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

Megan E. Garber[1, 2], Vered Frank[2], Alexey E. Kazakov[3], Hanqiao Zhang[2,4], 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. Lawrence Berkeley National Laboratory, Environmental Genomics and Systems Biology Division
4. University of California, Berkeley Department of Bioengineering

* Correspondence: amukhopadhyay@lbl.gov

Abstract

Functional diversity in bacteria is introduced by lineage specific expansion or horizontal gene transfer (HGT). Using modular bacterial signaling systems as a template, we experimentally validate domain swapping of modular proteins as an extension of the HGT model. We take a computational approach to explore the domain architecture of two-component systems (TCS) in select *Pseudomonads*. We find a transcriptional effector domain swap that reconstructed a duplicated sigma54-dependent TCS to a sigma70-dependent TCS. Through functional genomics approaches, we determine that the implicated TCSs are involved in consumption of short-chain carboxylic acids. We verify the relationship between the domain-swapped TCSs utilizing a mutational screen, in which we switch the specificity of the sigma70-dependent TCS output to the sigma54-dependent TCS input, and vice versa. Our findings suggest that this domain swap was

maintained throughout α-, β-, γ- proteobacteria, thus domain swapping has potential to lead to fitness advantages and 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

Bacteria use two-component systems (TCS) to sense and respond to signals in their environments. In a canonical TCS, a histidine kinase (HK) is auto-phosphorylated by ATP, upon stimulation (Sankhe et al., 2018). The phosphorylated HK can then engage in a phosphotransfer event to its cognate response regulator (RR) (Zschiedrich et al., 2016). Although sequences encoding TCSs can be highly redundant in a single bacterial genome (Galperin, 2005; Galperin et al., 2010; Grebe and Stock, 1999; Jung et al., 2012; Wuichet et al., 2010), biochemical and genetic evidence suggests that interactions between HKs and RRs are conserved between cognate pairs. Coevolutionary events of duplication and divergence have led to the apparent orthogonality of cognate pairs of HKs and RRs (Capra and Laub, 2012; Capra et al., 2012; Choi and Kim, 2011; Laub and Goulian, 2007; Laub et al., 2007; McClune and Laub, 2020; Podgornaia and Laub, 2013; Salazar and Laub, 2015; Skerker et al., 2008).

HKs and RRs are both modular proteins consisting of constant interacting domains, such as histidine-phosphotransfer (HPt), specifically HisKA, and catalytic ATPase (CA) domains in the HK and receiver (REC) domain in the RR (Figure 1a,b). Importantly, TCSs with HisKA domains are usually highly insulated (McClune and Laub, 2020), while HKs from other families, for instance with HisKA_2 or HWE domains, are otherwise promiscuous (Herrou et al., 2017; Lori et al., 2018).

46 HKs and RRs are often found accessorized with a variety of N-terminal input or C-terminal effector
47 domains (Dutta et al., 1999; Galperin, 2005, 2010; Kim et al., 2010; Ortega et al., 2017; Padilla-
48 Vaca et al., 2017). The input domains in HKs typically encode for sensing functionality, whereas
49 the appended domains in RRs confer output function. In the context of the coevolution of TCS,
50 the domains tethered to a pair of HKs and RRs are typically carried through to descendants during
51 the processes of lineage specific evolution. However, hypotheses of domain swapping in bacteria
52 have emerged as a potential mechanism for the diversification of TCSs (Alm et al., 2006; Capra
53 and Laub, 2012; Forslund et al., 2019; Laub and Goulian, 2007).

54 In a domain swap the core parts of the TCS (HPt, CA and REC domains) remain intact,
55 while the variable input or output domain is replaced by a different domain (Alm et al., 2006;
56 Forslund et al., 2019). Domain swaps are hypothesized to occur via horizontal gene transfer
57 (HGT) and homologous recombination (Forslund et al., 2019). A domain swap in a TCS can
58 change the input or sensing parts of a HK diversifying the molecular inputs for the cascade (Dutta
59 et al., 1999; Ortega et al., 2017), or it can alter the output domain of a RR modifying the function
60 of the entire signalling cascade (Galperin, 2010).

61 The output parts of RRs, the effector domains, dictate the final effect of a signaling
62 cascade. RRs can harbor output domains that lead to chemotactic responses (Briegel et al., 2009;
63 Lai and Parkinson, 2018), regulation of small molecules such as cyclic nucleotides (Ryjenkov et
64 al., 2005), or transcriptional regulation (Galperin, 2010; Wuichet et al., 2010). Transcriptional
65 effector domains (TEDs) interact primarily with DNA in a sequence-specific manner to regulate
66 the transcription of a gene. TEDs are diverse in structure, protein-DNA interaction, protein-protein
67 interaction, and function, and are therefore binned into separate sub-classes or families (Galperin,
68 2010). In this study we predominantly observe TEDs associated with RRs that confer activation
69 of genes with sigma54 sigma factors (AAA^+ tethered to HTH_8, NtrC family), and regulation of
70 genes with sigma70 sigma factors (GerE, NarL family or Trans_reg_C, OmpR family) (Figure 1c).

71 We take a computational approach to explore the domain architecture of RRs in a select
72 set of *Pseudomonads*, with the goal of identifying evolutionarily driven domain swaps. All
73 computational predictions are then substantiated by original experimental results. Our findings
74 lead us to observe a domain swapping event that altered the sigma factor dependence of a cluster
75 of carboxylic acid responsive sigma54-dependent, NtrC-like RRs native to proteobacteria.

76 Results

77 *Phylogenetic analysis of TCS domains from Pseudomonas RRs reveals TED swap*

78 We identified all RRs in five representative *Pseudomonads*: *P. aeruginosa* PAO1, *P.*
79 *stutzeri* RCH2, *P. putida* KT2240, *P. fluorescens* FW300-N2E2, and *P. fluorescens* FW300-N2C3.
80 We subdivided these RRs by their output domains, taking into account only RRs with TEDs
81 (Supplementary Table 1). The phylogenetic tree of the REC domains importantly agrees with a
82 previously published tree of *P. aeruginosa* PAO1 REC domains (Chen et al., 2004). Our results
83 demonstrate that the REC domains of RRs do not clade by species, but instead clade by TEDs
84 (Figure 2, left panel). This observation is consistent with the current understanding how highly-
85 redundant proteins evolve in bacteria (McClune and Laub, 2020; Voordeckers et al., 2015). A
86 domain swap can be hypothesized when RRs of one family cluster with RRs of another. Such an
87 event can be observed within the AAA^{+}-HTH_8/NtrC-like cluster (red and green) of the REC
88 phylogenetic tree, where REC domains with an alternative TED, GerE/NarL-like (blue) are present
89 (Figure 2, left panel).

90 To test a coevolutionary hypothesis between HKs and RRs, we applied the same
91 phylogenetic strategy to sequences of HPt-CA (HisKA-HATPase_C) domains of all of the HKs in
92 the same *Pseudomonas* strains. We observed that the leaves of the HPt-CA tree predominantly
93 cluster by the domain architecture of their cognate RRs (Figure 2, right panel). We noted that the
94 cognate HKs for the major clade of NarL-like RRs are hybrid HKs (grey bars), whereas the

95 cognate HKs of the NarL-like RRs within the swapped cluster are more similar in domain
96 architecture to the Ntr-like HKs. These results suggest that our initial observation and hypothesis
97 of a functional domain swap was probable.

98 We next wanted to determine the evolutionary distance of the domain swap, asking
99 whether it was present in *Pseudomonads* only, or all proteobacteria. Using representatives from
100 α-, β-, γ-, and δ- proteobacteria, selected for their high proportions of NtrC-like and NarL-like RRs,
101 we were able to recapitulate the results of the *Pseudomonas* REC domain-based phylogenetic
102 tree. The proteobacteria REC phylogenetic tree reveals that the swap is found in the select strains
103 of α-, β-, and γ- proteobacteria, but not observed in the selected δ-proteobacteria (Supplementary
104 Figure 1). This discrepancy could be the result of gene loss in the selected representatives, or it
105 could indicate that the swapped clade never existed in δ-proteobacteria. Notably, in the new view
106 of the tree of life *(Hug et al., 2016)*, it is hypothesized that δ-proteobacteria are more distantly
107 related to the rest of the proteobacterial clade than previously thought. Together these results
108 lead to the hypothesis that a domain swap likely occurred at the cuspis of the proteobacteria (α-,
109 β-, γ-) clade.

110

111 *Determining the Functional Similarity of the RRs in the Swapped Cluster*

112 While the phylogenetic analysis can be used to hypothesize a proteobacteria-specific
113 domain swap, it does not inform on the functionality of the representatives. Closer observation of
114 the swapped cluster reveals that there are four subclusters, three of which harbor AAA^+ domains
115 (NtrC-like/sigma54-dependent) and one of which harbors a GerE domain (NarL-like/sigma70-
116 dependent) (Figure 2 (close-up panel)). We hypothesized that if the domain swap is real, then the
117 signal cascades of the representatives within the clade should have similar functional inputs or
118 outputs. We focused our efforts on the functions of the TCSs in *P. putida* KT2440, because it is a

119 well studied model strain with a readily available genetic toolkit to enable follow-up and validation
120 studies.

121 Because the domain swap occurs within the NtrC-like cluster, we were able to hypothesize
122 that the signaling cascades were likely involved in nutrient uptake of carbon or nitrogen sources
123 (Cases et al., 2003; Hervás et al., 2009; Leech et al., 2008; Nishijyo et al., 2001). We also used
124 the annotations of the representatives within the swapped cluster that had been previously
125 identified as amino acid or di-carboxylic acid sensors (Cases et al., 2003; Lundgren et al., 2014;
126 Sonawane et al., 2006; Tatke et al., 2015). However the function of the NarL-like representatives
127 were unknown. We therefore applied high-throughput functional genomics to explore the input
128 signals and output genes of all of the relevant TCSs represented within the cluster.

129 We applied an automated pipeline of DNA-affinity purification tethered to next generation
130 sequencing (DAP-seq), commonly used to biochemically determine genomic sites of protein-DNA
131 interaction for transcription factors like RRs (Garber et al., 2018; Rajeev et al., 2020)
132 (Supplementary Figure 2a), to the entire subset of RRs in the swapped cluster. We found high-
133 confidence DNA binding targets for all of the NtrC-like RRs (Supplementary Figure 2b,
134 Supplementary Data 1,2). Aligning the high-confidence targets from all of the NtrC-like RRs
135 enabled us to manually assign a binding motif to the homologs. Using the motifs identified from
136 the high-confidence homologous hits, we were able to query each genome for additional DNA
137 binding targets (Supplementary Figure 2b, Figure 3a,c). This analysis enabled us to hypothesize
138 gene targets for the NtrC-like TCSs PP_1066, PP_0263, PP_1401 to be upstream of PP_2453,
139 PP_1188, and PP_1400 respectively. For all predicted binding targets see Supplementary Table
140 2. As we were only able to find genomic targets for AO356_22615 and AO356_25435 from
141 *Pseudomonas fluorescens* FW300-N2C3 (Supplementary Figure 2b, Supplementary Data 1,2),
142 we were unable to identify a binding motif or a conserved genetic output for the NarL-like RRs by
143 this method.

144 To determine the input signals for the cluster, we utilized a dataset of randomly barcoded
145 transposon libraries for next generation sequencing (RB Tn-seq) of *P. putida* KT2440 grown with
146 single carbon sources (Thompson et al., 2020a). The genes with low fitness scores (fitness < -2)
147 under a given growth condition have condition-specific essentiality (Price et al., 2018; Wetmore
148 et al., 2015). We observed that the gene clusters harboring the RRs of interest each had low
149 fitness scores for growth with either glutamic acid, succinic acid, a-ketoglutaric acid, or butyric
150 acid (Figure 3a). We speculated that genes and their corresponding TCSs with synonymous
151 fitness values represented a signaling cascade where the input signal was the carbon source.
152 Using this dataset, we proposed that the NtrC-like RRs, PP_1066, PP_0263, and PP_1401 are
153 regulated by glutamic acid (Sonawane et al., 2006), succinic acid, and a-ketoglutaric acid
154 (Lundgren et al., 2014; Tatke et al., 2015) respectively. We also hypothesized that the NarL-like
155 RR, PP_3551, is regulated by butyric acid. As we were not able to identify binding sites or a DNA
156 binding motif for PP_3551, we relied on the gene cluster and fitness data to hypothesize an output
157 for the butyric acid responsive TCS (Thompson et al., 2019, 2020b). The neighboring gene
158 PP_3553 demonstrated a low fitness score consistent with the TCS, making it a good candidate
159 for the signaling cascade's output.
160 From the above results we were able to generate hypotheses for the signals and the
161 regulated genes of the TCSs in question. To validate these hypotheses, we generated GFP
162 reporter strains by tethering the upstream regions (~200bp) of the hypothesized output genes to
163 a GFP coding sequence on a broad host range plasmid. The *P. putida* KT2440 reporter strains
164 responded to the proposed signals with increased fluorescence above background (Figure 3b).
165 In concert with our predictions, when the reporter strains were stimulated in genetic backgrounds
166 absent of the corresponding transcriptional RR, fluorescence response was ablated (Figure 3b).
167 Interestingly, we observed that the WT strain bearing the p2453 reporter plasmid was slightly
168 activated by cultivation in butyric acid, indicating either sensing promiscuity by PP_1066's cognate
169 HK, PP_1067, or cross-talk with PP_3552, the cognate HK for PP_3551. We also observed that

170 when p3553 was grown with glutamic acid, fluorescence decreased below the minimal media (-)
171 levels, leading us to hypothesize that there is another component either within or external to the
172 identified TCSs that downregulated GFP expression.

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

179 *Exploring the sequence space of the RRs in the Swapped Cluster*

180 We next explored whether the CarSR TCSs were closely related using a coevolution
181 hypothesis, which dictates that if HK-RR pairs interact then the interacting residues should
182 coevolve across orthologs. In this view, closely related TCSs share more sequence space than
183 do TCSs that are distantly related (Capra and Laub, 2012; McClune and Laub, 2020). We applied
184 co-variance analysis (Bakan et al., 2011, 2014) to HPt-CA-REC domains of cognate HK-RR pairs
185 to identify co-varying residues (Supplementary Figure 3) between HPt-CA and REC domains. Our
186 results are consistent with previous studies (Laub et al., 2007; Skerker et al., 2008), where the
187 first alpha helix of the REC domain, previously shown to share points of contact with HPt domain
188 (Jacob-Dubuisson et al., 2018), contained the highest scoring residues. Importantly, the CarR
189 cluster is well-insulated within the broader NtrC-like, sigma54-dependent, cluster of the REC
190 phylogenetic tree (Figure 2b,c), indicating that its closest relatives are NtrC-like. We hypothesized
191 that if the NarL-like and NtrC-like RRs in the CarR subcluster could be engineered to switch
192 specificity, despite their differences in TEDs and sigma factor preference, then it was more likely
193 that the RRs in the cluster share REC sequence space.

194 To test if we could switch the specificity of the extra-familial CarR RRs, we applied site
195 directed mutagenesis at the positions of the predicted interacting sites to both PP_1066 (hereafter
196 referred to as RR_g) and PP_3551 (hereafter referred to as RR_b) (Figure 4a). Native or mutant RRs
197 (indicated by *) were expressed under the control of an arabinose inducible system (pBAD) and
198 GFP was driven by either the glutamic acid or butyric acid responsive promoters, hereafter
199 referred to as pRR_g (p2453) and pRR_b (p3553). To validate the use of the single plasmid system,
200 we complemented the ΔRR_g or ΔRR_b background strains with heterologous expression of RR_g or
201 RR_b. We determined that both of the heterologous RRs can respond to their expected signal in a
202 dose-dependent manner (Supplementary Figure 4). To ensure that there was no interference
203 between the two TCSs, the specificity-switch experiments were performed in double knockout
204 (dKO) background strains ($\Delta RR_g \Delta RR_b$). We observed leaky levels of GFP expression from native
205 RR_g-pRR_g in the dKO background in minimal media (-) (Figure 4b). In contrast, GFP expression
206 under the control of native RR_b-pRR_b in minimal media was tightly regulated (Figure 4b).
207 Consistent with our previous results (Figure 3b), we observed increased levels of GFP expression
208 when RR_g-pRR_g was grown with either glutamic acid and butyric acid (Figure 4b). We also
209 observed an increase in fluorescence when RR_b-pRR_b was grown with butyric acid, however we
210 did not observe the expected decrease in baseline fluorescence (Figure 3B) when RR_b-pRR_b was
211 grown with glutamic acid. When cultivated in minimal media with or without the addition of glutamic
212 acid, mutant RR_b*-pRR_b behaved like RR_g-pRR_g, demonstrating the characteristic leaky
213 expression in minimal media, and an increase in GFP expression in glutamic acid. However, when
214 grown in butyric acid, expression increased above the levels of growth in glutamic acid (Figure
215 4b). These results suggest that RR_b* interacts with HK_g (PP_1067), but might also interact with
216 HK_b (PP_3552) when grown in butyric acid. GFP expression driven by RR_g*-pRR_g was leakier
217 than its counterpart, RR_b-pRR_b, which could be the result of either expression driven by a different
218 promoter with different dynamics or from interactions at a higher level of regulation. RR_g*-pRR_g
219 demonstrated the expected decrease in GFP expression below baseline observed in previous

220 experiments (Figure 3b) under glutamic acid growth conditions. Strikingly, RR_g*-pRR_g shows a
221 statistically significant increase in GFP expression when grown with butyric acid, suggesting that
222 RR_g* switched its specificity from HK_g to HK_b. Taken together, these results show, empirically,
223 that gradual changes in critical specificity determining residues made over time could be
224 responsible for the differentiation of the CarSR TCSs. Furthermore they demonstrate biologically
225 relevant evidence that these distinct extra-familial TCSs with different sigma factor dependencies,
226 the CarR RRs, are closely related.
227

228 *Interrogating an alternative hypothesis*

229 As an alternative to our working hypothesis, we conceived that if bacteria from other
230 distantly related phyla explored similar sequence space, the CarR subcluster could be derived
231 from HGT alone, and not a domain swap. To test this hypothesis, we queried the configuration of
232 RR families in bacterial phyla with high counts of NarL-like and NtrC-like RRs. Bacterial and
233 archaeal TCSs have been previously curated in a TCS census by Galperin et al. (Galperin, 2006,
234 2010). By querying this dataset, we were able to find representative strains from Firmicutes,
235 Bacteroides, Acidobacteria, δ-Proteobacteria, and γ-Proteobacteria (*P. Putida* KT2440) that have
236 high counts of both NarL-like and NtrC-like RRs in their genomes (Supplementary Figure 5a). To
237 test our alternative hypothesis, we identified all RRs and their corresponding REC domains from
238 the representative species, made a REC domain-based phylogenetic tree, and mapped onto it
239 the corresponding RR family (Supplementary Figure 5b). We observed a similar configuration to
240 pseudomonas and proteobacteria trees (Figure 2b, Supplementary Figure 1), indicating that the
241 domain architecture we observe in present day bacteria is as ancient as the bacterial domain of
242 life, and the alternative hypothesis of HGT fails.
243

244 *Observation of HK domain architecture in related TCS*

245 On closer observation of the cognate HKs to RRs in the carboxylic acid sensing subcluster,
246 we found that the cognate HKs of the NtrC-like RRs are transmembrane bound with a d_Cache
247 domain and the cognate HKs to the NarL-like are cytosolic with a PAS domain. While both of
248 these sensing domains directly interact with small molecule ligands, they likely did not evolve
249 through lineage specific expansion. To reconcile the discrepancy, we hypothesize that the HK
250 domain architecture observed in the pseudomonas carboxylic acid sensing subcluster is the result
251 of domain loss, where a common ancestor harbored both a periplasmic d_Cache domain and a
252 cytosolic PAS domain (Supplementary Figure 6). We anticipate that a closely related TCS might
253 exist in nature with the predicted domain architecture.

254 Discussion

255 In this work, we discover and provide evidence for a carboxylic acid sensing NtrC-like TCS
256 that underwent a domain swap, altering its sigma factor dependence (Figure 5). Our proposed
257 model assumes that an ancestral HK had promiscuous d_Cache and PAS carboxylic acid sensing
258 domains. Consistent with our characterization that the TCS sense carboxylic acids, both d_Cache
259 and PAS domains have been observed to interact directly with small molecules (Brewster et al.,
260 2016; Gavira et al., 2020; Henry and Crosson, 2011). We further propose that when the domain
261 swap occured, the output gene was promiscuous within carboxylic acid metabolism and could
262 moonlight a beneficial function (Pougach et al., 2014). The *in vivo* mutational screen provides
263 evidence that insulated, extra-familial RRs within the same species share sequence space and
264 are closely related. In a previous study, an extra-familial specificity swap of an Omp family HK
265 enabled interaction with an Ntr family RR in *C. crescentus*, leading the authors to hypothesize
266 that a duplicated Ntr family TCS may have overlapped in sequence space with an Omp family
267 TCS (Capra et al., 2012). Their results demonstrated how lineage-specific evolution of a

268 duplicated TCS facilitated insulation from unwanted cross-talk. In contrast, our work highlights
269 how domain swapping can alter the output TED (and classification) of an already well-insulated
270 TCS.

271 NtrC-like RRs require sigma54 transcription factors to activate genes in their regulons,
272 whereas NarL-like RRs interact with sigma70. It is well documented that cellular levels of small
273 nucleotide, (p)ppGpp, in combination with the protein DksA heavily regulate RNA polymerase's
274 (RNAP) preference for sigma70 (housekeeping or extracytoplasmic function (ECF) family)
275 (Casas-Pastor et al., 2019; Lee et al., 2012; Lonetto et al., 1998) or sigma54 (Bernardo et al.,
276 2006, 2009; Dalebroux and Swanson, 2012; Jurado et al., 2003; Ronneau and Hallez, 2019;
277 Wigneshweraraj et al., 2008). While the exact environmental conditions needed for
278 pseudomonads to reach the appropriate levels of (p)ppGpp and DksA for sigma54 occupation of
279 RNAP are unknown (Potvin et al., 2008; Ronneau and Hallez, 2019; Shingler, 2011), it can be
280 reasoned that the switch is lifestyle dependent. We find the functional switch, resulting from the
281 identified domain swap, to be especially notable, because the TCSs implicated in this study were
282 found to be metabolite responsive systems, and could therefore play a major role in survival and
283 fitness. We speculate that the capability to consume carboxylic acids without the regulatory
284 constraints dictated by sigma factor dependence, may have given a fitness advantage to the
285 ancestral organism in which the swap originated.

286 Domain swapping in general is an important biological phenomenon that can impact all
287 domains of life (Forslund et al., 2019). Complicating our interpretation of bacterial evolution, HGT
288 is known to mobilize genes or entire gene clusters that can be subsequently integrated into foreign
289 genomes (Bellieny-Rabelo et al., 2020; Linsky et al., 2020; Liu et al., 2017; Price et al., 2008;
290 Treangen and Rocha, 2011; Wu et al., 2011). The narrative of HGT becomes even more entwined
291 when we consider domain swapping. In this study we applied a simple and straightforward
292 strategy to identify domain swapping in a well-studied, modular signalling system. While we
293 focused our attention on a single domain swap, our results (Figure 2, Supplementary Figures 1,

294 6) and results of previous studies (Alm et al., 2006; Grebe and Stock, 1999) suggest that domain
295 swapping in TCS is prevalent. Our strategy, however, is not limited to signalling systems. Given
296 the potential ubiquity of domain swapping, we propose that this strategy can be applied to other
297 complex modular proteins and systems to detangle their intricate evolutions and cryptic functions.

298 Nature's pre-engineered enzymes and proteins provide attractive parts for practical
299 applications (Smanski et al., 2016; Way et al., 2014) and potential answers to plug-and-play
300 biology via synthetic domain swapping (Barajas et al., 2017; Maervoet and Briers, 2017). One
301 key and critical component of synthetic domain swapping is identifying domain boundaries for a
302 seamless swap (Barajas et al., 2017; Rhodius et al., 2013; Schmidl et al., 2019). In a previous
303 study a library based approach was implemented to identify domain boundaries between REC
304 domains and TEDs for *E. coli* RRs, however their successes were only shown for inter-family TED
305 swaps (Schmidl et al., 2019). In this study, we demonstrate that nature devised the appropriate
306 domain boundary between a REC domain and an extra-familial TED to engineer a functional
307 signaling cascade. Such evolutionarily driven domain swaps could be an untapped source for
308 identifying optimal domain boundaries in synthetic biology applications.

309 Taken together, this work establishes and validates domain swapping as a mechanism for
310 neofunctionalization of modular genes. By applying hypotheses grounded in the established
311 theories of evolution of TSCs, we were able to link an orphaned sigma70-dependent TCS to a
312 well-described family of sigma54-dependent TCSs. Our results validate the close relationship
313 between systems that would have otherwise been described as distantly related. Ultimately, our
314 work provides evidence for an extension to the current model of neofunctionalization to include
315 functional part sharing via domain swapping.

316 Acknowledgements

317 We would like to thank the following individuals for their contribution to this body of work for either
318 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, and Matt Incha; or providing expertise toward the development of high-throughput cloning, protein purification, and NGS techniques, Nurgul Kaplan, Jennifer Chiniquy, Joel M. Guenther, and Brett Garabedian.

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.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. A.M. provided resources supervision and support. All authors reviewed and edited the final draft.

Declarations of Interest

The authors declare no competing interests.

Figure and Figure Legends

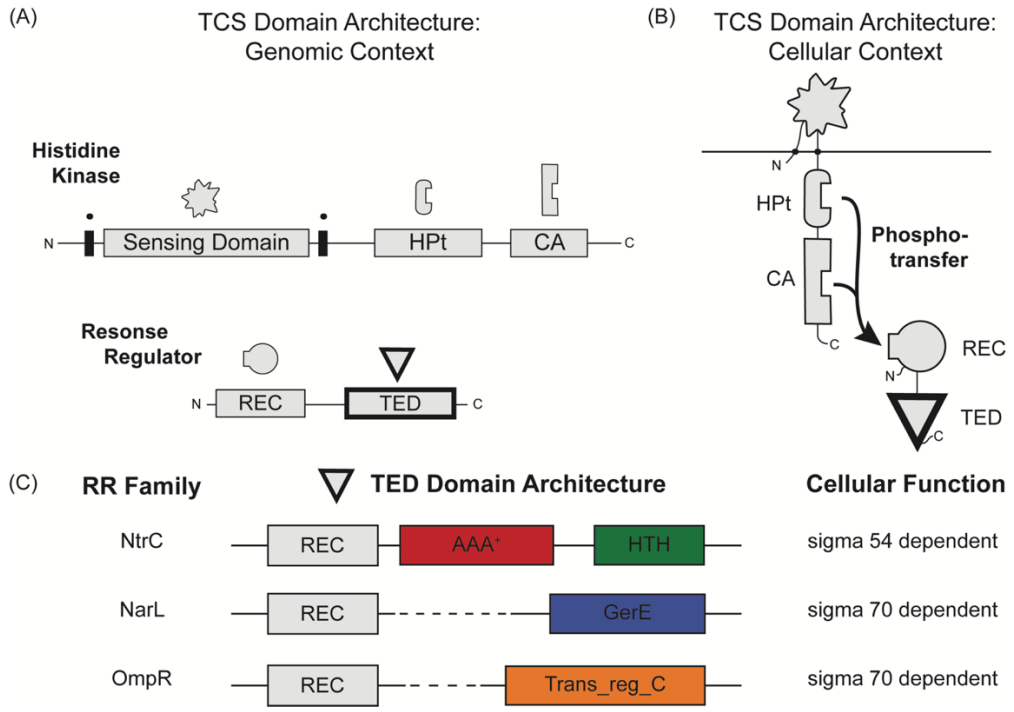


Figure 1: ***Domain architecture of TCSs with transcriptional effector domains (TEDs):*** Modularity of canonical TCS in its genomic (A) and cellular (B) contexts. (C) RR families designated by the domain architecture of transcriptional effector domains (TEDs) found in TCS relevant to this study annotated by their cellular function.

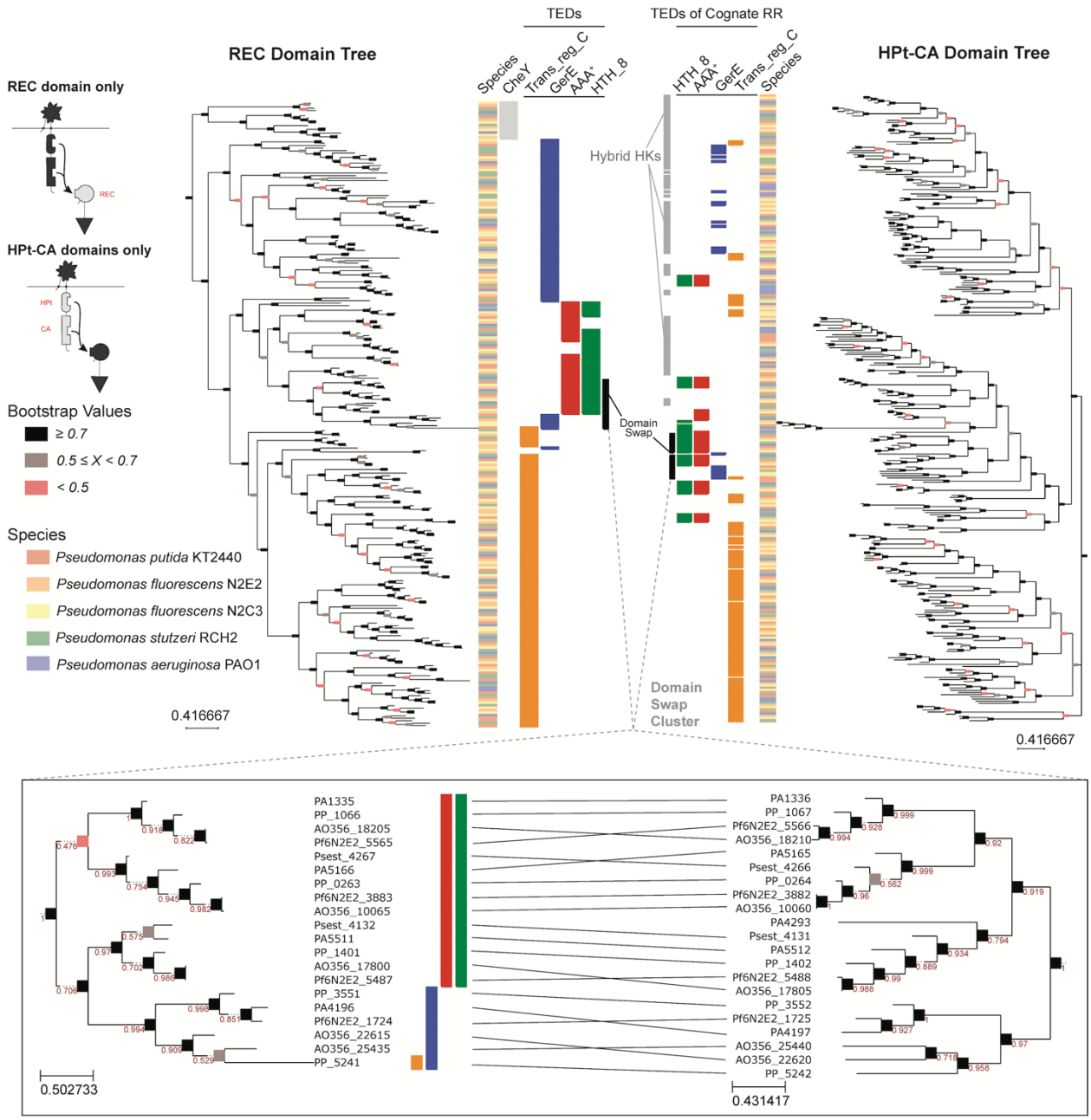
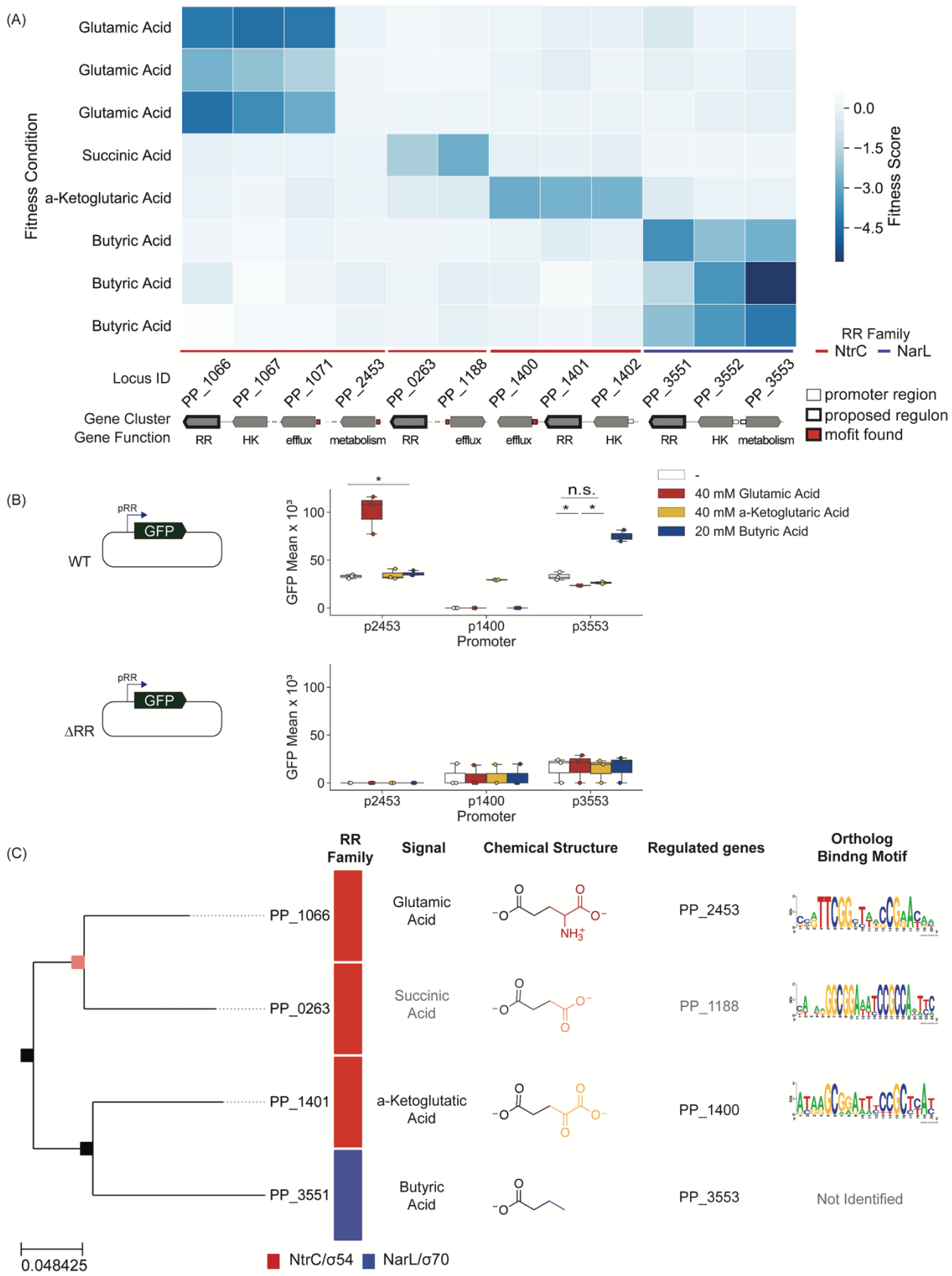


Figure 2: **Phylogenetic analysis of TCS domains from *Pseudomonas* RRs reveals TED domain swap:** Phylogenetic trees of REC (left) and HPt-CA (right) domains of RRs and HKs from select Pseudomonads. Trees are annotated with species and transcriptional effector domain (TED) at the leaves. Hybrid HKs 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)

357 domains trees annotated with protein name and/or TED. Lines are drawn to match predicted

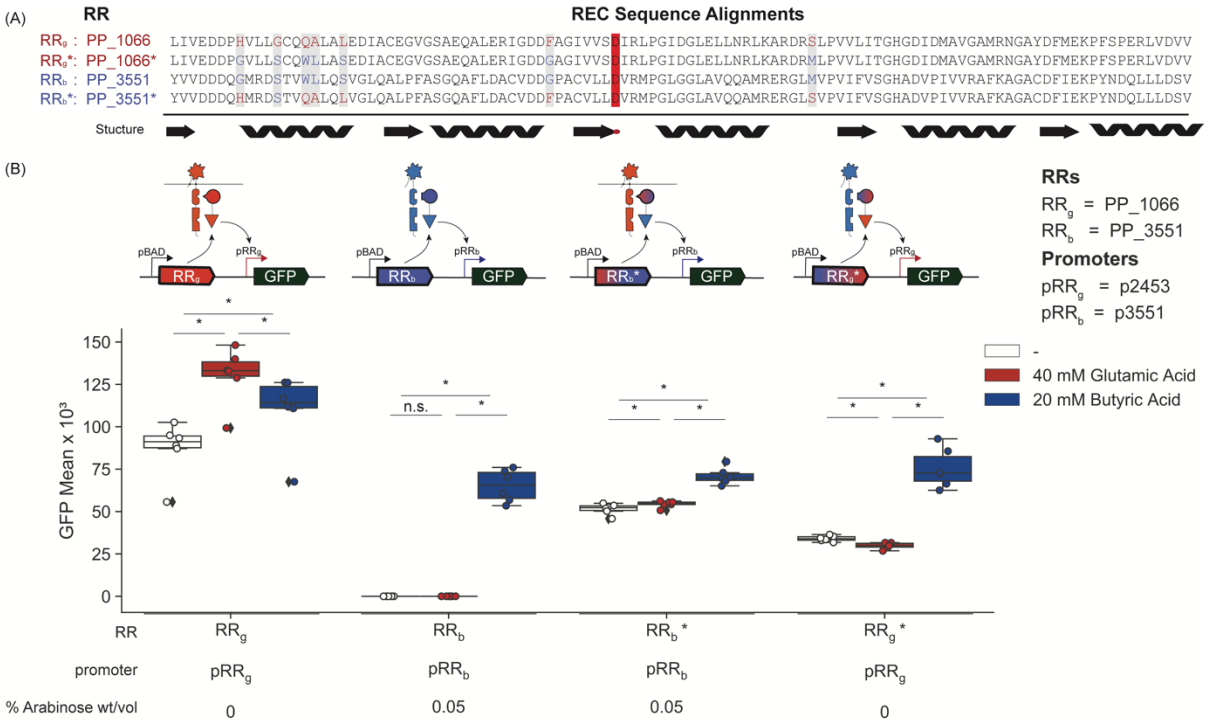
358 cognate HK-RR pairs.


(A)
Fitness Condition
Glutamic Acid
Glutamic Acid
Glutamic Acid
Succinic Acid
a-Ketoglutaric Acid
Butyric Acid
Butyric Acid
Butyric Acid
Fitness Score
0.0
−1.5
−3.0
−4.5
RR Family
NtrC
NarL
Locus ID
PP_1066
PP_1067
PP_1071
PP_2453
PP_0263
PP_1188
PP_1400
PP_1401
PP_1402
PP_3551
PP_3552
PP_3553
Gene Cluster
Gene Function
RR
HK
efflux
metabolism
RR
efflux
efflux
RR
HK
RR
HK
metabolism
promoter region
proposed regulon
mofit found
(B)
pRR
GFP
WT
ΔRR
GFP Mean x 10³
100
50
0
p2453
p1400
p3553
Promoter
*
n.s.
-
40 mM Glutamic Acid
40 mM a-Ketoglutaric Acid
20 mM Butyric Acid
(C)
RR Family
Signal
Chemical Structure
Regulated genes
Ortholog Bindng Motif
PP_1066
PP_0263
PP_1401
PP_3551
Glutamic Acid
Succinic Acid
a-Ketoglutatic Acid
Butyric Acid
PP_2453
PP_1188
PP_1400
PP_3553
Not Identified
0.048425
NtrC/σ54
NarL/σ70

359

360 Figure 3: ***Determining the functional similarity of the RRs in the swapped cluster***: (A)
361 Heatmap of fitness data for TCS gene clusters within the domain swapped cluster. A low fitness
362 score (fitness< - 2) indicates a gene has condition-specific essentiality. DAP-seq results are
363 briefly summarized with a red box upstream of the gene in which an orthogonal DNA binding motif
364 was identified. (B) GFP reporter plasmids in WT and ΔRR backgrounds were grown in minimal
365 media with glucose plus a second carbon source - glutamic acid, a-ketoglutaric acid or butyric
366 acid. Fluorescence was measured by flow cytometry and measurements are reported as GFP
367 Mean x 10^3after gating. Statistical significance was determined between conditions by t-test (* =
368 p-value < 0.05, n.s. = not significant). (C) Summary of findings from functional genomics studies.
369 The relationship of the *P. putida* TCSs is shown by a pruned phylogenetic tree to the left, and
370 results, RR family, validated sigal, signal chemical structure, validated regulated genes,
371 orthogonal binding motif identified by DAP-seq are drawn as a table to the right. If the findings
372 were not validated by reporter assay, the results are displayed in grey.
373

374



375

376 Figure 4: ***Exploring the sequence space of the RRs in the swapped cluster:*** (A) Sequence
377 alignments of native RRs PP_1066 (RR_g), PP_3551 (RR_b), and mutated RRs PP_1066* (RR_g*),
378 PP_3551* (RR_b*). Residues with high co-variance scores > 1.1 are marked in a grey background
379 and red text for RR_g-derived residues or blue text for RR_b-derived residues. Structural information
380 for the REC domain is shown below the sequences, where arrows are β-barrels and ribbons are
381 α-helices. (B) RR_g, driving glutamic acid responsive promoter (pRR_g) switches its specificity to
382 butyric acid responsive when mutated to RR_g*. RR_g, driving butyric acid responsive promoter
383 (pRR_b) partially switches its specificity to glutamic acid responsive when mutated to RR_b* .
384 Statistical significance was determined between conditions by t-test (* = p-value < 0.05, n.s. = not
385 significant).

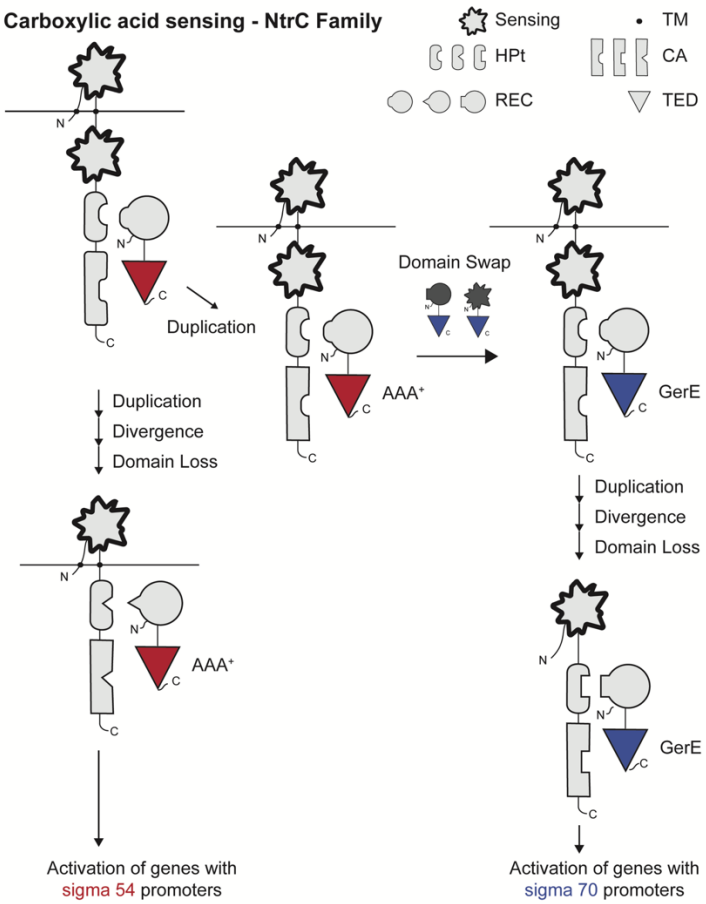


Figure 5: ***Proposed route for evolution of carboxylic acid sensing TCS with alternative sigma factor dependence*:** Within an ancestral host, an Ntr-like TCS with carboxylic acid sensing capabilities was duplicated. A transcription factor with a GerE DNA binding domain recombined with the duplicated gene to form an active and functional Ntr-like TCS with a GerE DNA binding domain. Both TCSs encoded in the same genome undergo duplication, divergence, and domain loss to achieve insulated pathways with insulated signal detection. Each system in its modern form can detect unique carboxylic acids and activate unique sets of genes. The key difference between the related systems is they interact with different types of promoters either sigma54 or sigma70, which are known to be active under different environmental constraints. We hypothesize that the ancestral strain in which the domain swap originated might have benefited from a lack of environmental constraints against consuming available environmental resources.

398 # Methods

399 **Phylogenetic analysis of REC and HPt-CA trees**

400 RRs and HKs were identified with hmmsearch from HMMER v3.1b2 (Mistry et al., 2013) to search
401 for all proteins with Response_reg (PF00072) and HisKA (PF00512) domains. Alternatively, HKs
402 and RRs were identified with the Microbial Signal Transduction database (MISTDB) (Gumerov et
403 al., 2020). Using HMMER again, we queried the domain architecture of the signaling proteins and
404 selected for RRs with HTH_8, Trans_reg_C, AAA^+, GerE, HTH_18 (AraC) domains. REC and
405 HPt-CA domain sequences were extracted from whole sequences based on coordinates
406 determined by the hmmsearch. Domains were aligned using the MAFFT-LINSI algorithm from
407 MAFFT v7.310 (Katoh and Standley, 2013). Phylogenetic trees were constructed using FastTree
408 2 (Price et al., 2010). Trees were visualized and annotated using python based ETE3 (Huerta-
409 Cepas et al., 2016).

410 **Automated DNA affinity purification - seq**

411 *DNA preparation for NGS*

412 Pseudomonas isolates were cultured in either LB or minimal media (see Supplementary Table 3
413 for strain specific minimal media recipes). Genomic DNA was purified with a promega wizard
414 genomic preparation kit (Promega, Madison, WI). DNA was sheared with covaris miniTUBE
415 (Covaris, Woburn, MA) to an average size of 200 bp. The DNA quality was confirmed by
416 Bioanalyzer high sensitivity DNA kit (Agilent, Santa Clara). Sheared DNA was then adapter-
417 ligated (AL) with NEBnext Ultra ii Library Preparation kit (New England Biolabs, Ipswich, MA). AL-
418 DNA quality was again confirmed by Bioanalyzer high sensitivity DNA kit (Agilent, Santa Clara).
419 AL-DNA was stored at -20˚C until required for downstream use.
420

421 *Expression Strain Design*

422 Pet28 expression vectors with N-terminal 6x-His-tagged RRs were cloned by Gibson assembly
423 (Gibson et al., 2009). Plasmid design was facilitated by j5 DNA assembly design (Hillson et al.,
424 2012) (diva.jbei.org), see Supplementary Table 4 for primers.

425

426 *Automated DNA affinity purification*

427 Quadruplicates of expression strains were grown in autoinduction media (Zyp-5052 (Studier,
428 2005)) at 37˚C, 250 RPM, for 5-6 hours and then transferred to grow at 17˚C, 250 RPM, overnight.
429 Cell pellets were harvested and lysed at 37˚C for 1 hour in a lysis buffer (1X TBS, 100 μM PMSF
430 (Millipore Sigma, Burlington MA), 2.5 units/mL Benzonase nuclease (Millipore Sigma, Burlington
431 MA), 1 mg/mL Lysozyme (Millipore Sigma, Burlington MA)). Lysed cells were then clarified by
432 centrifugation at 3214 x g and further filtered in 96-well filter plates by centrifugation at 1800 x g.
433 To enable high-throughput processing, protein-DNA purification steps were performed with IMAC
434 resin pipette tips (PhyNexus, San Jose, CA) using a custom automated platform with the Biomek
435 FX liquid handler (Beckman Coulter, Indianapolis, IN). The expressed RRs were individually
436 bound to metal affinity resin embedded within the IMAC resin pipette tips and washed in a wash
437 buffer (1X TBS, 10 mM Imidazole, 0.1% Tween 20). The bead bound RRs were then mixed with
438 60μL of DNA binding buffer (1X TBS, 10 mM magnesium chloride, 0.4 ng/μL AL-DNA, with or
439 without 50 mM acetyl phosphate (split into duplicates)). The protein bound to its target DNA was
440 then enriched in an enrichment buffer (1X TBS, 10 mM Imidazole, 0.1% Tween 20) and eluted in
441 an elution buffer (1X TBS, 180 mM Imidazole). The elution was stored at -20˚C for a minimum of
442 one day and up to a week before proceeding to the NGS library generation. See supplementary
443 methods for detailed protocol.

444

445 *NGS Library Generation*

446 3.2 µL of the elution from the previous step was added to 3.5 µL SYBR green ssoAdvanced
447 (Biorad, Hercules, CA) and 0.15 µL of each dual indexed NGS primers. NGS libraries were
448 prepared by following the protocols for fluorescent amplification of NGS libraries (Chiniquy et al.,
449 2020). Pooled libraries were sequenced by Illumina NovaSeq 6000 SP (100 cycles) (Illumina, San
450 Diego, CA).

451

452 *DAP-seq data analysis*

453 Sequenced reads were processed by a computational DAP-seq analysis pipeline as follows.
454 Adapters and low-quality bases were trimmed and reads shorter than 30 bp were filtered out using
455 Trimmomatic v.0.36 (Bolger et al., 2014). The resulting reads were checked for contamination
456 using FOCUS (Silva et al., 2014). Then the reads were aligned to the corresponding
457 *Pseudomonas* spp. genome using Bowtie v1.1.2 (Langmead et al., 2009) with –m 1 parameter
458 (report reads with single alignment only). Resulting SAM files were converted to BAM format and
459 sorted using samtools v 0.1.19 (Li et al., 2009). Peak calling was performed using SPP 1.16.0
460 (Kharchenko et al., 2008) with false discovery rate threshold of 0.01 and MLE enrichment ratio
461 threshold of 4.0. Enriched motifs were discovered in genome fragments corresponding to the
462 peaks using MEME (Bailey et al., 2009) with parameters –mod anr –minw 12 –maxw 30 –revcomp
463 –pal –nmotifs 1. Source code of the DAP-seq analysis pipeline is available at
464 https://github.com/novichkov-lab/dap-seq-utils.

465 For conserved RRs with small numbers of high-confidence peaks (1-2 per genome), binding
466 motifs were predicted manually by comparative genomics approach. Orthologous RRs were
467 identified by OrthoFinder2 (Emms and Kelly, 2019). For each of orthologous RRs, one genome
468 fragment corresponding to the peak with the highest enrichment value was selected for motif

469 search. Conserved motifs were discovered using the SignalX tool from GenomeExplorer package
470 (Mironov et al., 2000) with the “inverted repeat” option.

471

472 **Fitness Experiments**

473 Single carbon source fitness data is available at http://fit.genomics.lbl.gov (Thompson et al.,
474 2020a).

475 **Co-variance analysis**

476 Cognate RRs and HKs from *Pseudomonas* and *E. coli* strains were identified as pairs if they were
477 found neighboring each other in their respective genomes. HPt (HisKA), CA (HATPase_C), and
478 REC (Response_reg) domain boundaries were determined with hmmsearch from HMMER
479 v3.1b2 (Mistry et al., 2013). Fasta files of concatenated HPt-CA-REC domains from cognate and
480 randomized HK-RR pairs were aligned with the MAFFT-LINSI algorithm from MAFFT v7.310
481 (Katoh and Standley, 2013). Alignment files were then queried for coevolution with the ProDy *Evol*
482 suite (Bakan et al., 2011, 2014) in python and were plotted in a heatmap. The highest scoring
483 residues > 1.1 were used to inform hypotheses for specificity switch strains.

484 **GFP reporter strain generation and assays**

485 *Knockout strain generation*

486 1000bp homology fragments upstream and downstream of the target gene were cloned into
487 plasmid pKS18. Plasmids were then transformed into E. coli S17 and then mated into P. putida
488 via conjugation. Transconjugants were selected for on LB agar plates supplemented with 30
489 mg/ml kanamycin and 30 mg/ml chloramphenicol. Transconjugants were then grown overnight
490 on LB media and were then plated on LB agar with no NaCl that was supplemented with 10%
491 (wt/vol) sucrose. Putative deletions were screened on LB agar with no NaCl supplemented with

492 10% (wt/vol) sucrose and LB agar plate with kanamycin. Colonies that grew in the presence of
493 sucrose but had no resistance to kanamycin were further tested via PCR with primers flanking
494 the target gene to confirm gene deletion.

495 *GFP reporter strains*

496 Promoter boundaries for p2453, p1400, p3553 were identified as the region just upstream of the
497 gene's start codon up until the start or stop codon of the next nearest gene. The promoters were
498 cloned upstream of GFP on a broad host range plasmid with BBR1 origin and Kanamycin
499 resistance with Gibson cloning (Gibson et al., 2009), primers in Supplementary Table 4. The
500 plasmids were transformed into *P. putida* KT2440 or *P. putida* KT2440 mutant strains by
501 electroporation. Three biological replicates of each strain were cultured in LB and stored with 25%
502 (vol/vol) glycerol at -80 ˚C. Complementation plasmids (GFP reporter plasmids with full length RR
503 driven by pBAD promoter and constitutively expressed AraC) were combinatorially built
504 leveraging Golden Gate cloning (Engler et al., 2008) and j5 DNA assembly design (Hillson et al.,
505 2012) (diva.jbei.org), primers in Supplementary Table 4. The plasmids were transformed into
506 knockout strains of *P. putida* KT2440 by electroporation. 3-6 biological replicates of each strain
507 were cultured in LB and stored with 25 % (v/v) glycerol at -80 ˚C.

508 *RRs with switched specificity*

509 Gene blocks (TWIST Biosciences, San Francisco, CA) of REC domains (Supplementary Table
510 3) with co-varying mutations (co-variation score > 1.1) were cloned into the complementation
511 plasmids with Gibson assembly (Gibson et al., 2009). The plasmids were transformed into *P.*
512 *putida* KT2440 or knockout strains of *P. putida* KT2440 by electroporation. Six biological
513 replicates of each strain were cultured in LB and stored with 25 % (vol/vol) glycerol at -80 ˚C.

514 *GFP reporter Assays*

515 Reporter strains were adapted to M9 minimal media (MM) (see supplemental methods for strain
516 specific minimal media recipes) supplemented with 0.5 % (wt/vol) glucose as the sole carbon
517 source in 3 overnight passages, and were stored in MM at -80˚C in 25 % (vol/vol) glycerol.
518 Adapted strains were cultured in MM + 0.5 % (wt/vol) glucose and passaged to MM + 0.5 %
519 (wt/vol) glucose with or without a second carbon-source (40 mM glutamic Acid, 40 mM a-
520 ketoglutaric acid, or 20 mM butyric acid, unless otherwise specified). After 24-hours of growth,
521 fluorescence was measured by flow cytometry on the BD Accuri C6 (BD Biosciences, San Jose,
522 CA). Autofluorescence was gated out with FlowJo (BD Biosciences, San Jose, CA), using a non-
523 fluorescent strain of P. putida KT2440 carrying an empty vector plasmid for reference. To remove
524 noise, the GFP mean for samples with less than 150 events after gating was set to 0. Otherwise,
525 the GFP mean of the remaining events after gating was reported. Statistical significance was
526 determined by T-test between replicates.

527 **Resource Availability**

528 *Data and Code Availability*

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

531 DAP-seq data is have been deposited in NCBI's Gene Expression Omnibus (Edgar et al., 2002)
532 and are accessible through GEO Series accession number GSE157075
533 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157075).

534 ## References

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

537 Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and
538 Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res.
539 *37*, W202-8.

540 Bakan, A., Meireles, L.M., and Bahar, I. (2011). ProDy: protein dynamics inferred from theory
541 and experiments. Bioinformatics *27*, 1575–1577.

542 Bakan, A., Dutta, A., Mao, W., Liu, Y., Chennubhotla, C., Lezon, T.R., and Bahar, I. (2014). Evol
543 and ProDy for bridging protein sequence evolution and structural dynamics. Bioinformatics *30*,
544 2681–2683.

545 Barajas, J.F., Blake-Hedges, J.M., Bailey, C.B., Curran, S., and Keasling, J.D. (2017).
546 Engineered polyketides: Synergy between protein and host level engineering. Synthetic and
547 Systems Biotechnology *2*, 147–166.

548 Bellieny-Rabelo, D., Nkomo, N.P., Shyntum, D.Y., and Moleleki, L.N. (2020). Horizontally
549 Acquired Quorum-Sensing Regulators Recruited by the PhoP Regulatory Network Expand the
550 Host Adaptation Repertoire in the Phytopathogen Pectobacterium brasiliense. MSystems *5*.

551 Bernardo, L.M.D., Johansson, L.U.M., Solera, D., Skärfstad, E., and Shingler, V. (2006). The
552 guanosine tetraphosphate (ppGpp) alarmone, DksA and promoter affinity for RNA polymerase
553 in regulation of sigma-dependent transcription. Mol. Microbiol. *60*, 749–764.

554 Bernardo, L.M.D., Johansson, L.U.M., Skärfstad, E., and Shingler, V. (2009). sigma54-promoter
555 discrimination and regulation by ppGpp and DksA. J. Biol. Chem. *284*, 828–838.

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

558 Brewster, J.L., McKellar, J.L.O., Finn, T.J., Newman, J., Peat, T.S., and Gerth, M.L. (2016).
559 Structural basis for ligand recognition by a Cache chemosensory domain that mediates
560 carboxylate sensing in Pseudomonas syringae. Sci. Rep. *6*, 35198.

561 Briegel, A., Ortega, D.R., Tocheva, E.I., Wuichet, K., Li, Z., Chen, S., Müller, A., Iancu, C.V.,
562 Murphy, G.E., Dobro, M.J., et al. (2009). Universal architecture of bacterial chemoreceptor
563 arrays. Proc Natl Acad Sci USA *106*, 17181–17186.

564 Capra, E.J., and Laub, M.T. (2012). Evolution of two-component signal transduction systems.
565 Annu. Rev. Microbiol. *66*, 325–347.

566 Capra, E.J., Perchuk, B.S., Skerker, J.M., and Laub, M.T. (2012). Adaptive mutations that
567 prevent crosstalk enable the expansion of paralogous signaling protein families. Cell *150*, 222–
568 232.

569 Casas-Pastor, D., Müller, R.R., Becker, A., Buttner, M., Gross, C., Mascher, T., Goesmann, A.,
570 and Fritz, G. (2019). Expansion and re-classification of the extracytoplasmic function (ECF) σ
571 factor family. BioRxiv.

572 Cases, I., Ussery, D.W., and De Lorenzo, V. (2003). The σ^{54} regulon (sigmulon)
573 of*Pseudomonas putida*. Environ. Microbiol. *5*, 1281–1293.

574 Chen, Y.-T., Chang, H.Y., Lu, C.L., and Peng, H.-L. (2004). Evolutionary analysis of the two-
575 component systems in Pseudomonas aeruginosa PAO1. J. Mol. Evol. *59*, 725–737.

576 Chiniquy, J., Garber, M.E., Mukhopadhyay, A., and Hillson, N.J. (2020). Fluorescent
577 amplification for next generation sequencing (FA-NGS) library preparation. BMC Genomics *21*,
578 85.

579 Choi, K., and Kim, S. (2011). Building interacting partner predictors using co-varying residue
580 pairs between histidine kinase and response regulator pairs of 48 bacterial two-component
581 systems. Proteins *79*, 1118–1131.

582 Dalebroux, Z.D., and Swanson, M.S. (2012). ppGpp: magic beyond RNA polymerase. Nat. Rev.
583 Microbiol. *10*, 203–212.

584 Dutta, R., Qin, L., and Inouye, M. (1999). Histidine kinases: diversity of domain organization.
585 Mol. Microbiol. *34*, 633–640.

586 Edgar, R., Domrachev, M., and Lash, A.E. (2002). Gene Expression Omnibus: NCBI gene
587 expression and hybridization array data repository. Nucleic Acids Res. *30*, 207–210.

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

590 Engler, C., Kandzia, R., and Marillonnet, S. (2008). A one pot, one step, precision cloning
591 method with high throughput capability. PLoS ONE *3*, e3647.

592 Forslund, S.K., Kaduk, M., and Sonnhammer, E.L.L. (2019). Evolution of protein domain
593 architectures. Methods Mol. Biol. *1910*, 469–504.

594 Galperin, M.Y. (2005). A census of membrane-bound and intracellular signal transduction
595 proteins in bacteria: bacterial IQ, extroverts and introverts. BMC Microbiol. *5*, 35.

596 Galperin, M.Y. (2006). Structural classification of bacterial response regulators: diversity of
597 output domains and domain combinations. J. Bacteriol. *188*, 4169–4182.

598 Galperin, M.Y. (2010). Diversity of structure and function of response regulator output domains.
599 Curr. Opin. Microbiol. *13*, 150–159.

600 Galperin, M.Y., Higdon, R., and Kolker, E. (2010). Interplay of heritage and habitat in the
601 distribution of bacterial signal transduction systems. Mol. Biosyst. *6*, 721–728.

602 Garber, M.E., Rajeev, L., Kazakov, A.E., Trinh, J., Masuno, D., Thompson, M.G., Kaplan, N.,
603 Luk, J., Novichkov, P.S., and Mukhopadhyay, A. (2018). Multiple signaling systems target a
604 core set of transition metal homeostasis genes using similar binding motifs. Mol. Microbiol. *107*,
605 704–717.

606 Gavira, J.A., Gumerov, V.M., Rico-Jiménez, M., Petukh, M., Upadhyay, A.A., Ortega, A., Matilla,
607 M.A., Zhulin, I.B., and Krell, T. (2020). How bacterial chemoreceptors evolve novel ligand
608 specificities. MBio *11*.

609 Gibson, D.G., Young, L., Chuang, R.-Y., Venter, J.C., Hutchison, C.A., and Smith, H.O. (2009).
610 Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat. Methods *6*, 343–
611 345.

612 Grebe, T.W., and Stock, J.B. (1999). The histidine protein kinase superfamily. Adv. Microb.
613 Physiol. *41*, 139–227.

614 Gumerov, V.M., Ortega, D.R., Adebali, O., Ulrich, L.E., and Zhulin, I.B. (2020). MiST 3.0: an
615 updated microbial signal transduction database with an emphasis on chemosensory systems.
616 Nucleic Acids Res. *48*, D459–D464.

617 Henry, J.T., and Crosson, S. (2011). Ligand-binding PAS domains in a genomic, cellular, and
618 structural context. Annu. Rev. Microbiol. *65*, 261–286.

619 Herrou, J., Crosson, S., and Fiebig, A. (2017). Structure and function of HWE/HisKA2-family
620 sensor histidine kinases. Curr. Opin. Microbiol. *36*, 47–54.

621 Hervás, A.B., Canosa, I., Little, R., Dixon, R., and Santero, E. (2009). NtrC-dependent
622 regulatory network for nitrogen assimilation in *Pseudomonas putida*. J. Bacteriol. *191*, 6123–
623 6135.

624 Hillson, N.J., Rosengarten, R.D., and Keasling, J.D. (2012). j5 DNA assembly design
625 automation software. ACS Synth. Biol. *1*, 14–21.

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

628 Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., Butterfield,
629 C.N., Hernsdorf, A.W., Amano, Y., Ise, K., et al. (2016). A new view of the tree of life. Nat.
630 Microbiol. *1*, 16048.

631 Jacob-Dubuisson, F., Mechaly, A., Betton, J.-M., and Antoine, R. (2018). Structural insights into
632 the signalling mechanisms of two-component systems. Nat. Rev. Microbiol. *16*, 585–593.

633 Jung, K., Fried, L., Behr, S., and Heermann, R. (2012). Histidine kinases and response
634 regulators in networks. Curr. Opin. Microbiol. *15*, 118–124.

635 Jurado, P., Fernández, L.A., and de Lorenzo, V. (2003). Sigma 54 levels and physiological
636 control of the Pseudomonas putida Pu promoter. J. Bacteriol. *185*, 3379–3383.

637 Katoh, K., and Standley, D.M. (2013). MAFFT multiple sequence alignment software version 7:
638 improvements in performance and usability. Mol. Biol. Evol. *30*, 772–780.

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

641 Kim, S., Hirakawa, H., Muta, S., and Kuhara, S. (2010). Identification and classification of a two-
642 component system based on domain structures in bacteria and differences in domain structure
643 between Gram-positive and Gram-negative bacteria. Biosci. Biotechnol. Biochem. *74*, 716–720.

644 Lai, R.-Z., and Parkinson, J.S. (2018). Monitoring Two-Component Sensor Kinases with a
645 Chemotaxis Signal Readout. Methods Mol. Biol. *1729*, 127–135.

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

648 Laub, M.T., and Goulian, M. (2007). Specificity in two-component signal transduction pathways.
649 Annu. Rev. Genet. *41*, 121–145.

650 Laub, M.T., Biondi, E.G., and Skerker, J.M. (2007). Phosphotransfer profiling: systematic
651 mapping of two-component signal transduction pathways and phosphorelays. Meth. Enzymol.
652 *423*, 531–548.

653 Leech, A.J., Sprinkle, A., Wood, L., Wozniak, D.J., and Ohman, D.E. (2008). The NtrC family
654 regulator AlgB, which controls alginate biosynthesis in mucoid Pseudomonas aeruginosa, binds

655 directly to the algD promoter. J. Bacteriol. *190*, 581–589.

656 Lee, D.J., Minchin, S.D., and Busby, S.J.W. (2012). Activating transcription in bacteria. Annu.
657 Rev. Microbiol. *66*, 125–152.

658 Linsky, M., Vitkin, Y., and Segal, G. (2020). A Novel Legionella Genomic Island Encodes a
659 Copper-Responsive Regulatory System and a Single Icm/Dot Effector Protein Transcriptionally
660 Activated by Copper. MBio *11*.

661 Liu, Y., Zhao, L., Yang, M., Yin, K., Zhou, X., Leung, K.Y., Liu, Q., Zhang, Y., and Wang, Q.
662 (2017). Transcriptomic dissection of the horizontally acquired response regulator EsrB reveals
663 its global regulatory roles in the physiological adaptation and activation of T3SS and the
664 cognate effector repertoire in Edwardsiella piscicida during infection toward turbot. Virulence *8*,
665 1355–1377.

666 Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G.,
667 Durbin, R., and 1000 Genome Project Data Processing Subgroup (2009). The Sequence
668 Alignment/Map format and SAMtools. Bioinformatics *25*, 2078–2079.

669 Lonetto, M.A., Rhodius, V., Lamberg, K., Kiley, P., Busby, S., and Gross, C. (1998).
670 Identification of a contact site for different transcription activators in region 4 of the Escherichia
671 coli RNA polymerase sigma70 subunit. J. Mol. Biol. *284*, 1353–1365.

672 Lori, C., Kaczmarczyk, A., de Jong, I., and Jenal, U. (2018). A Single-Domain Response
673 Regulator Functions as an Integrating Hub To Coordinate General Stress Response and
674 Development in Alphaproteobacteria. MBio *9*.

675 Lundgren, B.R., Villegas-Peñaranda, L.R., Harris, J.R., Mottern, A.M., Dunn, D.M., Boddy, C.N.,
676 and Nomura, C.T. (2014). Genetic analysis of the assimilation of C5-dicarboxylic acids in
677 Pseudomonas aeruginosa PAO1. J. Bacteriol. *196*, 2543–2551.

678 Maervoet, V.E.T., and Briers, Y. (2017). Synthetic biology of modular proteins. Bioengineered *8*,
679 196–202.

680 McClune, C.J., and Laub, M.T. (2020). Constraints on the expansion of paralogous protein
681 families. Curr. Biol. *30*, R460–R464.

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

684 Mistry, J., Finn, R.D., Eddy, S.R., Bateman, A., and Punta, M. (2013). Challenges in homology
685 search: HMMER3 and convergent evolution of coiled-coil regions. Nucleic Acids Res. *41*, e121.

686 Nishijyo, T., Haas, D., and Itoh, Y. (2001). The CbrA-CbrB two-component regulatory system
687 controls the utilization of multiple carbon and nitrogen sources in Pseudomonas aeruginosa.
688 Mol. Microbiol. *40*, 917–931.

689 Ortega, Á., Zhulin, I.B., and Krell, T. (2017). Sensory repertoire of bacterial chemoreceptors.
690 Microbiol. Mol. Biol. Rev. *81*.

691 Padilla-Vaca, F., Mondragón-Jaimes, V., and Franco, B. (2017). General Aspects of Two-
692 Component Regulatory Circuits in Bacteria: Domains, Signals and Roles. Curr. Protein Pept.
693 Sci. *18*, 990–1004.

694 Podgornaia, A.I., and Laub, M.T. (2013). Determinants of specificity in two-component signal
695 transduction. Curr. Opin. Microbiol. *16*, 156–162.

696 Potvin, E., Sanschagrin, F., and Levesque, R.C. (2008). Sigma factors in Pseudomonas
697 aeruginosa. FEMS Microbiol. Rev. *32*, 38–55.

698 Pougach, K., Voet, A., Kondrashov, F.A., Voordeckers, K., Christiaens, J.F., Baying, B., Benes,
699 V., Sakai, R., Aerts, J., Zhu, B., et al. (2014). Duplication of a promiscuous transcription factor
700 drives the emergence of a new regulatory network. Nat. Commun. *5*, 4868.

701 Price, M.N., Dehal, P.S., and Arkin, A.P. (2008). Horizontal gene transfer and the evolution of
702 transcriptional regulation in Escherichia coli. Genome Biol. *9*, R4.

703 Price, M.N., Dehal, P.S., and Arkin, A.P. (2010). FastTree 2 — approximately maximum-
704 likelihood trees for large alignments. PLoS ONE *5*, e9490.

705 Price, M.N., Wetmore, K.M., Waters, R.J., Callaghan, M., Ray, J., Liu, H., Kuehl, J.V., Melnyk,
706 R.A., Lamson, J.S., Suh, Y., et al. (2018). Mutant phenotypes for thousands of bacterial genes
707 of unknown function. Nature *557*, 503–509.

708 Rajeev, L., Garber, M.E., and Mukhopadhyay, A. (2020). Tools to map target genes of bacterial
709 two-component system response regulators. Environ. Microbiol. Rep. *12*, 267–276.

710 Rhodius, V.A., Segall-Shapiro, T.H., Sharon, B.D., Ghodasara, A., Orlova, E., Tabakh, H.,
711 Burkhardt, D.H., Clancy, K., Peterson, T.C., Gross, C.A., et al. (2013). Design of orthogonal
712 genetic switches based on a crosstalk map of σs, anti-σs, and promoters. Mol. Syst. Biol. *9*,
713 702.

714 Ronneau, S., and Hallez, R. (2019). Make and break the alarmone: regulation of (p)ppGpp
715 synthetase/hydrolase enzymes in bacteria. FEMS Microbiol. Rev. *43*, 389–400.

716 Ryjenkov, D.A., Tarutina, M., Moskvin, O.V., and Gomelsky, M. (2005). Cyclic diguanylate is a
717 ubiquitous signaling molecule in bacteria: insights into biochemistry of the GGDEF protein
718 domain. J. Bacteriol. *187*, 1792–1798.

719 Salazar, M.E., and Laub, M.T. (2015). Temporal and evolutionary dynamics of two-component
720 signaling pathways. Curr. Opin. Microbiol. *24*, 7–14.

721 Sankhe, G.D., Dixit, N.M., and Saini, D.K. (2018). Activation of Bacterial Histidine Kinases:
722 Insights into the Kinetics of the cis Autophosphorylation Mechanism. MSphere *3*.

723 Schmidl, S.R., Ekness, F., Sofjan, K., Daeffler, K.N.-M., Brink, K.R., Landry, B.P., Gerhardt,
724 K.P., Dyulgyarov, N., Sheth, R.U., and Tabor, J.J. (2019). Rewiring bacterial two-component
725 systems by modular DNA-binding domain swapping. Nat. Chem. Biol. *15*, 690–698.

726 Shingler, V. (2011). Signal sensory systems that impact σ^{54} -dependent transcription. FEMS
727 Microbiol. Rev. *35*, 425–440.

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

730 Skerker, J.M., Perchuk, B.S., Siryaporn, A., Lubin, E.A., Ashenberg, O., Goulian, M., and Laub,
731 M.T. (2008). Rewiring the specificity of two-component signal transduction systems. Cell *133*,
732 1043–1054.

733 Smanski, M.J., Zhou, H., Claesen, J., Shen, B., Fischbach, M.A., and Voigt, C.A. (2016).
734 Synthetic biology to access and expand nature's chemical diversity. Nat. Rev. Microbiol. *14*,
735 135–149.

736 Sonawane, A.M., Singh, B., and Röhm, K.-H. (2006). The AauR-AauS two-component system
737 regulates uptake and metabolism of acidic amino acids in Pseudomonas putida. Appl. Environ.
738 Microbiol. *72*, 6569–6577.

739 Studier, F.W. (2005). Protein production by auto-induction in high density shaking cultures.
740 Protein Expr. Purif. *41*, 207–234.

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

744 Thompson, M.G., Costello, Z., Hummel, N., Cruz-Morales, P., Blake-Hedges, J.M., Krishna, R.,
745 Skyrud, W., Pearson, A., Incha, M., Shih, P., et al. (2019). Robust characterization of two
746 distinct glutarate sensing transcription factors of *Pseudomonas putida* L-lysine metabolism. ACS
747 Synth. Biol.

748 Thompson, M.G., Incha, M.R., Pearson, A.N., Schmidt, M., Sharpless, W.A., Eiben, C.B., Cruz-
749 Morales, P., Blake-Hedges, J.M., Liu, Y., Adams, C.A., et al. (2020a). Functional analysis of the
750 fatty acid and alcohol metabolism of Pseudomonas putida using RB-TnSeq. Appl. Environ.
751 Microbiol.

752 Thompson, M.G., Pearson, A.N., Barajas, J.F., Cruz-Morales, P., Sedaghatian, N., Costello, Z.,
753 Garber, M.E., Incha, M.R., Valencia, L.E., Baidoo, E.E.K., et al. (2020b). Identification,
754 Characterization, and Application of a Highly Sensitive Lactam Biosensor from Pseudomonas
755 putida. ACS Synth. Biol. *9*, 53–62.

756 Treangen, T.J., and Rocha, E.P.C. (2011). Horizontal transfer, not duplication, drives the
757 expansion of protein families in prokaryotes. PLoS Genet. *7*, e1001284.

758 Voordeckers, K., Pougach, K., and Verstrepen, K.J. (2015). How do regulatory networks evolve
759 and expand throughout evolution? Curr. Opin. Biotechnol. *34*, 180–188.

760 Way, J.C., Collins, J.J., Keasling, J.D., and Silver, P.A. (2014). Integrating biological redesign:
761 where synthetic biology came from and where it needs to go. Cell *157*, 151–161.

762 Wetmore, K.M., Price, M.N., Waters, R.J., Lamson, J.S., He, J., Hoover, C.A., Blow, M.J.,
763 Bristow, J., Butland, G., Arkin, A.P., et al. (2015). Rapid quantification of mutant fitness in
764 diverse bacteria by sequencing randomly bar-coded transposons. MBio *6*, e00306-15.

765 Wigneshweraraj, S., Bose, D., Burrows, P.C., Joly, N., Schumacher, J., Rappas, M., Pape, T.,
766 Zhang, X., Stockley, P., Severinov, K., et al. (2008). Modus operandi of the bacterial RNA
767 polymerase containing the sigma54 promoter-specificity factor. Mol. Microbiol. *68*, 538–546.

768 Wuichet, K., Cantwell, B.J., and Zhulin, I.B. (2010). Evolution and phyletic distribution of two-
769 component signal transduction systems. Curr. Opin. Microbiol. *13*, 219–225.

770 Wu, X., Monchy, S., Taghavi, S., Zhu, W., Ramos, J., and van der Lelie, D. (2011). Comparative
771 genomics and functional analysis of niche-specific adaptation in Pseudomonas putida. FEMS

772 Microbiol. Rev. *35*, 299–323.

773 Zschiedrich, C.P., Keidel, V., and Szurmant, H. (2016). Molecular Mechanisms of Two-
774 Component Signal Transduction. J. Mol. Biol. *428*, 3752–3775.