outliers; n = 6.

In an alternative hypothesis, CarSR IV - the NarL-like constituents of the CarSR cluster - could have originated from horizontal gene transfer (HGT) instead of domain swapping. Although CarSR homologs are predominantly found in Proteobacteria, other phyla, such as Planctomycetes and Acidobacteria, harbor homologs (Figure 1, Fig. S2, S8). Considering only the conserved domains of CarSR II and CarSR IV homologs, we were able to demonstrate that CarSR IV likely originated in Alphaproteobacteria (Fig. S9). We speculate that the homologs observed in other phyla are derived from HGT. Given the clustering configuration within Proteobacteria TCSs (Figure 1, Fig. S2), domain swapping, not HGT, is the most likely explanation for the origins of CarSR IV.

Another important consideration in evaluating the evolution of TCSs is the domain architecture of HKs. CarS I-IV in *P. putida* KT2440 have divergent upstream domain architectures. CarS I-III are transmembrane bound with a dCache_1 domain, whereas CarS IV is

cytosolic with a PAS domain. Considering this inconsistency, we hypothesized that the observed domain architecture stems from domain loss, where a common ancestor harbored both a periplasmic dCache_1 domain and a cytosolic PAS domain. We discovered that CarSR homologs with both sensing domains exist in nature (Fig. S10a). Consistent with our CarSR origins hypothesis, the dual domained CarSR homologs are found primarily in Proteobacteria (Fig. S10b,c).

## Discussion

Highly paralogous signaling pathways such as bacterial two-component systems insulate from each other as they evolve(Capra and Laub 2012). In this study we showed that domain swaps, which can rapidly change the effector function of signaling systems, can drive insulation via sequence-space expansion. We developed a holistic pipeline to map the multi-dimensional landscape of TCSs across the entire bacterial clade, and discovered newly emerged signaling pathways that were the result of domain swapping (Figure 1). Mapping TCSs sequence-space of extant species was facilitated by the massive amount of sequencing data available and the abundance of divergent TCSs across single genomes(Galperin 2005; Gumerov et al. 2020). Many other studies utilize deep-mutational scanning(Starr et al. 2017; McClune et al. 2019; Repecka et al. 2021) to generate screenable libraries that cover the sequence diversity we observe for extant TCSs. One such study found that TCSs from *Escherichia coli* are sparsely distributed across sequence-space, enabling countless sequence combinations for orthogonal pathway engineering(McClune et al. 2019). Although we observed that bacterial TCS sequence-space is densely distributed (Figure 1e), we found the sequence-space for the lower taxonomic levels is much more sparse (Figure 1f, Fig. S3). When we dropped to the genera level of Pseudomonads

(Fig. S3), the clusters were so segregated in sequence-space that we can draw a parallel to the conclusions from the *E. coli* study. Specifically, bacterial strains have the capacity to evolve into new unoccupied sequence-space and generate orthogonal pathways. We can now claim that individual strains can trigger their potential to expand into new sequence-space when new, discordant pathways emerge.

We utilized evolutionary hypotheses from our pipeline to functionally characterize members of a newly emerged pathway alongside its closest relatives. We found that the newly emerged pathway is biochemically linked to its nearest neighbors in sequence-space and that the signals detected have similar chemical structures. We therefore propose that when this new pathway emerged from domain swapping, the resultant paralogs were functionally incompatible and competed for sequence-space (Figure 1c,d). Our study supports the more general theory that sequence-space expansion is driven by the emergence of new signaling systems in order to avoid cross-talk between paralogous relatives.

In the case of CarSR, sequence-space expansion was driven by the emergence of a new signaling pathway with a new effector function. Our computational and empirical results inform a working model for the emergence of the new pathway (Fig. S11). Our framework assumes the existence of an ancestral promiscuous NtrBC-like TCS which sensed a spectrum of carboxylic acids and coordinated metabolism. In an ancestral alphaproteobacteria strain, a duplicated CarR underwent a domain swap, which switched the TED and rewired its cis-regulation. Domain swapping built functionally divergent paralogs of the promiscuous carboxylic acid sensing systems that were unable to share sequence-space (Figure 1c). These systems were therefore driven to speciation, leading to the extant descendents observed in this study (Figure 1c,d).

It is notable that domain swapping can alter sigma factor dependence. Because of its change in sigma factor dependence, the domain swapped CarSR paralog was functionally different from its parent. NtrC-like RRs require sigma54 transcription factors to activate genes in their regulons, whereas NarL-like RRs interact with sigma70 (Mejía-Almonte et al. 2020). Cellular levels of small nucleotide, (p)ppGpp, in combination with the protein DksA regulate RNA polymerase (RNAP) preference for sigma70 (extracytoplasmic function family)(Casas-Pastor et al. 2021) or sigma54(Dalebroux and Swanson 2012; Ronneau and Hallez 2019). While the environmental conditions needed for pseudomonads to reach the appropriate levels of (p)ppGpp and DksA for sigma54 occupation of RNAP are unknown(Potvin et al. 2008), it can be reasoned that the switch is lifestyle dependent. CarSR are metabolite responsive systems, and could play a major role in survival and fitness. We propose 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. We speculate that this important functional difference was the primary driver for maintenance of the new pathway and sequence-space expansion.

Our study also addresses a major bottleneck in the study of bacterial two-component systems, the relationships and connections between functionally annotated and unannotated bacterial signaling systems. We use the CarSR domain swap, discovered and functionally annotated in this study, to demonstrate that a comprehensive evolutionary model can be used to pair domain swapped or orphan two-component systems to their closest relatives in sequence-space. We show that we can use evolutionary information to derive functional hypotheses and use this information to define the chemical-space(Coley 2021) for functional characterization assays. With our holistic view of TCSs, we demonstrated that evolution and function are tethered. By characterizing anomalies in TCS clustering, we were able to provide empirical evidence for (i)

transcriptional rewiring by domain swapping (ii) conservation of upstream signal detection after duplication and divergence of co-evolving parts. With experimental characterization of CarSR we were able to highlight how ligand specificity is a function of conservation of downstream genes rather than signaling modules themselves. Characterization using co-evolutionary hypotheses may be the key to distinguishing the many still unknown functions of TCSs.

# Methods

**Phylogenetic analysis of REC and DHp-CA trees**

*Pseudomonas* RRs and HKs were identified with hmmsearch from HMMER v3.1b2(Mistry et al. 2013) to search for all proteins with Response_reg (PF00072) and HisKA (PF00512) domains. Hmmsearch was used to query the domain architecture of the signaling proteins and selected for RRs with HTH_8, Trans_reg_C, $AAA^+$, GerE, HTH_18 (AraC) domains. Alternatively, we extracted a diverse set of genomes from given genera, phyla, or kingdoms (external databases 1, 3, 5) found in the Microbial Signal Transduction database (MISTDB-API)(Gumerov et al. 2020). We again used MISTDB-API to extract amino acid sequences of HKs, RRs, and their respective domain architectures (external databases 2, 4, 6). Sequences with greater >90 or >95% homology identified with CD-HIT web server(Huang et al. 2010) were removed from further analysis. DHp-CA domain sequences were extracted from whole sequences based on coordinates determined by the hmmsearch. These domains were aligned using the MAFFT-LINSI algorithm from MAFFT v7.310(Katoh and Standley 2013). REC domains were aligned with Hmmalign from HMMER v3.1b2(Mistry et al. 2013) to Response_reg.hmm (PF00072). Phylogenetic trees were constructed using FastTree 2(Price et al. 2010) with parameters, pseudo, lg. Trees were visualized and annotated using python based ETE3(Huerta-Cepas et al. 2016).

## Two-dimensional visualization of protein sequences

Multiple sequence alignments from above were encoded into a discrete variable vector representation(Jing et al. 2020). This process is performed for each sequence 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 one. 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 and dimensionally reduced to Nx500 dimensions by principal component analysis before being reduced to Nx2 dimensions using the t-Distributed Stochastic Neighbor Embedding algorithm(van der Maaten and Hinton 2008), with parameters n_components=2, verbose=0, n_iter=1000. Perplexity was determined as a function of the number of strains being visualized.

## Automated DNA affinity purification - seq

*DNA preparation for NGS*

Pseudomonas isolates were cultured in either LB or minimal media (see Table S3 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 assembly(Gibson et al. 2009). Plasmid design was facilitated by j5 DNA assembly design(Hillson et al. 2012), see Table S4 for primers.

*Automated DNA affinity purification*

Quadruplicates of expression strains were grown in autoinduction media (Zyp-5052(Studier 2005)) 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 libraries(Chiniquy et al. 2020). 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.0.36(Bolger et al. 2014). The resulting reads were checked for contamination using FOCUS (Silva et al. 2014). Then the reads were aligned to the corresponding *Pseudomonas* spp. genome using Bowtie v1.1.2 (Langmead et al. 2009) with –m 1 parameter (report reads with single alignment only). Resulting SAM files were converted to BAM format and sorted using samtools v 0.1.19(Li et al. 2009). Peak calling was performed using SPP 1.16.0(Kharchenko et al. 2008) 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 MEME(Bailey et al. 2009) 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 a comparative genomics approach. Orthologous RRs were identified by OrthoFinder2(Emms and Kelly 2019). 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 package(Mironov et al. 2000) with the “inverted repeat” option.

## Fitness Experiments

As described in prior reports(Thompson et al. 2019), the *P. putida* KT2440 RB-TnSeq library, JBEI-1, was thawed, inoculated into 25 mL of LB medium supplemented with 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 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(Wetmore et al. 2015; Rand et al. 2017; Incha et al. 2020). Single carbon source fitness data is available at http://fit.genomics.lbl.gov(Thompson et al. 2020).

## 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. DHp (HisKA), CA (HATPase_C), and REC (Response_reg) domain boundaries were determined with hmmsearch from HMMER v3.1b2(Mistry et al. 2013). 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(Katoh and Standley 2013). Alignment files were then queried for coevolution with the ProDy *Evol* suite(Bakan et al. 2011; Bakan et al. 2014) 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 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, 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(Gibson et al. 2009). Primers in Table S4. 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 cloning(Engler et al. 2008) and j5 DNA assembly design(Hillson et al. 2012) (diva.jbei.org), primers in Table S4. The plasmids were transformed into gene 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 (Table S3) with co-varying mutations (co-variation score > 1.1) were cloned into the complementation plasmids with Gibson assembly(Gibson et al. 2009). 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 a-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 for reference (Fig. S12). To remove noise, the GFP mean for samples with less than 150 events after gating was set to 0 (Fig. S13). Otherwise, the GFP mean of the remaining events after gating was reported. Statistical significance was determined by t-test between replicates.

## Phylogeny of CarSR homologs

Homologs of *P. putida* KT2440 CarS II, IV, CarR II, and IV were identified with web-based BLASTP(Johnson et al. 2008) excluding non-redundant RefSeq proteins with 5000 max targets sequences and >30% homology. To remove unrelated TCSs, sequences were further pruned including homologs between 40-50% homology. The cut-off was determined based on the phylogenetic distance from *Pseudomonas*; more closely related, higher cut-off and more distantly related, lower cut-off. Sequences with greater >90 or >95% homology identified with CD-HIT web server(Huang et al. 2010) were removed from further analysis. In few CarS homologs, occasional N-terminal extensions upstream of conserved methionine start codon were trimmed and excluded from subsequent alignments. Sequences were aligned with the MAFFT algorithm from MAFFT v7.310(Katoh and Standley 2013) using parameters --maxiterate 16 and --localpair. A phylogenetic tree was reconstructed from the multiple alignment using the maximum likelihood method implemented in the FastTree software v2.1.10(Price et al. 2010).

Files containing homologs of CarR II and IV or CarS II and IV were concatenated and aligned with Hmmalign from HMMER v3.1b2(Mistry et al. 2013), CarR with Response_reg.hmm (PF00072) and CarS with HisKA.hmm (PF00512) or HATPase_c.hmm (PF02518). The aligned CarS sequences were then concatenated, resulting in alignments containing both HisKA and HAPTase_c domains. Phylogenetic trees were constructed using FastTree 2(Price et al. 2010) with parameters, pseudo, lg.

## CarS homologs with dCache_1 and PAS domains

We identified 56 HKs with the predicted domain architecture, dCache_1-PAS-DHp-CA, in the pfam database (Mistry et al. 2021). The 56 HKs were queried with BLASTp(Altschul et al. 1990) against the DHp-CA domain of either CarS II or CarS IV. A subset of the 56 HKs were

selected for their range of percent identities to CarS II and CarS IV. DHp-CA domains from this subset, CarS II, CarS IV and other unrelated HKs belonging to other TCS families were aligned with the MAFFT algorithm from MAFFT v7.310(Katoh and Standley 2013). To represent the clustering of the diverse set of HKs, phylogenetic trees were constructed using FastTree 2(Price et al. 2010).

## Acknowledgements

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; 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. 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 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.E.K., A.A.N., and H.Z. performed phylogenetic and 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.E.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. and A.M. designed the experiments and composed the initial drafts. J.D.K provided support and resources. A.M. provided resources supervision and support. All authors reviewed and edited the final draft.

## Competing Interests

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

## Data and materials 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 Omnibus[59] 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

Bibliography

Alm E, Huang K, Arkin A. 2006. The evolution of two-component systems in bacteria reveals different strategies for niche adaptation. PLoS Comput. Biol. 2:e143.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J. Mol. Biol. 215:403–410.

Ashby MK, Houmard J. 2006. Cyanobacterial two-component proteins: structure, diversity, distribution, and evolution. Microbiol. Mol. Biol. Rev. 70:472–509.

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.

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.

Bakan A, Meireles LM, Bahar I. 2011. ProDy: protein dynamics inferred from theory and experiments. Bioinformatics 27:1575–1577.

Bolger AM, Lohse M, Usadel B. 2014. Trimmomatic: a flexible trimmer for Illumina sequence data. Bioinformatics 30:2114–2120.

Capra EJ, Laub MT. 2012. Evolution of two-component signal transduction systems. Annu. Rev. Microbiol. 66:325–347.

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.

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.

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.

Cases I, Ussery DW, De Lorenzo V. 2003. The $\sigma^{54}$ regulon (sigmulon) of*Pseudomonas putida*. Environ. Microbiol. 5:1281–1293.

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.

Chiniquy J, Garber ME, Mukhopadhyay A, Hillson NJ. 2020. Fluorescent amplification for next generation sequencing (FA-NGS) library preparation. BMC Genomics 21:85.

Coley CW. 2021. Defining and exploring chemical spaces. Trends in Chemistry 3:133–145.

Dalebroux ZD, Swanson MS. 2012. ppGpp: magic beyond RNA polymerase. Nat. Rev. Microbiol. 10:203–212.

Emms DM, Kelly S. 2019. OrthoFinder: phylogenetic orthology inference for comparative genomics. Genome Biol. 20:238.

Engler C, Kandzia R, Marillonnet S. 2008. A one pot, one step, precision cloning method with high throughput capability. PLoS ONE 3:e3647.

Forslund SK, Kaduk M, Sonnhammer ELL. 2019. Evolution of protein domain architectures. Methods Mol. Biol. 1910:469–504.

Galperin MY. 2005. A census of membrane-bound and intracellular signal transduction proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol. 5:35.

Galperin MY. 2010. Diversity of structure and function of response regulator output domains. Curr. Opin. Microbiol. 13:150–159.

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.

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.

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.

Hillson NJ, Rosengarten RD, Keasling JD. 2012. j5 DNA assembly design automation software. ACS Synth. Biol. 1:14–21.

Huang Y, Niu B, Gao Y, Fu L, Li W. 2010. CD-HIT Suite: a web server for clustering and comparing biological sequences. Bioinformatics 26:680–682.

Huberts DHEW, van der Klei IJ. 2010. Moonlighting proteins: an intriguing mode of multitasking. Biochim. Biophys. Acta 1803:520–525.

Huerta-Cepas J, Serra F, Bork P. 2016. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol. Biol. Evol. 33:1635–1638.

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.

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.

Johnson M, Zaretskaya I, Raytselis Y, Merezhuk Y, McGinnis S, Madden TL. 2008. NCBI BLAST: a better web interface. Nucleic Acids Res. 36:W5-9.

Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7: improvements in performance and usability. Mol. Biol. Evol. 30:772–780.

Kharchenko PV, Tolstorukov MY, Park PJ. 2008. Design and analysis of ChIP-seq experiments for DNA-binding proteins. Nat. Biotechnol. 26:1351–1359.

Koretke KK, Lupas AN, Warren PV, Rosenberg M, Brown JR. 2000. Evolution of two-component signal transduction. Mol. Biol. Evol. 17:1956–1970.

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.

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.

Lundgren BR, Villegas-Peñaranda LR, Harris JR, Mottern AM, Dunn DM, Boddy CN, Nomura CT. 2014. Genetic analysis of the assimilation of C5-dicarboxylic acids in Pseudomonas aeruginosa PAO1. J. Bacteriol. 196:2543–2551.

van der Maaten L, Hinton G. 2008. Visualizing data using t-SNE. J Mach Learn Res 8:2579−2605.

McClune CJ, Alvarez-Buylla A, Voigt CA, Laub MT. 2019. Engineering orthogonal signalling pathways reveals the sparse occupancy of sequence space. Nature.

McClune CJ, Laub MT. 2020. Constraints on the expansion of paralogous protein families. Curr. Biol. 30:R460–R464.

Mejía-Almonte C, Busby SJW, Wade JT, van Helden J, Arkin AP, Stormo GD, Eilbeck K, Palsson BO, Galagan JE, Collado-Vides J. 2020. Redefining fundamental concepts of

transcription initiation in bacteria. Nat. Rev. Genet. 21:699–714.

Mironov AA, Vinokurova NP, Gelfand MS. 2000. Software for analysis of bacterial genomes. Mol Biol (NY) 34:222–231.

Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, Tosatto SCE, Paladin L, Raj S, Richardson LJ, et al. 2021. Pfam: The protein families database in 2021. Nucleic Acids Res. 49:D412–D419.

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.

Pao GM, Saier MH. 1995. Response regulators of bacterial signal transduction systems: selective domain shuffling during evolution. J. Mol. Evol. 40:136–154.

Potvin E, Sanschagrin F, Levesque RC. 2008. Sigma factors in Pseudomonas aeruginosa. FEMS Microbiol. Rev. 32:38–55.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 — approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490.

Rajeev L, Garber ME, Mukhopadhyay A. 2020. Tools to map target genes of bacterial two-component system response regulators. Environ. Microbiol. Rep. 12:267–276.

Rand JM, Pisithkul T, Clark RL, Thiede JM, Mehrer CR, Agnew DE, Campbell CE, Markley AL, Price MN, Ray J, et al. 2017. A metabolic pathway for catabolizing levulinic acid in bacteria. Nat. Microbiol. 2:1624–1634.

Repecka D, Jauniskis V, Karpus L, Rembeza E, Rokaitis I, Zrimec J, Poviloniene S, Laurynenas A, Viknander S, Abuajwa W, et al. 2021. Expanding functional protein sequence spaces using generative adversarial networks. Nat. Mach. Intell.

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.

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.

Skerker JM, Perchuk BS, Siryaporn A, Lubin EA, Ashenberg O, Goulian M, Laub MT. 2008. Rewiring the specificity of two-component signal transduction systems. Cell 133:1043–1054.

Sonawane AM, Singh B, Röhm K-H. 2006. The AauR-AauS two-component system regulates uptake and metabolism of acidic amino acids in Pseudomonas putida. Appl. Environ. Microbiol. 72:6569–6577.

Starr TN, Picton LK, Thornton JW. 2017. Alternative evolutionary histories in the sequence space of an ancient protein. Nature 549:409–413.

Studier FW. 2005. Protein production by auto-induction in high density shaking cultures. Protein Expr. Purif. 41:207–234.

Tatke G, Kumari H, Silva-Herzog E, Ramirez L, Mathee K. 2015. Pseudomonas aeruginosa MifS-MifR Two-Component System Is Specific for α-Ketoglutarate Utilization. PLoS ONE 10:e0129629.

Thompson MG, Blake-Hedges JM, Cruz-Morales P, Barajas JF, Curran SC, Eiben CB, Harris NC, Benites VT, Gin JW, Sharpless WA, et al. 2019. Massively parallel fitness profiling reveals multiple novel enzymes in Pseudomonas putida lysine metabolism. MBio 10.

Thompson MG, Incha MR, Pearson AN, Schmidt M, Sharpless WA, Eiben CB, Cruz-Morales P, Blake-Hedges JM, Liu Y, Adams CA, et al. 2020. Fatty acid and alcohol metabolism in Pseudomonas putida: functional analysis using random barcode transposon sequencing. Appl. Environ. Microbiol. 86.

Wetmore KM, Price MN, Waters RJ, Lamson JS, He J, Hoover CA, Blow MJ, Bristow J, Butland G, Arkin AP, et al. 2015. Rapid quantification of mutant fitness in diverse bacteria by sequencing randomly bar-coded transposons. MBio 6:e00306-15.