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], Matthew R. Incha[2,5], Jay D. Keasling[2,4,6], 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
5. University of California, Berkeley Department of Plant and Microbial Biology
6. University of California, Berkeley Department of Chemical and Biomolecular Engineering

* Correspondence: amukhopadhyay@lbl.gov

Abstract

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

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

Keywords

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

Introduction

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

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

phosphotransfer to its cognate response regulator (RR). Although TCSs can be highly redundant in a single bacterial genome[1,2], biochemical and genetic evidence suggests that interactions between cognate HK and RR pairs are conserved. Coevolutionary events of duplication and divergence led to orthogonality of cognate pairs[3–5]. TCS are modular, input domains in HKs typically encode for sensing functionality, whereas effector domains in RRs confer output function (Figure 1a). The domains of HK-RR pair are typically carried through to descendants during VGT.

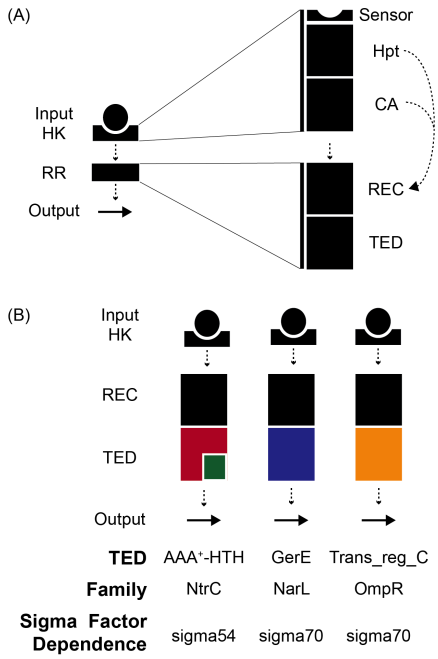

Figure 1: ***Domain architecture of TCS with transcriptional effector domains (TEDs):*** (A) Canonical TCS have input signals that activate histidine kinases (HK) and response regulators (RR) to coordinate some output response. An expanded view of HKs and RRs reveals their modularity. (B) Relevant RR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

Hypotheses of domain swapping have emerged as a potential mechanism for the diversification of TCSs. Domain swaps occur via HGT and homologous recombination[6]. Core parts of TCSs (HPt, CA, REC domains) remain intact, while swaps of the input or sensing parts diversify the signals[7] or swaps of output or effector domains modify the function of the entire cascade[1].

61 RR effector domains dictate the final effect of a signaling cascade. RRs can harbor effector
62 domains that lead to chemotactic responses[8], regulation of small molecules[9], or transcriptional
63 regulation[1,2]. Transcriptional effector domains (TEDs) interact primarily with DNA in a sequence-
64 specific manner to regulate transcription. TEDs are diverse in structure, protein-DNA interaction,
65 protein-protein interaction, and function, and are binned into subclasses or families[1]. In this study
66 we predominantly observe TEDs associated with RRs that confer activation of genes with sigma54
67 factors (AAA^+ tethered to HTH_8, NtrC family), and regulation of genes with sigma70 factors
68 (GerE, NarL family or Trans_reg_C, OmpR family) (Figure 1b).

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

74 Results

75 We identified all RRs in five representative pseudomonads, subdividing them by their
76 output domains, considering only RRs with TEDs (Supplementary Table 1). The phylogenetic tree
77 of the REC domains agrees with a previously published tree of *P. aeruginosa* PAO1 REC
78 domains[10]. Consistent with the current understanding how highly redundant proteins evolve in
79 bacteria[4,11], our results demonstrate that the REC domains of RRs do not clade by species, but
80 instead clade by TEDs (Figure 2). A domain swap, hypothesized when RRs of one family cluster
81 with RRs of another, is observed within the AAA^+-HTH_8/NtrC-like cluster (red and green) of the
82 REC phylogenetic tree, where REC domains with an alternative TED, GerE/NarL-like (blue) are
83 present (Figure 2).

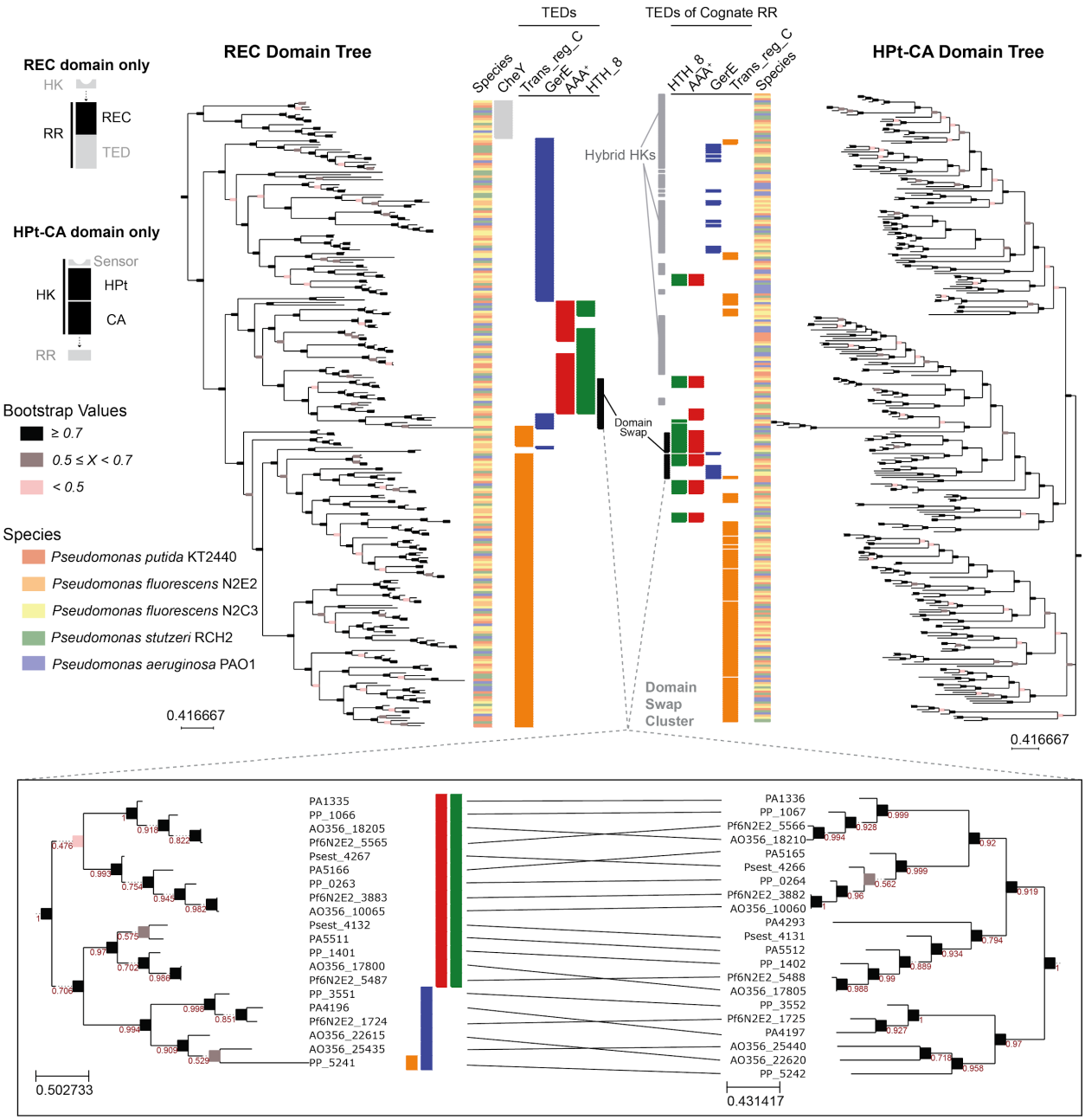

Figure 2: **Phylogenetic trees of REC (left) and HPt-CA (right) domains of RRs and HKs from select pseudomonads**. Trees are annotated with species and transcriptional effector domain (TED) at the leaves. Hybrid HKs on the HPt-CA tree are highlighted in grey, and are notably associated with TCS that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

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

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

99 We next determined the evolutionary distance of the domain swap, asking whether it was
100 present in pseudomonads only, or all proteobacteria. Using representatives from α-, β-, γ-, and δ-
101 proteobacteria, selected for their high proportions of NtrC-like and NarL-like RRs, we
102 recapitulated the results of the *Pseudomonas* REC domain-based phylogenetic tree. The
103 proteobacteria REC phylogenetic tree reveals that the swap is found in the select strains of α-, β-,
104 and γ- proteobacteria, but not observed in the selected δ-proteobacteria (Supplementary Figure 1).
105 This discrepancy could be the result of gene loss in the selected representatives, or could indicate
106 that the swapped clade never existed in δ-proteobacteria. Notably, in the new view of the tree of
107 life[12], δ-proteobacteria are more distantly related to the rest of the proteobacterial clade than
108 previously thought. Together these results imply that a domain swap likely occurred at the cuspis
109 of the proteobacteria (α-, β-, γ-) clade.

110 Functional similarity between the extra-familial representatives in the swapped cluster
111 would indicate a genuine relationship. Closer observation of the swapped cluster reveals that there
112 are four subclusters, three of which harbor AAA^+ domains (NtrC-like/sigma54-dependent) and
113 one which harbors a GerE domain (NarL-like/sigma70-dependent) (Figure 2). We focused our
114 efforts on the functions of the TCSs in *P. putida* KT2440, because it is a well-studied model strain
115 with a readily available genetic toolkit to enable validation studies. Because the domain swap
116 occurs within the NtrC-like cluster, we hypothesized that the signaling cascades were involved in
117 nutrient uptake of carbon or nitrogen sources[13]. Additionally, NtrC representatives within the

118 swapped cluster were previously characterized as amino acid or di-carboxylic acid sensors[13–16].
119 The functions of the NarL-like representatives were unknown.

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

130 To determine the signals for the cluster, we utilized randomly barcoded transposon libraries
131 for next generation sequencing of *P. putida* KT2440 grown with single carbon sources. The genes
132 with low fitness scores (fitness < -2) under a given growth condition have condition-specific
133 essentiality. We observed that the gene clusters harboring the RRs of interest each had low fitness
134 scores for growth with either glutamic, succinic, a-ketoglutaric, or butyric acids (Figure 3a). We
135 speculated that genes and their corresponding TCSs with synonymous fitness values represented
136 a signaling cascade where the signal was the carbon source. We proposed that the NtrC-like RRs,
137 PP_1066, PP_0263, and PP_1401 are regulated by glutamic[16], succinic, and a-ketoglutaric
138 acids[14,15] respectively. We also hypothesized that the NarL-like RR, PP_3551, is regulated by
139 butyric acid. We were not able to identify binding sites or a DNA binding motif for PP_3551 and
140 relied on the gene cluster and fitness data to hypothesize an output for the butyric acid responsive

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

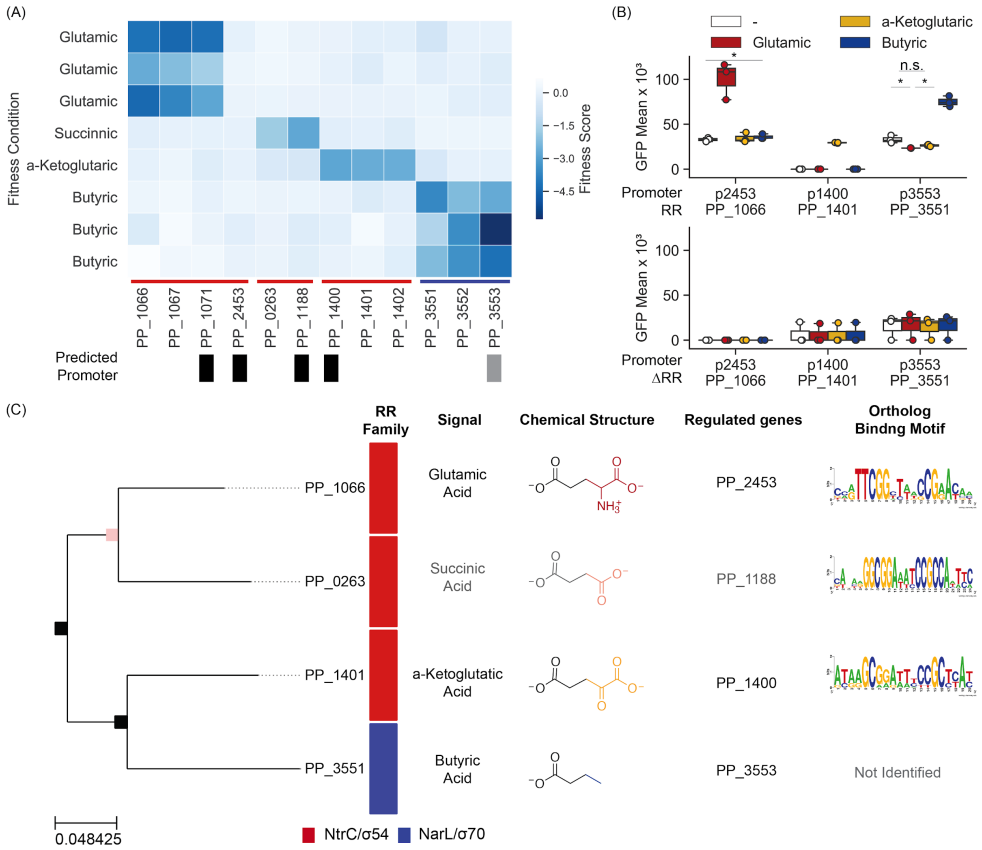

Figure 3: ***Functional genomics for determining input signal and output genes*:** (A) Heatmap of fitness data for TCS gene clusters within the domain swapped cluster. A low fitness score (fitness< - 2) indicates a gene has condition-specific essentiality. Predicted promoters are represented by a rectangle below downstream genes, black fill indicates promoters with DAP-seq identified motifs. (B) GFP reporter plasmids in WT and 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 10^3after gating. Statistical significance was determined between conditions by t-test (* = p-value < 0.05, n.s. = not significant). Center line, median; box limits, upper and lower quartiles; whiskers, 1.5x interquartile range; points with black diamonds, outliers; n = 3. (C) Summary of findings from functional genomics studies. The relationship of the *P. putida* TCSs is shown by a pruned phylogenetic tree to the left, and results, RR family, validated signal, signal chemical structure, validated regulated genes, orthogonal binding motif identified by DAP-seq are drawn as a table to the right. If the findings were not validated by reporter assay, the results are displayed in grey.

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

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

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

169 We proposed an ancestral model in which common ancestors of CarSR were generalists,
170 able to sense a broad range of carboxylic acids and interact promiscuously with cognate pairs
171 (Figure 4a). To demonstrate the model we coerced the TCSs into a pseudo-ancestral state (Figure
172 4b,c). We applied co-variance analysis to HPt-CA-REC domains of cognate HK-RR pairs to
173 identify co-varying residues (Supplementary Figure 3) between HPt-CA and REC domains.
174 Consistent with previous studies[19], the first alpha helix of the REC domain, previously shown to
175 share points of contact with HPt domain, contained the highest scoring residues. We applied site-
176 directed mutagenesis switching the highest scoring residues of the REC domain of PP_1066
177 (hereafter RR_g) and PP_3551 (hereafter RR_b) (Figure 4b).

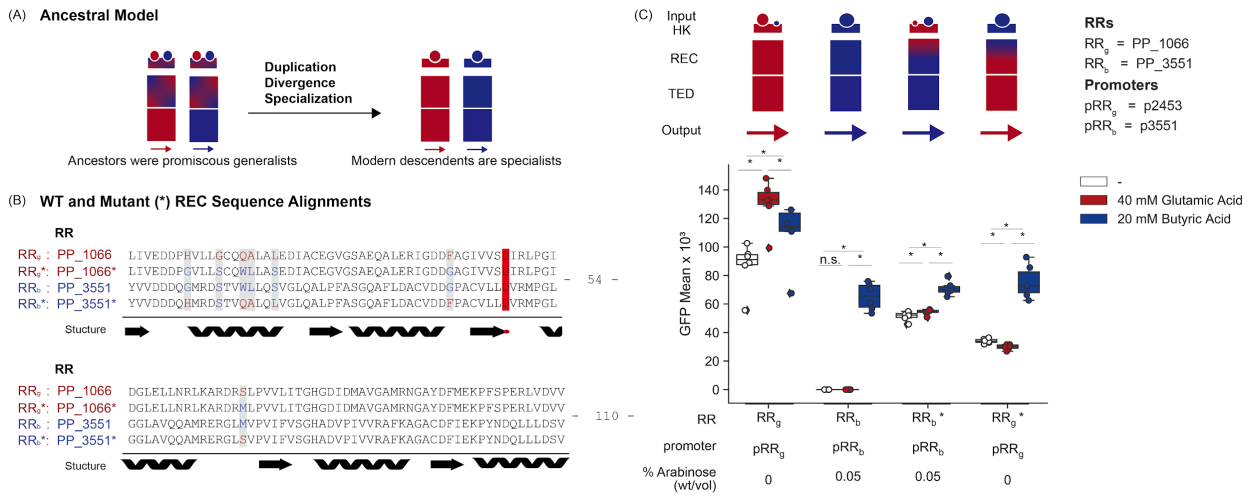

Figure 4: ***Covarying residue swaps alter REC specificity of butyric acid responsive TCS:*** (A) Model where hypothetical generalist ancestors of domain swapped cluster evolved to become modern day specialists. (B) Sequence alignments of REC domains of native RRs PP_1066 (RR_g), PP_3551 (RR_b), and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

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

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

204 In an alternative hypothesis, the CarR subcluster could be derived from HGT alone, not
205 domain swapping. We queried the configuration of RR families in bacterial phyla with high counts
206 of NarL-like and NtrC-like RRs, and found representative strains from Firmicutes, Bacteroides,
207 Acidobacteria, δ-Proteobacteria, and γ-Proteobacteria (*P. Putida* KT2440) (Supplementary Figure
208 5a). We identified all RRs and their corresponding REC domains from the representative species,
209 and repeated phylogenetic analysis and domain mapping (Supplementary Figure 5b). We observed
210 a similar configuration to pseudomonads and proteobacteria trees (Figure 2b, Supplementary
211 Figure 1), indicating that the organization of TCS families observed in present day bacteria is as
212 ancient as the bacterial domain of life, and the alternative hypothesis of HGT fails.

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

221 Discussion

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

225 domains, and that when the domain swap occured, the output gene was promiscuous within
226 carboxylic acid metabolism and could moonlight a beneficial function[20]. The *in vivo* mutational
227 screen provides evidence that insulated, extra-familial RRs within the same species share sequence
228 space and are closely related. In a previous study, an extra-familial specificity swap of an Omp
229 family HK enabled interaction with an Ntr family RR in *C. crescentus*, leading to the conclusion
230 that a duplicated Ntr family TCS overlapped in sequence space with an Omp family TCS[21]. Their
231 results demonstrated how VGT of a duplicated TCS facilitated insulation from unwanted cross-
232 talk, while our work highlights how domain swapping can alter the output TED of an already well-
233 insulated TCS.

234 NtrC-like RRs require sigma54 transcription factors to activate genes in their regulons,
235 whereas NarL-like RRs interact with sigma70. It is well documented that cellular levels of small
236 nucleotide, (p)ppGpp, in combination with the protein DksA regulate RNA polymerase's (RNAP)
237 preference for sigma70 (extracytoplasmic function family)[22] or sigma54[23,24]. While the
238 environmental conditions needed for pseudomonads to reach the appropriate levels of (p)ppGpp
239 and DksA for sigma54 occupation of RNAP are unknown[25], it can be reasoned that the switch is
240 lifestyle dependent. TCSs implicated in this study were found to be metabolite responsive systems,
241 and could play a major role in survival and fitness. We speculate that the capability to consume
242 carboxylic acids without regulatory constraints dictated by sigma factor dependence, may have
243 given a fitness advantage to the ancestral strain.

244 Domain swapping in general is an important biological phenomenon that can impact all
245 domains of life[6]. HGT is known to mobilize genes or entire gene clusters that can be subsequently
246 integrated into foreign genomes[26]. Including domain swapping in HGT models implies that
247 bacteria have another mechanism to develop new proteins. While we focused on a single domain

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

252 Native enzymes and proteins provide attractive parts for practical applications[27,28] and
253 potential answers to plug-and-play biology via synthetic domain swapping. One critical
254 component of synthetic domain swapping is identifying domain boundaries for a seamless swap.
255 A previous study identified domain boundaries between REC domains and TEDs for *E. coli* RRs,
256 engineering only inter-family TED swaps[29]. Our study demonstrates that evolution synthesized an
257 extra-familial swap in a functional signaling cascade. Evolutionarily driven domain swaps could
258 be an untapped source for identifying domain boundaries for synthetic biology.

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

264 Methods

265 Phylogenetic analysis of REC and HPt-CA trees

266 RRs and HKs were identified with hmmsearch from HMMER v3.1b2[30] to search for all proteins
267 with Response_reg (PF00072) and HisKA (PF00512) domains. Alternatively, HKs and RRs were
268 identified with the Microbial Signal Transduction database (MISTDB)[31]. Using HMMER again,
269 we queried the domain architecture of the signaling proteins and selected for RRs with HTH_8,

270 Trans_reg_C, AAA^{+}, GerE, HTH_18 (AraC) domains. REC and HPt-CA domain sequences were
271 extracted from whole sequences based on coordinates determined by the hmmsearch. Domains
272 were aligned using the MAFFT-LINSI algorithm from MAFFT v7.310[32]. Phylogenetic trees were
273 constructed using FastTree 2[33]. Trees were visualized and annotated using python based ETE3[34].

274 **Automated DNA affinity purification - seq**

275 *DNA preparation for NGS*

276 Pseudomonas isolates were cultured in either LB or minimal media (see Supplementary Table 3
277 for strain specific minimal media recipes). Genomic DNA was purified with a promega wizard
278 genomic preparation kit (Promega, Madison, WI). DNA was sheared with covaris miniTUBE
279 (Covaris, Woburn, MA) to an average size of 200 bp. The DNA quality was confirmed by
280 Bioanalyzer high sensitivity DNA kit (Agilent, Santa Clara). Sheared DNA was then adapter-
281 ligated (AL) with NEBnext Ultra ii Library Preparation kit (New England Biolabs, Ipswich, MA).
282 AL-DNA quality was again confirmed by Bioanalyzer high sensitivity DNA kit (Agilent, Santa
283 Clara). AL-DNA was stored at -20˚C until required for downstream use.
284

285 *Expression Strain Design*

286 Pet28 expression vectors with N-terminal 6x-His-tagged RRs were cloned by Gibson assembly[35].
287 Plasmid design was facilitated by j5 DNA assembly design[36] (diva.jbei.org), see Supplementary
288 Table 4 for primers.
289

290 *Automated DNA affinity purification*

291 Quadruplicates of expression strains were grown in autoinduction media (Zyp-5052[37]) at 37 ˚C,
292 250 RPM, for 5-6 hours and then transferred to grow at 17˚C, 250 RPM, overnight. Cell pellets
293 were harvested and lysed at 37 ˚C for 1 hour in a lysis buffer (1X TBS, 100 µM PMSF (Millipore
294 Sigma, Burlington MA), 2.5 units/mL Benzonase nuclease (Millipore Sigma, Burlington MA), 1
295 mg/mL Lysozyme (Millipore Sigma, Burlington MA)). Lysed cells were then clarified by
296 centrifugation at 3214 x g and further filtered in 96-well filter plates by centrifugation at 1800 x g.
297 To enable high-throughput processing, protein-DNA purification steps were performed with
298 IMAC resin pipette tips (PhyNexus, San Jose, CA) using a custom automated platform with the
299 Biomek FX liquid handler (Beckman Coulter, Indianapolis, IN). The expressed RRs were
300 individually bound to metal affinity resin embedded within the IMAC resin pipette tips and washed
301 in a wash buffer (1X TBS, 10 mM Imidazole, 0.1% Tween 20). The bead bound RRs were then
302 mixed with 60 µL of DNA binding buffer (1X TBS, 10 mM magnesium chloride, 0.4 ng/µL AL-
303 DNA, with or without 50 mM acetyl phosphate (split into duplicates)). The protein bound to its
304 target DNA was then enriched in an enrichment buffer (1X TBS, 10 mM Imidazole, 0.1% Tween
305 20) and eluted in an elution buffer (1X TBS, 180 mM Imidazole). The elution was stored at -20˚C
306 for a minimum of one day and up to a week before proceeding to the NGS library generation. See
307 supplementary methods for detailed protocol.

308

309 *NGS Library Generation*

310 3.2 µL of the elution from the previous step was added to 3.5 µL SYBR green ssoAdvanced
311 (Biorad, Hercules, CA) and 0.15 µL of each dual indexed NGS primers. NGS libraries were

312 prepared by following the protocols for fluorescent amplification of NGS libraries[38]. Pooled
313 libraries were sequenced by Illumina NovaSeq 6000 SP (100 cycles) (Illumina, San Diego, CA).
314

315 *DAP-seq data analysis*

316 Sequenced reads were processed by a computational DAP-seq analysis pipeline as follows.
317 Adapters and low-quality bases were trimmed and reads shorter than 30 bp were filtered out using
318 Trimmomatic v.0.36[39]. The resulting reads were checked for contamination using FOCUS [40]. Then
319 the reads were aligned to the corresponding *Pseudomonas* spp. genome using Bowtie v1.1.2 [41]
320 with –m 1 parameter (report reads with single alignment only). Resulting SAM files were
321 converted to BAM format and sorted using samtools v 0.1.19[42]. Peak calling was performed using
322 SPP 1.16.0[43] with false discovery rate threshold of 0.01 and MLE enrichment ratio threshold of
323 4.0. Enriched motifs were discovered in genome fragments corresponding to the peaks using
324 MEME[44] with parameters –mod anr –minw 12 –maxw 30 –revcomp –pal –nmotifs 1. Source code
325 of the DAP-seq analysis pipeline is available at https://github.com/novichkov-lab/dap-seq-utils.

326 For conserved RRs with small numbers of high-confidence peaks (1-2 per genome), binding motifs
327 were predicted manually by comparative genomics approach. Orthologous RRs were identified by
328 OrthoFinder2[45]. For each of orthologous RRs, one genome fragment corresponding to the peak
329 with the highest enrichment value was selected for motif search. Conserved motifs were
330 discovered using the SignalX tool from GenomeExplorer package[46] with the "inverted repeat"
331 option.

332

333 **Fitness Experiments**

334 The *P. putida* KT2440 RB-TnSeq library, JBEI-1[47], was thawed, inoculated into 25 mL
335 of LB medium supplemented with kanamycin, and grown to OD_{600} of 0.5. Three 1 mL samples
336 were taken after this step to serve as t_0 records of barcode abundance. The library was then
337 washed via centrifugation and resuspension in an equal volume of MOPS minimal medium. The
338 washed cells were then diluted 1:50 in MOPS minimal medium with 10 mM L-glutamate serving
339 as the sole carbon source. The library was cultured in a 96-well deep well plate sealed with a gas
340 permeable membrane (VWR, USA). The plate was shaken (700 rpm) in an INFORS HT
341 Multitron (Infors USA Inc.) at 30 °C for 24 hrs. Duplicate 600 µL samples were then combined
342 and BarSeq analysis was conducted as described previously[48–50].
343 Single carbon source fitness data is available at http://fit.genomics.lbl.gov[51].

344 **Co-variance analysis**

345 Cognate RRs and HKs from *Pseudomonas* and *E. coli* strains were identified as pairs if they were
346 found neighboring each other in their respective genomes. HPt (HisKA), CA (HATPase_C), and
347 REC (Response_reg) domain boundaries were determined with hmmsearch from HMMER
348 v3.1b2[30]. Fasta files of concatenated HPt-CA-REC domains from cognate and randomized HK-
349 RR pairs were aligned with the MAFFT-LINSI algorithm from MAFFT v7.310[32]. Alignment files
350 were then queried for coevolution with the ProDy *Evol* suite[52,53] in python and were plotted in a
351 heatmap. The highest scoring residues > 1.1 were used to inform hypotheses for specificity switch
352 strains.

353 **GFP reporter strain generation and assays**

354 *Knockout strain generation*

355 1000bp homology fragments upstream and downstream of the target gene were cloned into
356 plasmid pKS18. Plasmids were then transformed into E. coli S17 and then mated into P. putida via
357 conjugation. Transconjugants were selected for on LB agar plates supplemented with 30 mg/ml
358 kanamycin and 30 mg/ml chloramphenicol. Transconjugants were then grown overnight on LB
359 media and were then plated on LB agar with no NaCl that was supplemented with 10% (wt/vol)
360 sucrose. Putative deletions were screened on LB agar with no NaCl supplemented with 10%
361 (wt/vol) sucrose and LB agar plate with kanamycin. Colonies that grew in the presence of sucrose
362 but had no resistance to kanamycin were further tested via PCR with primers flanking the target
363 gene to confirm gene deletion.

364 *GFP reporter strains*

365 Promoter boundaries for p2453, p1400, p3553 were identified as the region just upstream of the
366 gene's start codon up until the start or stop codon of the next nearest gene. The promoters were
367 cloned upstream of GFP on a broad host range plasmid with BBR1 origin and Kanamycin
368 resistance with Gibson cloning[35], primers in Supplementary Table 4. The plasmids were
369 transformed into *P. putida* KT2440 or *P. putida* KT2440 mutant strains by electroporation. Three
370 biological replicates of each strain were cultured in LB and stored with 25% (vol/vol) glycerol at
371 -80 ˚C. Complementation plasmids (GFP reporter plasmids with full length RR driven by pBAD
372 promoter and constitutively expressed AraC) were combinatorially built leveraging Golden Gate
373 cloning[54] and j5 DNA assembly design[36] (diva.jbei.org), primers in Supplementary Table 4. The
374 plasmids were transformed into knockout strains of *P. putida* KT2440 by electroporation. 3-6

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

377 *RRs with switched specificity*

378 Gene blocks (TWIST Biosciences, San Francisco, CA) of REC domains (Supplementary Table 3)
379 with co-varying mutations (co-variation score > 1.1) were cloned into the complementation
380 plasmids with Gibson assembly[35]. The plasmids were transformed into *P. putida* KT2440 or
381 knockout strains of *P. putida* KT2440 by electroporation. Six biological replicates of each strain
382 were cultured in LB and stored with 25 % (vol/vol) glycerol at -80 ˚C.

383 *GFP reporter Assays*

384 Reporter strains were adapted to M9 minimal media (MM) (see supplemental methods for strain
385 specific minimal media recipes) supplemented with 0.5 % (wt/vol) glucose as the sole carbon
386 source in 3 overnight passages, and stored in MM at -80 ˚C in 25 % (vol/vol) glycerol. Adapted
387 strains were cultured in MM + 0.5 % (wt/vol) glucose and passaged to MM + 0.5 % (wt/vol)
388 glucose with or without a second carbon-source (40 mM glutamic Acid, 40 mM a-ketoglutaric
389 acid, or 20 mM butyric acid, unless otherwise specified). After 24-hours of growth, samples were
390 diluted 1:100 in 1X PBS, and fluorescence was measured by flow cytometry on the BD Accuri C6
391 (BD Biosciences, San Jose, CA). Autofluorescence was gated out with FlowJo (BD Biosciences,
392 San Jose, CA), using a non-fluorescent strain of P. putida KT2440 carrying an empty vector
393 plasmid for reference (Supplementary Figure 8). To remove noise, the GFP mean for samples with
394 less than 150 events after gating was set to 0 (Supplementary Figure 9). Otherwise, the GFP mean
395 of the remaining events after gating was reported. Statistical significance was determined by T-
396 test between replicates.

397 **Resource Availability**

398 *Data and Code Availability*

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

401 DAP-seq data have been deposited in NCBI's Gene Expression Omnibus[55] and are accessible
402 through GEO Series accession number GSE157075
403 (https://www.ncbi.nlm.nih.gov/geo/query/acc.cgi?acc=GSE157075).

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

405 ## References

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

418 architectures. *Methods Mol. Biol.* **1910,** 469–504 (2019).

419 7. Ortega, Á., Zhulin, I. B. & Krell, T. Sensory repertoire of bacterial chemoreceptors.
420 *Microbiol. Mol. Biol. Rev.* **81,** (2017).

421 8. Briegel, A. *et al.* Universal architecture of bacterial chemoreceptor arrays. *Proc Natl Acad*
422 *Sci USA* **106,** 17181–17186 (2009).

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

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

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

430 12. Hug, L. A. *et al.* A new view of the tree of life. *Nat. Microbiol.* **1,** 16048 (2016).

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

433 14. Tatke, G., Kumari, H., Silva-Herzog, E., Ramirez, L. & Mathee, K. Pseudomonas
434 aeruginosa MifS-MifR Two-Component System Is Specific for α-Ketoglutarate Utilization.
435 *PLoS ONE* **10,** e0129629 (2015).

436 15. Lundgren, B. R. *et al.* Genetic analysis of the assimilation of C5-dicarboxylic acids in
437 Pseudomonas aeruginosa PAO1. *J. Bacteriol.* **196,** 2543–2551 (2014).

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

441 17. Rajeev, L., Garber, M. E. & Mukhopadhyay, A. Tools to map target genes of bacterial two-
442 component system response regulators. *Environ. Microbiol. Rep.* **12,** 267–276 (2020).
443 18. Garber, M. E. *et al.* Multiple signaling systems target a core set of transition metal
444 homeostasis genes using similar binding motifs. *Mol. Microbiol.* **107,** 704–717 (2018).
445 19. Skerker, J. M. *et al.* Rewiring the specificity of two-component signal transduction systems.
446 *Cell* **133,** 1043–1054 (2008).
447 20. Pougach, K. *et al.* Duplication of a promiscuous transcription factor drives the emergence of
448 a new regulatory network. *Nat. Commun.* **5,** 4868 (2014).
449 21. Capra, E. J., Perchuk, B. S., Skerker, J. M. & Laub, M. T. Adaptive mutations that prevent
450 crosstalk enable the expansion of paralogous signaling protein families. *Cell* **150,** 222–232
451 (2012).
452 22. Casas-Pastor, D. *et al.* Expansion and re-classification of the extracytoplasmic function
453 (ECF) σ factor family. *BioRxiv* (2019). doi:10.1101/2019.12.11.873521
454 23. Ronneau, S. & Hallez, R. Make and break the alarmone: regulation of (p)ppGpp
455 synthetase/hydrolase enzymes in bacteria. *FEMS Microbiol. Rev.* **43,** 389–400 (2019).
456 24. Dalebroux, Z. D. & Swanson, M. S. ppGpp: magic beyond RNA polymerase. *Nat. Rev.*
457 *Microbiol.* **10,** 203–212 (2012).
458 25. Potvin, E., Sanschagrin, F. & Levesque, R. C. Sigma factors in Pseudomonas aeruginosa.
459 *FEMS Microbiol. Rev.* **32,** 38–55 (2008).
460 26. Soucy, S. M., Huang, J. & Gogarten, J. P. Horizontal gene transfer: building the web of life.
461 *Nat. Rev. Genet.* **16,** 472–482 (2015).
462 27. Smanski, M. J. *et al.* Synthetic biology to access and expand nature's chemical diversity.
463 *Nat. Rev. Microbiol.* **14,** 135–149 (2016).

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

487 for next generation sequencing (FA-NGS) library preparation. *BMC Genomics* **21,** 85
488 (2020).
489 39. Bolger, A. M., Lohse, M. & Usadel, B. Trimmomatic: a flexible trimmer for Illumina
490 sequence data. *Bioinformatics* **30,** 2114–2120 (2014).
491 40. Silva, G. G. Z., Cuevas, D. A., Dutilh, B. E. & Edwards, R. A. FOCUS: an alignment-free
492 model to identify organisms in metagenomes using non-negative least squares. *PeerJ* **2,**
493 e425 (2014).
494 41. Langmead, B., Trapnell, C., Pop, M. & Salzberg, S. L. Ultrafast and memory-efficient
495 alignment of short DNA sequences to the human genome. *Genome Biol.* **10,** R25 (2009).
496 42. Li, H. *et al.* The Sequence Alignment/Map format and SAMtools. *Bioinformatics* **25,** 2078–
497 2079 (2009).
498 43. Kharchenko, P. V., Tolstorukov, M. Y. & Park, P. J. Design and analysis of ChIP-seq
499 experiments for DNA-binding proteins. *Nat. Biotechnol.* **26,** 1351–1359 (2008).
500 44. Bailey, T. L. *et al.* MEME SUITE: tools for motif discovery and searching. *Nucleic Acids*
501 *Res.* **37,** W202-8 (2009).
502 45. Emms, D. M. & Kelly, S. OrthoFinder: phylogenetic orthology inference for comparative
503 genomics. *Genome Biol.* **20,** 238 (2019).
504 46. Mironov, A. A., Vinokurova, N. P. & Gelfand, M. S. Software for analysis of bacterial
505 genomes. *Mol Biol (NY)* **34,** 222–231 (2000).
506 47. Thompson, M. G. *et al.* Massively Parallel Fitness Profiling Reveals Multiple Novel
507 Enzymes in Pseudomonas putida Lysine Metabolism. *MBio* **10,** (2019).
508 48. Rand, J. M. *et al.* A metabolic pathway for catabolizing levulinic acid in bacteria. *Nat.*
509 *Microbiol.* **2,** 1624–1634 (2017).

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

525 Acknowledgements

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

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

Funding source:

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

Author Contributions

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

554 Declarations of Interest

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