Characterization of an ECF56-family sigma factor from *Streptomyces venezuelae* reveals a highly conserved regulome

Mitchell G. Thompson*[1,2,3], Amin Zargar*[1,2,4], Pablo Cruz-Morales[1,2,4], Tristan De Rond[1,2,5], Samantha Chang[1,2], Allison N. Pearson[1,2], Garima Goyal[1,2,6], Jesus F. Barajas[1,2,6], Jacquelyn M. Blake-Hedges[1,2,5], Ryan M. Phelan[1,2], Victor Reyes-Umana[1,2,3], Amanda C. Hernández[1,2], Nathan J. Hillson[1,2,6], Patrick M. Shih[1,2,7,8], Jay D. Keasling[1-2,9,10,11,12,13]

[1]Joint BioEnergy Institute, Emeryville, CA, USA

[2]Biological Systems & Engineering Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[3]Department of Plant and Microbial Biology, University of California-Berkeley, Berkeley, CA, USA

[4]QB3 Institute, University of California-Berkeley, 5885 Hollis Street, 4th Floor, Emeryville, CA 94608, United States

[5]Department of Chemistry, University of California-Berkeley, Berkeley, CA, USA

[6]Department of Energy Agile BioFoundry, Emeryville, CA, USA

[7]Department of Plant Biology, University of California-Davis, Davis, CA, USA

[8]Genome Center, University of California-Davis, Davis, CA, USA

[9]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA

[10]Department of Bioengineering, University of California-Berkeley, Berkeley, CA 94720, USA

[11]Department of Chemical and Biomolecular Engineering, University of California-Berkeley, Berkeley, CA, USA

[12]The Novo Nordisk Foundation Center for Biosustainability, Technical University of Denmark, Denmark

[13]Center for Synthetic Biochemistry, Shenzhen Institutes for Advanced Technologies, Shenzhen, China

*Authors contributed equally

Corresponding author
Jay D. Keasling
jdkeasling@lbl.gov

Abstract

Bacteria often possess alternative sigma factors that initiate the transcription of specific genes under environmental stresses, the largest and most diverse group being the extracytoplasmic function (ECF) sigma factors. The regulation of ECF activity is crucial for ensuring the distinct transcription of stress responsive genes only occurs under the appropriate conditions. While most ECFs are comprised of only the core σ_2 and σ_4 regions, a unique form of ECF sigma factor regulation also contains a C-terminal extension bearing homology to the NTF2 superfamily of protein domains. While previous work has shown that this NTF2 domain can affect transcriptional activity *in vivo* in ECF41 and ECF42, its role in the newly classified ECF56 subgroup is unknown. In this work, we show that truncation of the C-terminus of the ECF56 sigma factor SVEN_4562 of *Streptomyces venezuelae* upregulates its activity in a hybrid assay. Through transcriptomics in *S. venezuelae*, we found that this truncated ECF56 sigma factor has a highly conserved promoter sequence *in vivo*. Bioinformatic assays illustrated that deep branches of the *Actinobacteria* phylum contained putative ECF56 promoter motifs identical to those found in the *S. venezuelae* ECF56 regulon. We validated these findings through *ex situ* hybrid assays illustrating that truncated ECF56 sigma factors from phylogenetically diverse *Actinobacteria* activate transcription from these promoters. Importantly, our work shows that the genetic infrastructure of the

ECF56 family of sigma factors is highly conserved and performs important functions yet to be understood in *Actinobacteria*.

Importance

Most ECF sigma-factors rely on anti-sigma factor regulation; in contrast, the unique classes of ECF sigma-factors that contain a C-terminal extension are thought to respond directly to an environmental signal. Here we show that the *cis*-acting regulatory element of the ECF56 regulon is likely highly conserved in many *Actinobacteria*, with exact nucleotide level conservation over ~2 billion years of evolution. The high conservation of this genetic architecture, as well as a conserved gene content within the regulon, strongly point to a specialized and important role in *Actinobacteria* biology.

Introduction

Sigma factors are required for the initiation of transcription in all known bacteria. They are able to recruit the RNA polymerase (RNAP) core, direct it to a specific promoter region, and initiate transcription by enabling promoter melting (1). While most genes in a given bacterium are transcribed using a “housekeeping” sigma factor, bacteria often possess alternative sigma factors that initiate transcription of specific subsets of genes in response to specific environmental stresses (1). Families of alternative sigma factors are involved in responses to stress such as responding to heat shock (RpoH-like), flagellar synthesis (FliA-like), and nitrogen limitation (RpoN-like) (1). The largest and most diverse group of alternative sigma factors are the Extracytoplasmic Function (ECF) sigma factors (1). ECF sigma factors are smaller in size compared to housekeeping σ^{70} sigma factors, and normally possess only two of the

four conserved domains, σ_2 and σ_4. These domains are sufficient to recognize the -10 and -35 elements of a bacterial promoter, recruit RNAP core, and initiate transcription via promoter melting (1). Despite their reduced size, ECF sigma factors play a role of outsized importance in the transduction of environmental signals in bacteria (2–4).

The regulation of ECF activity is critical to ensure that the transcription of stress responsive genes only occurs under the appropriate conditions. In many cases, ECFs are negatively regulated by cognate anti-sigma factors, which are usually co-transcribed as part of a negative feedback loop (5). Anti-sigma factors physically sequester ECFs until an appropriate environmental signal either leads to the physical destruction of the anti-sigma factor or alters the affinity of the anti-sigma factor for its cognate ECF (6). The mechanism that causes dissociation varies widely depending on the ECF sigma factor, but can take the form of a proteolysis event, export of the anti-sigma factor, or inactivation of the anti-sigma factor through phosphorylation (5–7).

Another form of ECF sigma factor regulation has been described more recently, via C-terminal extensions of ~100 amino acids that show homology to members of the NTF2 superfamily of protein domains (8). The NTF2 superfamily of domains is widely distributed on the tree of life and shows little conservation at the primary sequence identity level; but members include many enzymes and a recent review of their function has implicated allosteric regulation of DNA binding proteins (8). Studies of these NTF2 domains in ECF41 and ECF42 found it was critical to initiation of gene transcription and strongly implicate that it directly responds to an external signal (9–12). In 2015, a genomic survey of ECFs and other signal transduction proteins in *Actinobacteria* revealed another group of ECFs with similar C-terminal extensions: the ECF56

subgroup (13). Previous work with the *Mycobacterium tuberculosis* ECF56 family sigma factor Rv0182c (*sigG*) implicated involvement in DNA damage response pathways (14, 15). While genes regulated by SigG were identified, neither the physiological role these genes play nor the environmental cue SigG responds to were identified. To date, the specific signal that is transduced by the NTF2 domain of ECF56 or any other ECF family members has yet to be identified.

In an attempt to shed light on the physiological role of the ECF56 subfamily, we characterized the regulatory activity of the NTF2 domain of ECF56 sigma factors *ex situ* in *Escherichia coli* and *in vivo* in *S. venezuelae*. Beyond its attributes as a fast-growing, genetically tractable *Actinobacteria*, *S. venezuelae* has also been previously used to study the NTF2 domain of the ECF42 sigma factor subfamily (10). In this work, we identify a highly conserved promoter that defines an ECF56 regulon in *S. venezuelae* and demonstrate through bioinformatics that the regulon is highly conserved through deep branches of *Actinobacteria*. Lastly, we further validate our findings regarding the role of these C-terminal extensions by determining their role in diverse ECF56 sigma factors.

Results

The C-terminal extension of ECF56 sigma factors inhibits transcription

We first aimed to determine the regulatory role of these C-terminal extensions in ECF56 sigma factors. To start, we sought to determine the differential transcriptional activity with and without the C-terminal extension of the ECF56 sigma factor from *S. venezuelae*, SVEN_4562 (**Figure 1A**). Previously, Wecke and colleagues had identified

a conserved DGGGK motif downstream of the σ_4 domain in ECF41 family sigma factors (**Figure S1A**). Truncating the C-terminus of these sigma factors immediately after this motif resulted in increased transcription from cognate promoters (12). By aligning >900 ECF56 sigma factors, we observed a highly conserved MPP(F/Y/L) motif similarly positioned relative to the σ_4 domain (**Figure S1B**). As ~60% of ECF56 sigma factor sequences contained an aromatic phenylalanine or tyrosine after the highly conserved MPP, we truncated SVEN_4562 immediately after its internal MPPY to maintain the aromatic residue, resulting in SVEN_4562$_{\Delta 237\text{-}323}$.

To test this hypothesis we measured transcription from a previously identified *Rv0182c* target promoter (*Rv0887cp*) from *M. tuberculosis* in an *ex situ* assay (15). In *E. coli*, the transcription of each sigma factor was driven by a leaky uninduced lacUV5 promoter, and its ability to bind and drive expression of the *Rv0887c* promoter was determined through a GFP reporter protein (**Figure 1B**). Compared to the empty vector control, we observed a 2.4-fold increase in GFP expression in the strain overexpressing the truncated ECF56 sigma factor. In contrast, overexpression of the full ECF56 sigma factor resulted in a 1.1-fold increase in GFP expression. These observations suggest that the C-terminal extension of ECF56 sigma factors inhibits promoter binding. Interestingly, this contrasts with other investigations of this domain in *Actinobacteria*, which found that the full C-terminal extension was required for complete activation of transcription (10–12).

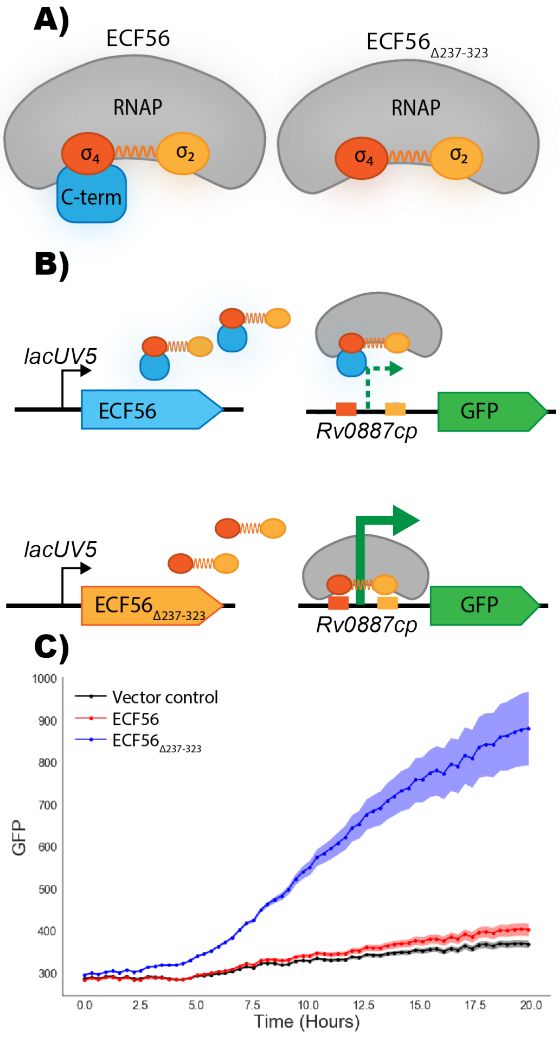


Figure 1. Architecture, and characterization of ECF56 sigma factors. A) Schematic illustrates the interactions between RNA polymerase and ECF sigma factors with and without the NTF2 domain. **B)** The schematic illustrates the genetic circuit built in *E. coli*. Wild-type ECF56 sigma factor(SVEN_4562) or a truncated variant is constitutively expressed and results in the transcription of a GFP cassette driven by the known ECF56-dependent promoter. **C)** Graph shows GFP fluorescence in arbitrary units over time in strains expressing variants of SVEN_4562 or an empty vector control. Shaded area shows 95% confidence intervals, n=3.

Defining the ECF56 Regulon of *S. venezuelae*

As the ECF56 family is widely distributed in *Actinobacteria (13)*, we hypothesized that they must play an important physiological role in these organisms, such as *S. venezuelae*. Therefore, we cloned the truncated SVEN_4562 under the strong constitutive promoter kasOp and genetically inserted it into the S. venezuelae genome at the *phiC* attachment site. Constitutive expression of truncated SVEN_4562 produced a bald phenotype (**Figure 2A**). This is especially interesting because C-terminal truncations of ECF42 in *S. venezuelae* showed no observable differences in morphology (10).

Through transcriptomics, we sought to develop a greater understanding of the physiological role of ECF56 in S. venezuelae. Using RNA sequencing (RNA-seq), we compared the transcriptomes of a strain overexpressing the truncated ECF56 to the wild-type strain and observed a population of genes that were heavily upregulated (**Figure 2B**). The complete list of differentially expressed genes can be found as **Supplemental Attachment 1**. Our results showed that 22 loci had a *b* of >3.5 and a *q-value* less than $1x10^{-10}$, representing the most highly upregulated genes. However, previous work in studies where transcription factors were overexpressed had noticed significant read through between genes that could potentially confound results (16). To account for this possibility we manually investigated reads mapped to the *S. venezuelae* chromosome to identify instances in which increased differential expression was likely due to read through (or direct overexpression in the case of SVEN_4562) and identified 12 genes likely to be false hits (**Table S1**). The resulting genes that were specifically upregulated by the truncated ECF56 sigma factor were amongst the most highly expressed in the cell, with transcript per million (TPM) values either close to or as high

as genes such as *rpoD* and *EF-Tu* **(Figure 2C).** Unfortunately, none of the genes found to be upregulated and controlled by SVEN_4562 directly had a known function. Of the ten genes, six were annotated as proteins of hypothetical function, whereas the remaining genes were annotated in broadly functional families such as glyoxalase, hydrolase, dihydrofolate reductase, and deaminase-reductase (**Table 1**). Interrogation of the upstream sequences of the genes in **Table 1** revealed a nearly perfectly conserved promoter motif (CGATGAGTTN$_{14}$GGTC - **Figure 2D**). Even more striking was that this promoter motif was also perfectly conserved in the distantly related *M. tuberculosis* ECF56 sigma factor regulated *Rv0887c* (**Figure 2D**).

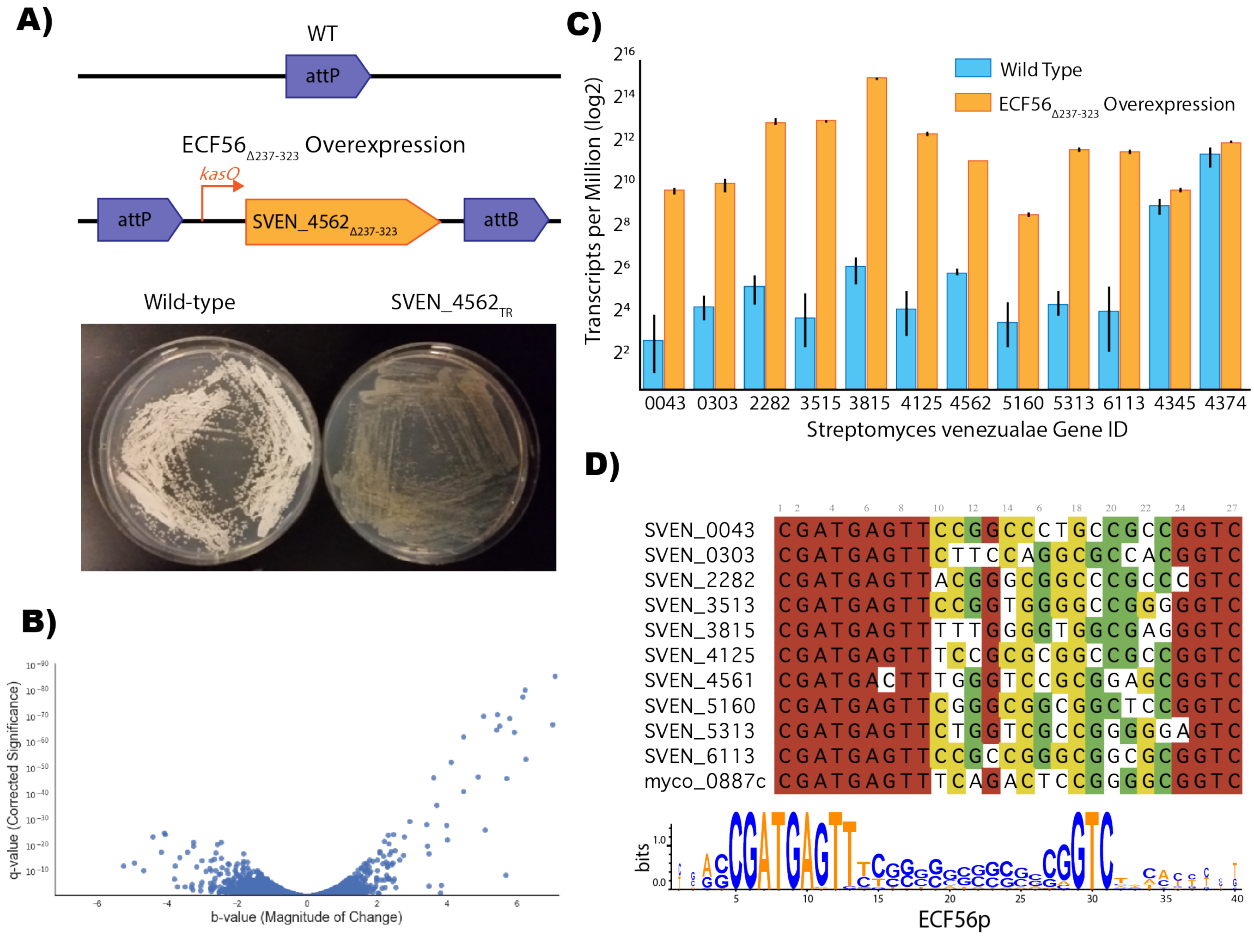


Figure 2. Identification of genes regulated by SVEN_4562. A) A truncated version of ECF56 from *S. venezuelae* (SVEN_4562) driven by the kasOp promoter was inserted into the genome of *S. venezuelae*

188 resulting in a bald phenotype when compared to the wild-type control. **B)** Volcano plot showing RNA-Seq
189 analysis of transcripts of the truncated SVEN_4562 overexpressing strain compared to wild-type *S.*
190 *venezuelae*. The y-axis shows the q-value (corrected significance), while the x-axis shows b-value
191 (magnitude of change). Results represent n=3 **C)** Genes believed to be directly upregulated when
192 truncated SVEN_4562 is overexpressed, in addition to *rpoD* (SVEN_4345) and *EF-Tu* (SVEN_4374).
193 Error bars show 95% confidence interval, n=3. **D)** Top panel illustrates an alignment of the upstream
194 sequences of these upregulated genes and the bottom panel represents a position weight matrix of these
195 sequences.
196

Gene ID	**Description**	***b***	***q-value***
SVEN_3513	putative hydrolase	7.08	3.89E-86
SVEN_3815	hypothetical protein	6.52	2.47E-93
SVEN_4125	hypothetical protein	6.24	8.38E-81
SVEN_6113	Dihydrofolate reductase	5.92	2.27E-64
SVEN_2282	Glyoxalase family protein	5.79	8.97E-70
SVEN_0043	hypothetical protein	5.71	2.58E-46
SVEN_5313	Bifunctional deaminase-reductase domain protein	5.50	1.41E-66
SVEN_0303	hypothetical protein	4.47	1.98E-41
SVEN_5160	hypothetical protein	4.13	1.33E-52
SVEN_4561	hypothetical protein	3.70	4.14E-36

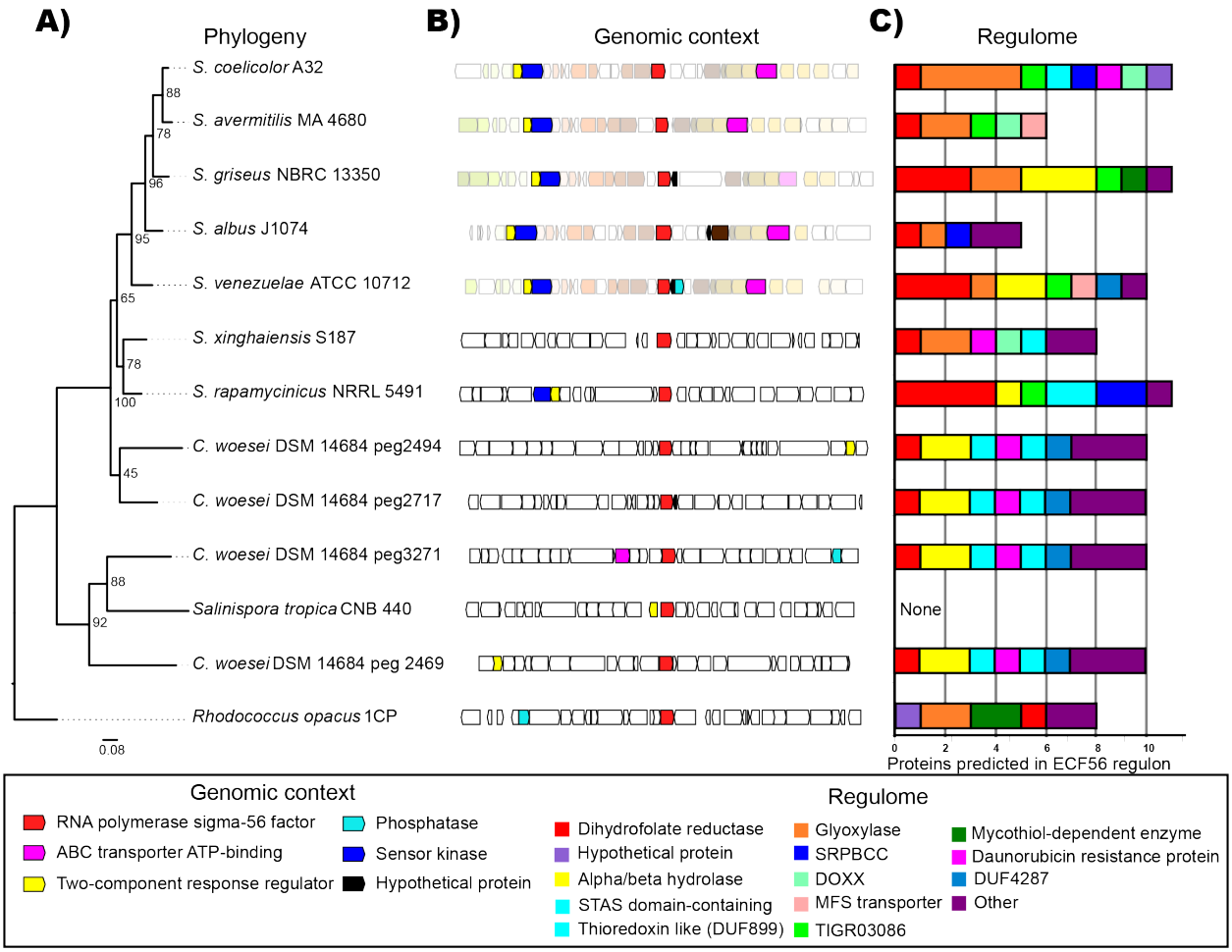
197

198 **Table 1: Genes predicted to be regulated by SVEN_4562.** Genes that were found to be highly
199 upregulated during the overexpression of truncated SVEN_4562, and not to be the result of suspected
200 readthrough. The b-value is a measure of change in expression (17), where as the qval is a corrected
201 significance value.
202
203 Predicting the ECF56 Regulome
204 Given the highly conserved nature of the *S. venezuelae*, and *M. tuberculosis*
205 promoters we sought to investigate the genomic context of ECF56 homologs in selected
206 Actinobacterial genomes. We searched for homologs of ECF56 in a database including
207 720 high quality genomes from actinobacterial species. We retrieved 599 homologs
208 from 414 genomes, mostly from Streptomycetaceae, Mycobacteria and
209 Pseudonocardiaceae. We found that most *Streptomyces spp.* (75%) included a single
210 ECF56 homolog. A phylogenetic reconstruction (**Supplemental Attachment 2**)
211 showed that these single copy-ECF56 homologs are monophyletic, we also found that
212 these homologs share a high degree of synteny among streptomycetes.
213 After confirming the broad distribution and conservation of the ECF56 we then
214 sought to identify putative members of regulons in distinct species. For this analysis we
215 selected ten representative genomes and searched for the ECF56 consensus binding
216 sequence "CGATGAGTTN$_{15}$GTC". This search revealed that many proteins belonging
217 to the glyoxylase, dehydrofolate reductase and alpha/beta hydrolase families among
218 others are under the control of ECF56 across largely divergent actinobacterial species.
219 Two interesting exceptions were found: only one protein (a member of the glyoxylase
220 family) is located upstream of the ECF56 binding sequence in *M tuberculosis* while

221 none could be found in *Salinispora tropica* which nevertheless includes a bonafide

222 ECF56 homolog.

223



224

225 **Figure 3. Distribution of ECF56 across Actinobacteria. A,** Phylogenetic distribution of representative

226 orthologs from selected Actinobacteria species, Most *Streptomyces* genomes include only one copy,

227 while C. woesei includes four. **B**. Genomic neighborhood for each ECF56 ortholog in red. The genes are

228 colored according to their homology. Faded colors represent the genome neighborhood conserved in

229 recently diverged orthologs. Solid colors represent conserved genes within the genomic neighborhood,

230 their annotation is shown at the bottom. **C.** Abundance and functions of the genes regulated by ECF56

231 in each species. Each bar accounts for the number of the proteins predicted (and confirmed in the case of

232 *S. venezuelae*), to be regulated by ECF56. The functions are represented by different colors shown in
233 the box at the bottom.
234
235 Conservation of target promoter of ECF56 across phyla
236 Lastly, we wanted to investigate the potential upregulation of transcription of the
237 truncated ECF56 in various species, as similar studies in ECF41 have shown varying
238 results depending on the coding sequence. As such, we cloned truncated versions of
239 two more ECF56 from *C. woesei* and *M. tuberculosis* (**Figure 4A**). *C. woesei* was
240 selected due to its divergence from Streptomycetes ~2 billion years ago (18), yet
241 bioinformatic analysis revealed the identical promoter sequences were conserved in its
242 genome. *M. tuberculosis* was selected due to its importance as a public health threat,
243 and previous work that has been done on the Rv0182c regulon (14, 15). As with
244 SVEN_4562, both of these homologs were truncated immediately past a conserved
245 MPPF motif in the case of Cwoe_2709, and an MPPY motif for Rv0182c. We also
246 cloned an additional ECF dependent promoter, SVEN_3513p, to compare the
247 differences due to the regions varying between the -35 and -10 elements. All ECF56
248 truncations performed similarly in the *ex situ* assay, while the promoter of M.
249 tuberculosis showed greater upregulation than the promoter of SVEN_3513, suggesting
250 that the intergenic -35 and -10 region does affect transcription.

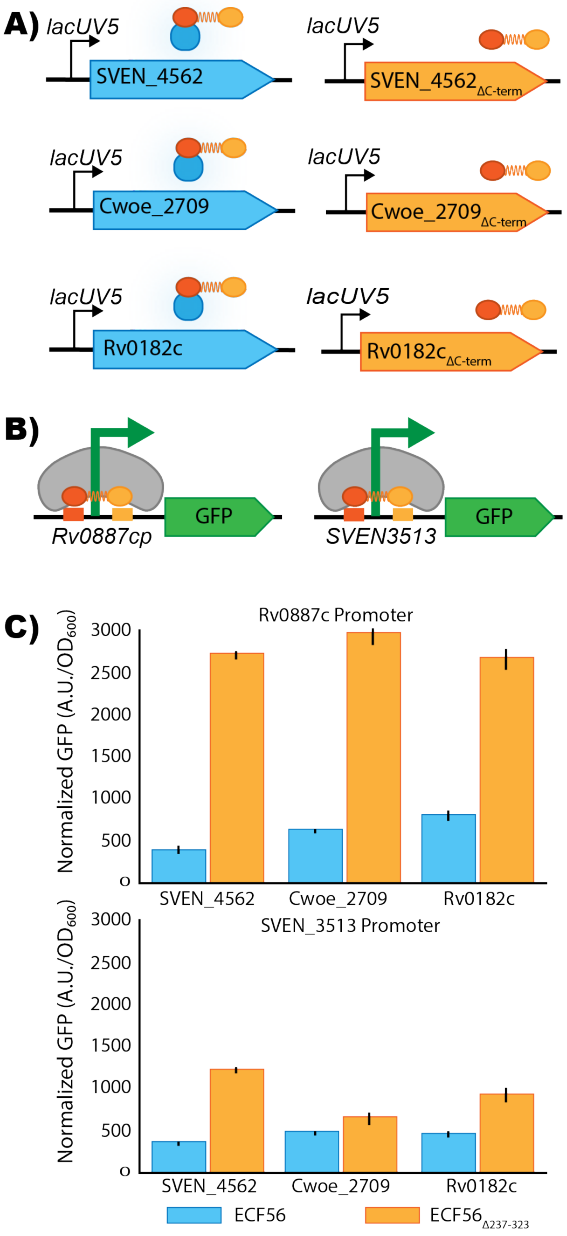


Figure 4. Genetic architecture of ECF56 across phyla. A) A wild-type and truncated version of three phylogenetically distinct ECF56 promoters are driven by the lacUV5 promoter were constructed **B)** Promoters from *M. tuberculosis* (RV0887c) and *S. venezuelae* (SVEN_3513) transcribe the GFP reporter in the presence of the ECF sigma factor **C)** *ex situ* assay illustrating the overnight GFP expression levels in *E. coli* normalized to cell density. Error bars represent 95% confidence intervals, n=3.

Discussion

Traditionally, sigma factors have been thought to rely on one or more trans-acting partners to transduce specific environmental cues in order to activate responsive gene expression (5). Recent work has challenged this notion, with the discovery that many C-terminal extensions modulate the activity of a variety of ECF sigma factor families (19). These domains presumably are able to sense environmental signals, which results in conformational changes that activate expression of their cognate regulon, effectively turning them into one-component systems (19). However despite recent efforts, which have defined multiple putative regulons of these ECF sub-families, the physiological roles that these ECFs regulate and the signals that they respond to remain a mystery (19).

With its unique role in ECF transcription, efforts have been made to study the NTF2 domain and its interaction with the target promoter and RNA polymerase. Initial studies in the ECF41 subfamily in *B. lincheniformis* (Firmicutes) and *R. sphaeroides* (α-proteobacteria) found that truncations of the NTF2 domain drastically upregulated transcription of a highly conserved promoter (12). Interestingly, biochemical and crystal structure studies of the ECF41 σ^J of *M. tuberculosis* (*Actinobacteria*) and the ECF42 of S. venezuelae (*Actinobacteria*) revealed that the NTF2 domain was necessary for transcription (9, 10). While a direct coupling analysis of ECF41 and ECF42 revealed that the NTF2 region bind and regulate the core ECF regions of σ_2 and σ_4 responsible for transcription (11), the role of these NTF2 domains in the ECF56 subfamily has not yet been studied. Crystallographic studies of ECF56 homologs may provide crucial information on how NTF2 domains impose negative regulation.

The only previously studied member of the ECF56 subfamily is Rv0182c (SigG) of *M. tuberculosis*. SigG was shown to be the most weakly expressed sigma factor in *M. tuberculosis* when grown in laboratory conditions, potentially making it challenging to study (20). However microarray analysis revealed *sigG* to be one of the most upregulated genes when *M. tuberculosis* was grown during human macrophage infection prompting further research (21). Further work then showed that SigG was upregulated in the presence of the DNA damaging agent mitomycin C, and that it controlled the expression of two proteins that contained glyoxalase domains (14, 15). This connection to the DNA-damage response was recently confirmed when it was found that SigG is regulated by the PafBC system (22). PafBC has recently been shown to be the master transcriptional regulator of the LexA/RecA-independent DNA repair response in *M. tuberculosis (22)*. One of the main functions of this response is to proteolytically degrade RecA, to temporally control the stress response (22). This may suggest that the C-terminus of ECF56 sigma factors respond to specific DNA damage signals to activate a highly conserved stress response (**Figure 5**). Our transcriptomic investigation in *S. venezuelae* revealed putative glyoxalase proteins strongly upregulated with constitutive expression of the truncated ECF56 sigma factor. As our bioinformatic studies revealed many other *Actinobacteria* also had glyoxalases, dihydrofolate reductases, and α/β hydrolases in their predicted ECF56 regulons, it is conceivable that these enzymes are required to metabolize toxic byproducts of DNA damage. However, these protein families are known to have a wide range of functions, and further experimentation is clearly required to predict their exact function. We believe

303 that future work should focus heavily on the physiological and biochemical roles these
304 proteins play, which will further elucidate the signals that activate ECF56 sigma factors.
305 Rhodius et al. had previously proposed that ECF sigma factors may make
306 excellent parts for synthetic biology by demonstrating their exceptional orthogonality
307 (23). They further went on to show that domains from different ECF subfamilies can be
308 combined to generate functional proteins that recognize hybrid promoters (23). From
309 this work various groups have used ECF sigma factors to build highly insulated genetic
310 circuits (24), ECF sigma factor based toolkits for both *E. coli* and *Bacillus subtilis* (25,
311 26), and highly complex synthetic timers (27). All of these synthetic circuits take
312 advantage of the highly specific interaction ECF sigma factors have with their cognate
313 promoters. Based on the extreme conservation of the ECF56 promoter, the ECF56
314 subfamily is also likely to be highly orthogonal to other ECF promoters. While this
315 orthogonality is highly desirable in genetic engineering, the activity of most ECF sigma
316 factors relies on control by either cognate anti-sigma factors or synthetic genetic
317 circuitry. This excess genetic baggage may complicate some engineering tasks. The
318 prospect that C-terminal extended ECF sigma factor may be able to transduce a variety
319 of environmental cues is extremely intriguing. If the signal could be identified, these
320 proteins could function as highly orthogonal, tunable, stand alone genetic parts. Further
321 work on these unique transcription factors should heavily focus on deorphaning these
322 receptors.
323
324
325

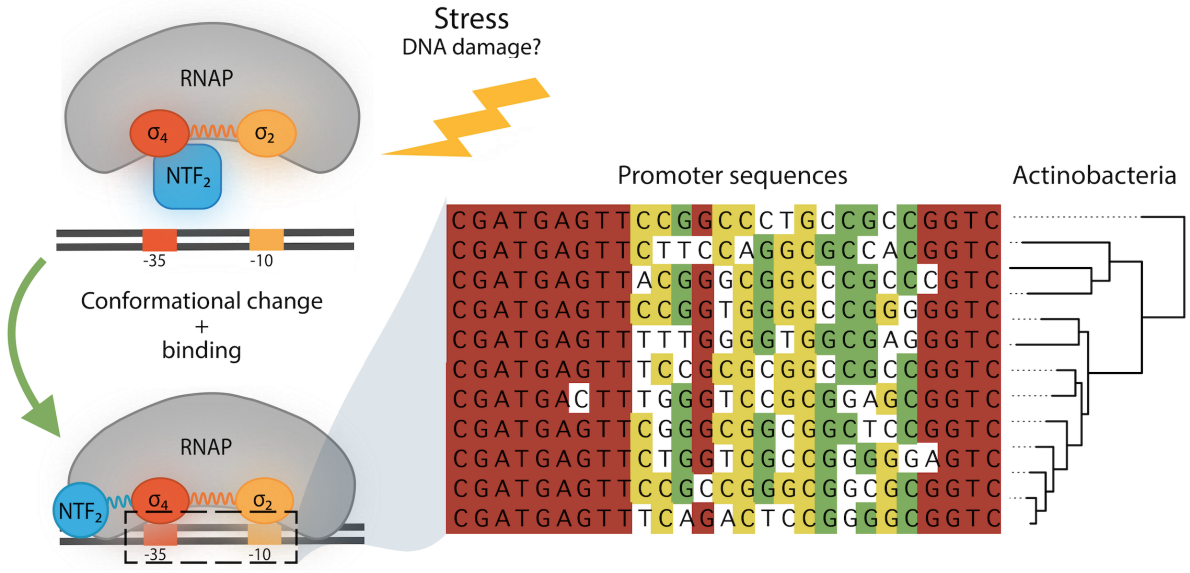


Figure 5: Investigations into sigma factor ECF56 in *S. venezuelae* uncovers that the removal of the inhibitory NTF2 domain results in constitutive transcription of a highly conserved promoter across vast phylogenetic distances.

Methods

Media, Chemicals, and Strains

E. coli bacterial cultures were grown in Luria-Bertani (LB) Miller medium (BD Biosciences, USA) at 37 °C. *S. venezuelae* ATCC 10712 was grown on MYM media (28), and sporulated on mannitol soy (MS) agar media supplemented with 10 mM magnesium chloride (28) . Cultures were supplemented with carbenicillin (100 mg/L, Sigma Aldrich, USA), kanamycin (50 mg/L, Sigma Aldrich, USA), or spectinomycin (100 mg/L for *E. coli* or 400 mg/L for *S. venezuelae*, Sigma Aldrich, USA) All bacterial strains

343 and plasmids used in this work are listed in **Table 2** and are available through the public

344 instance of the JBEI registry. (https://public-registry.jbei.org/).

345

346 **Table 2: Strains and Plasmids used in this study**

Strains	**JBEI Number**	**Source**	**Notes**
E. coli DH10B		Thermo Fisher	
E. coli ET12567		ATCC BAA-525	
S. venezuelae		ATCC 10712	
Plasmids			
pUZ8002		(29)	
pSC-KasOp-mCherry		(30)	Spectinomycin
pSC-KasOp-SVEN_4562_trunc	TBD	This work	Spectinomycin
pBbE5k-RFP		(31)	Kan
pBbB5a-RFP		(31)	Amp
pBbE5k-SVEN_4562	TBD	This work	Kan
pBbE5k-SVEN_4562_trunc	TBD	This work	Kan
pBbE5k-Cwoe_2709	TBD	This work	Kan

pBbE5k-Cwoe_2709_trunc	TBD	This work	Kan
pBbE5k-Rv0182c	TBD	This work	Kan
pBbE5k-Rv0182c_trunc	TBD	This work	Kan
pBbBa Rv0877cp::GFP	TBD	This work	Amp
pBbBa SVEN_3513p::GFP	TBD	This work	Amp

347

348 DNA Manipulation

349 All plasmids were designed using Device Editor and Vector Editor software, while
350 all primers used for the construction of plasmids were designed using j5 software (32–
351 34). Plasmids were assembled via Gibson Assembly using standard protocols (35), or
352 Golden Gate Assembly using standard protocols(36). Plasmids were routinely isolated
353 using the Qiaprep Spin Miniprep kit (Qiagen, USA), and all primers were purchased
354 from Integrated DNA Technologies (IDT, Coralville, IA). *E. coli* codon optimized gene
355 blocks for *Rv0182c* and Cwoe_2709 were purchased from Integrated DNA
356 Technologies (IDT, Coralville, IA).

357

358 Conjugation into *Streptomyces venezuelae*

359 *E. coli* ET12567/pUZ8002 was transformed with an appropriate vector and
360 selected for on LB agar containing kanamycin (25 µg/mL), chloramphenicol (15 µg/mL),
361 and apramycin (50 µg/mL). A single colony was used to inoculate 5 mL of LB containing
362 kanamycin (25 µg/mL), chloramphenicol (15 µg/mL), and apramycin (50 µg/mL) at
363 37°C. The overnight culture was used to seed 10 mL of LB containing the same
364 antibiotics, which was grown at 37°C to an OD600 of 0.4-0.6. The *E. coli* cells were

pelleted by centrifugation, washed twice with LB, and resuspended in 500 µL of LB. Fresh *S. venezuelae* J1074 spores were collected from a mannitol soy agar plate with 5 mL of 2xYT and incubated at 50°C for 10 min. The spores (500 µL) and the *E. coli* cells (500 µL) were mixed, spread onto mannitol soy agar, and incubated at 30°C for 16 hours. After 1 mL addition of nalidixic acid (20 µg/mL) and apramycin (40 µg/mL) was added and allowed to dry, the plate was further incubated for 3-4 days at 30°C. A single colony was used to inoculate into TSB containing nalidixic acid (25 µg/mL) and apramycin (25 µg/mL). The overnight culture was spread onto a MS plate and incubated at 30^{o}C for 2-3 days. The spores were collected from the plate with 3-4 mL of water and mixed with glycerol to prepare 25% glycerol stock. The glycerol stock was stored at -80°C for long-term storage.

RNA Isolation and RNA-Seq

RNA was isolated from mycelial cultures grown for 24 hours on MYM media at 30 °C. Briefly, 1 mL of culture was mixed with 2 volumes of RNAlater (Invitrogen, USA) for five minutes at room temperature, pelleted in a Eppendorf 5245 centrifuge for 2 minutes at 10,000 rpm and decanted. Pelleted mycelia were then stored at -80 °C until RNA was extracted. RNA was isolated using QIAGEN RNeasy kit (Qiagen, USA) per manufacturer's instructions for gram-positive bacteria. Genomic DNA was removed with Ambion Turbo DNase (Invitrogen, USA). Bacterial rRNA was removed from total RNA using a Ribo-Zero RNA removal kit (Illumina, USA), and libraries for RNAseq were prepared using the NEBNext Ultra II RNA Library Prep Kit for Illumina (New England Biolabs, USA) according to the manufacturer's instructions. Libraries were then sequenced on a Illumina MiSeq using kit v3 (2 x 75-cycle) with paired-end reads.

388 RNAseq data was analyzed with Kallisto (37) which pseudo-aligned reads to the *S.*
389 *venezuelae* ATCC 10712 coding sequences (GCA_008639165.1). Sleuth was then
390 used to calculate differential expression (17). Raw transcript per million (TPM) values,
391 genome wide differential expression, and the top 100 most upregulated genes during
392 SVEN_4562 _trunc overexpression can be found in **Supplemental Attachment 1**. In
393 order to identify potential promoter readthrough, raw reads were mapped to the *S.*
394 *venezuelae* ATCC 10712 genome using Bowtie 2 (38), and mapped reads were
395 visualized with the Integrated Genomics Viewer (39). Genes that showed upregulation
396 as a consequence from readthrough were not considered for further promoter analysis.
397
398 Fluorescence Assays
399 Fluorescence assays were conducted as previously described with minor
400 changes (40). All endpoint assays were conducted in 96-deep well plates (Corning
401 Costar, 3960), with each well containing 500 **µ**L of LB medium with appropriate
402 antibiotics inoculated at 1% v/v from overnight cultures. Plates were sealed with
403 AeraSeal film (Excel Scientific, AC1201-02 and incubated at 37 C in a 250 rpm shaker
404 rack. After 24 hours, 100 **µ**L from each well was aliquoted into a black, clear-bottom 96-
405 well plate for measurements of optical density and fluorescence using an Infinite
406 F200Pro (Tecan Life Sciences, San Jose, CA) plate reader. Optical density was
407 measured at 600 nm (OD_{600}), while fluorescence was measured using an excitation
408 wavelength of 470 nm, an emission wavelength of 530 nm, and a manually set gain of
409 25. To perform time course assays, overnight cultures were inoculated into 10 mL of LB
410 medium from single colonies, and were grown at 37 °C. These cultures were then

411 diluted 1:100 into 100 uL of LB medium in 96-well plates (Falcon, 353072). Plates were
412 sealed with a gas-permeable microplate adhesive film (VWR, USA), and then optical
413 density and fluorescence were monitored for 20 hours in a Tecan Infinite F200Pro
414 (Tecan Life Sciences, San Jose, CA) plate reader. Optical density was measured at 600
415 nm (OD_{600}), while fluorescence was measured using an excitation wavelength of 470
416 nm, an emission wavelength of 530 nm, and a manually set gain of 25.
417
418 Bioinformatics
419 Sequences of ECF sigma factors were downloaded from Pfam -
420 https://pfam.xfam.org/family/Sigma70_ECF. An ECF56 sigma factor HMM was
421 generated as described previously (4), and ECF56 family sigma factors were identified
422 using hmmscan (41). ECF56 alignments were done using the MAFFT-LINSI algorithm
423 (42) and visually inspected using UGENE software (43). Alignments were restricted to
424 positions that had greater than 50% occupancy using Protein Dynamics and Sequence
425 Analysis (ProDy) (44, 45). All conserved sequence motifs were visualized with Weblogo
426 (46).
427 We mined for ECF56 homologs in actinobacteria using the amino acid sequence
428 of SVEN_4562 as query for searches using BlastP with an E-value cutoff of1E-12 and a
429 bit-score cutoff of 200. Independent BlastP searches were done with each one of 720
430 high quality genomes. The retrieved protein sequences were aligned using muscle
431 (47), and trimmed using jalview (48). An ECF56 phylogenetic tree was obtained using
432 IQtree with the resulting amino acid matrix the best amino acid substitution model for
433 the tree was selected using the ModelFinder utility implemented in IQtree (49). Node

support comes from 10000 bootstrap generations. For analysis of the genome context of the selected ECF56 homologs shown in figure 3 we used CORASON with an adhoc genomes database including selected organisms (50). the parameters for the search were e-value 1E-12 and bit-score cutoff of 200, the graphical output was modified to follow the order of a phylogenetic tree generated with the protein sequences homologs identified by CORASON with identical parameters to those described above.

The regulon of each organism shown in figure 3 was obtained by searching for the ECF56 binding consensus sequence “CGATGAGTTnnnnnnnnnnnnnnGTC” in each representative genome using the Artemis genome browser. Genes were considered ECF56 regulon members if located immediately downstream of the ECF56 binding sequence. Members of the regulon were manually annotated by performing blast searches against the NR database and searching for annotated domains in the hmmscan server (51).

Acknowledgements

The authors would like to thank Matthew Traxler for helpful discussions during the preparation of this manuscript. This work was part of the DOE Joint BioEnergy Institute (https://www.jbei.org) supported by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, and was part of the Agile BioFoundry (http://agilebiofoundry.org) supported by the U.S. Department of Energy, Energy Efficiency and Renewable Energy, Bioenergy Technologies Office, through contract DE-AC02-05CH11231 between Lawrence Berkeley National Laboratory and the U.S. Department of Energy. This work was also funded by National Institutes of

Health Awards, F32GM125179. The views and opinions of the authors expressed herein do not necessarily state or reflect those of the United States Government or any agency thereof. Neither the United States Government nor any agency thereof, nor any of their employees, makes any warranty, expressed or implied, or assumes any legal liability or responsibility for the accuracy, completeness, or usefulness of any information, apparatus, product, or process disclosed, or represents that its use would not infringe privately owned rights.

Contributions

Conceptualization, M.G.T., A.Z.; Methodology, M.G.T., A.Z., P.C.M., T.D.R.;Investigation, M.G.T., A.Z., P.C.M., T.D.R., S.C., R.M.P., G.G., A.N.P., J.F.B., J.M.B., V.R.; Writing – Original Draft, M.G.T, A.Z., P.C.M.; Writing – Review and Editing, All authors.; Resources and supervision N.J.H., P.M.S., J.D.K.

Supplementary Material

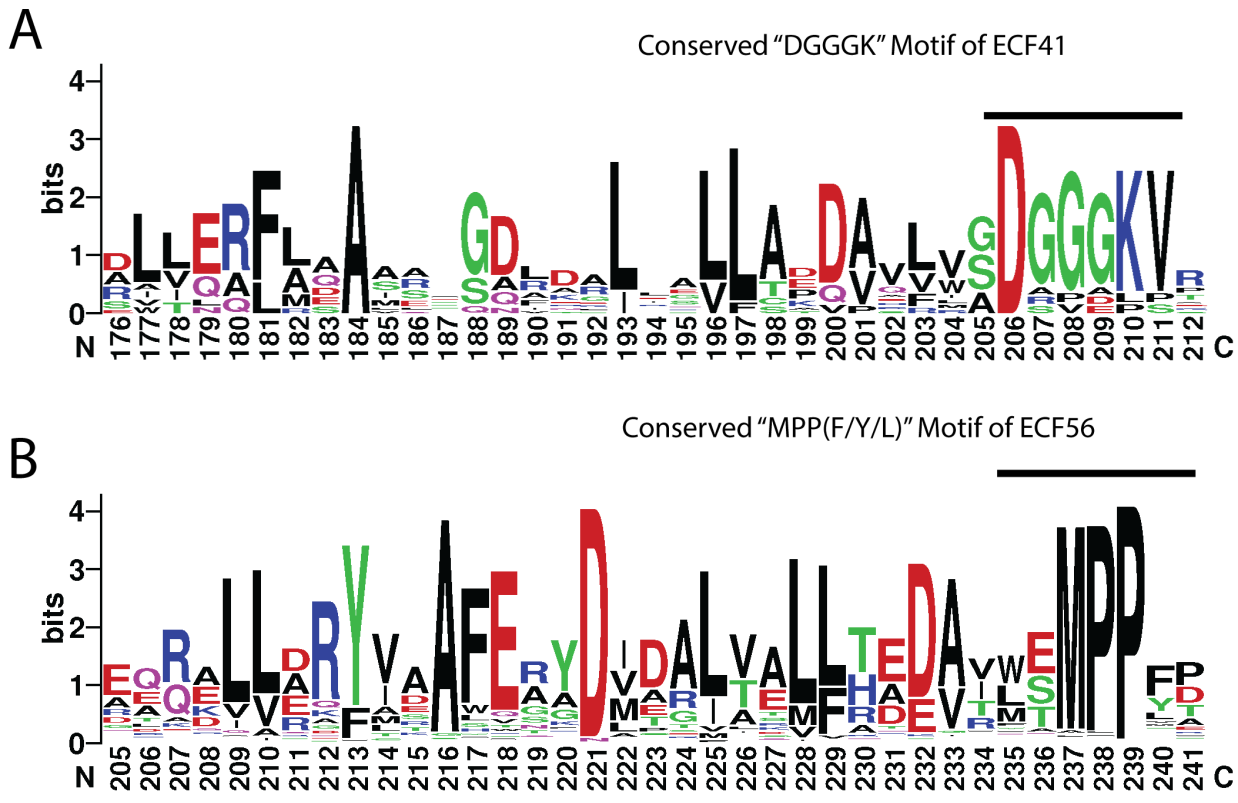


Figure S1. Conserved motifs of ECF41 and ECF56 that delineate truncation sites. Conserved residues of the beginning of the NTF2 domain of both ECF41 (A), and ECF56 (B) sigma factors. The conserved motifs DGGGK of ECF41 and MPP(F/Y/L) of ECF56 are highlighted.

Table S1: Suspected read through or overexpression artifacts from RNA-Seq analysis. Genes that were found to be highly upregulated as a consequence of read through during the overexpression of truncated SVEN_4562. The b-value is a measure of change in expression (17), where as the qval is a corrected significance value

Gene ID	Description	*b*	*q-value*
SVEN_2281	3-oxoacyl-[acyl-carrier protein] reductase	7.02	2.23E-67
SVEN_3512	Thermophilic serine proteinase precursor	6.25	6.09E-54

SVEN_3511	hypothetical protein	6.16	6.63E-78
SVEN_6114	Permease of the drug or metabolite transporter superfamily	5.44	3.63E-71
SVEN_3814	hypothetical protein	5.42	3.85E-65
SVEN_5314	putative integral membrane protein	5.09	1.69E-26
SVEN_3510	Transcriptional regulator, GntR family	5.04	2.09E-70
SVEN_4124	aminoglycoside phosphotransferase	4.890	4.92E-47
SVEN_3509	D-3-phosphoglycerate dehydrogenase	4.47	1.69E-62
SVEN_0042	hypothetical protein	4.00	1.81E-22
SVEN_4562	RNA polymerase sigma-70 factor	3.99	2.22E-28
SVEN_4122	secreted protein	3.61	1.32E-46

Bibliography

1. Feklístov A, Sharon BD, Darst SA, Gross CA. 2014. Bacterial sigma factors: a historical, structural, and genomic perspective. Annu Rev Microbiol 68:357–376.

2. Helmann JD. 2019. Where to begin? Sigma factors and the selectivity of transcription initiation in bacteria. Mol Microbiol 112:335–347.

3. Sineva E, Savkina M, Ades SE. 2017. Themes and variations in gene regulation by extracytoplasmic function (ECF) sigma factors. Curr Opin Microbiol 36:128–137.

4. Staroń A, Sofia HJ, Dietrich S, Ulrich LE, Liesegang H, Mascher T. 2009. The

third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) sigma factor protein family. Mol Microbiol 74:557–581.

5. Ho TD, Ellermeier CD. 2012. Extra cytoplasmic function σ factor activation. Curr Opin Microbiol 15:182–188.

6. Mascher T. 2013. Signaling diversity and evolution of extracytoplasmic function (ECF) σ factors. Curr Opin Microbiol 16:148–155.

7. Campbell EA, Tupy JL, Gruber TM, Wang S, Sharp MM, Gross CA, Darst SA. 2003. Crystal structure of Escherichia coli sigmaE with the cytoplasmic domain of its anti-sigma RseA. Mol Cell 11:1067–1078.

8. Eberhardt RY, Chang Y, Bateman A, Murzin AG, Axelrod HL, Hwang WC, Aravind L. 2013. Filling out the structural map of the NTF2-like superfamily. BMC Bioinformatics 14:327.

9. Goutam K, Gupta AK, Gopal B. 2017. The fused SnoaL_2 domain in the Mycobacterium tuberculosis sigma factor σJ modulates promoter recognition. Nucleic Acids Res 45:9760–9772.

10. Liu Q, Pinto D, Mascher T. 2018. Characterization of the Widely Distributed Novel ECF42 Group of Extracytoplasmic Function σ Factors in Streptomyces venezuelae. J Bacteriol 200.

11. Wu H, Liu Q, Casas-Pastor D, Dürr F, Mascher T, Fritz G. 2019. The role of C-terminal extensions in controlling ECF σ factor activity in the widely conserved groups ECF41 and ECF42. Mol Microbiol 112:498–514.

12. Wecke T, Halang P, Staroń A, Dufour YS, Donohue TJ, Mascher T. 2012.

Extracytoplasmic function σ factors of the widely distributed group ECF41 contain a fused regulatory domain. Microbiologyopen 1:194–213.

13. Huang X, Pinto D, Fritz G, Mascher T. 2015. Environmental sensing in actinobacteria: a comprehensive survey on the signaling capacity of this phylum. J Bacteriol 197:2517–2535.

14. Lee J-H, Geiman DE, Bishai WR. 2008. Role of stress response sigma factor SigG in Mycobacterium tuberculosis. J Bacteriol 190:1128–1133.

15. Gaudion A, Dawson L, Davis E, Smollett K. 2013. Characterisation of the Mycobacterium tuberculosis alternative sigma factor SigG: its operon and regulon. Tuberculosis (Edinb) 93:482–491.

16. Stringer AM, Currenti S, Bonocora RP, Baranowski C, Petrone BL, Palumbo MJ, Reilly AA, Zhang Z, Erill I, Wade JT. 2014. Genome-scale analyses of Escherichia coli and Salmonella enterica AraC reveal noncanonical targets and an expanded core regulon. J Bacteriol 196:660–671.

17. Pimentel H, Bray NL, Puente S, Melsted P, Pachter L. 2017. Differential analysis of RNA-seq incorporating quantification uncertainty. Nat Methods 14:687–690.

18. Lewin GR, Carlos C, Chevrette MG, Horn HA, McDonald BR, Stankey RJ, Fox BG, Currie CR. 2016. Evolution and ecology of actinobacteria and their bioenergy applications. Annu Rev Microbiol 70:235–254.

19. Pinto D, Liu Q, Mascher T. 2019. ECF σ factors with regulatory extensions: the one-component systems of the σ universe. Mol Microbiol 112:399–409.

20. Manganelli R, Dubnau E, Tyagi S, Kramer FR, Smith I. 1999. Differential

533 expression of 10 sigma factor genes in Mycobacterium tuberculosis. Mol Microbiol
534 31:715–724.

535 21. Cappelli G, Volpe E, Grassi M, Liseo B, Colizzi V, Mariani F. 2006. Profiling of
536 Mycobacterium tuberculosis gene expression during human macrophage
537 infection: upregulation of the alternative sigma factor G, a group of transcriptional
538 regulators, and proteins with unknown function. Res Microbiol 157:445–455.

539 22. Müller AU, Imkamp F, Weber-Ban E. 2018. The Mycobacterial LexA/RecA-
540 Independent DNA Damage Response Is Controlled by PafBC and the Pup-
541 Proteasome System. Cell Rep 23:3551–3564.

542 23. Rhodius VA, Segall-Shapiro TH, Sharon BD, Ghodasara A, Orlova E, Tabakh H,
543 Burkhardt DH, Clancy K, Peterson TC, Gross CA, Voigt CA. 2013. Design of
544 orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and
545 promoters. Mol Syst Biol 9:702.

546 24. Zong Y, Zhang HM, Lyu C, Ji X, Hou J, Guo X, Ouyang Q, Lou C. 2017. Insulated
547 transcriptional elements enable precise design of genetic circuits. Nat Commun
548 8:52.

549 25. Bervoets I, Van Brempt M, Van Nerom K, Van Hove B, Maertens J, De Mey M,
550 Charlier D. 2018. A sigma factor toolbox for orthogonal gene expression in
551 Escherichia coli. Nucleic Acids Res 46:2133–2144.

552 26. Pinto D, Dürr F, Froriep F, Araújo D, Liu Q, Mascher T. 2019. Extracytoplasmic
553 Function σ Factors Can Be Implemented as Robust Heterologous Genetic
554 Switches in Bacillus subtilis. iScience 13:380–390.

555 27. Pinto D, Vecchione S, Wu H, Mauri M, Mascher T, Fritz G. 2018. Engineering
556 orthogonal synthetic timer circuits based on extracytoplasmic function σ factors.
557 Nucleic Acids Res 46:7450–7464.

558 28. Shepherd MD, Kharel MK, Bosserman MA, Rohr J. 2010. Laboratory
559 maintenance of Streptomyces species. Curr Protoc Microbiol Chapter 10:Unit
560 10E.1.

561 29. Bierman M, Logan R, O'Brien K, Seno ET, Rao RN, Schoner BE. 1992. Plasmid
562 cloning vectors for the conjugal transfer of DNA from Escherichia coli to
563 Streptomyces spp. Gene 116:43–49.

564 30. Phelan RM, Sachs D, Petkiewicz SJ, Barajas JF, Blake-Hedges JM, Thompson
565 MG, Reider Apel A, Rasor BJ, Katz L, Keasling JD. 2017. Development of Next
566 Generation Synthetic Biology Tools for Use in Streptomyces venezuelae. ACS
567 Synth Biol 6:159–166.

568 31. Lee TS, Krupa RA, Zhang F, Hajimorad M, Holtz WJ, Prasad N, Lee SK, Keasling
569 JD. 2011. BglBrick vectors and datasheets: A synthetic biology platform for gene
570 expression. J Biol Eng 5:12.

571 32. Ham TS, Dmytriv Z, Plahar H, Chen J, Hillson NJ, Keasling JD. 2012. Design,
572 implementation and practice of JBEI-ICE: an open source biological part registry
573 platform and tools. Nucleic Acids Res 40:e141.

574 33. Chen J, Densmore D, Ham TS, Keasling JD, Hillson NJ. 2012. DeviceEditor
575 visual biological CAD canvas. J Biol Eng 6:1.

576 34. Hillson NJ, Rosengarten RD, Keasling JD. 2012. j5 DNA assembly design

577 automation software. ACS Synth Biol 1:14–21.

578 35. Gibson DG, Young L, Chuang R-Y, Venter JC, Hutchison CA, Smith HO. 2009.
579 Enzymatic assembly of DNA molecules up to several hundred kilobases. Nat
580 Methods 6:343–345.

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

583 37. Bray NL, Pimentel H, Melsted P, Pachter L. 2016. Near-optimal probabilistic RNA-
584 seq quantification. Nat Biotechnol 34:525–527.

585 38. Langmead B, Salzberg SL. 2012. Fast gapped-read alignment with Bowtie 2. Nat
586 Methods 9:357–359.

587 39. Robinson JT, Thorvaldsdóttir H, Winckler W, Guttman M, Lander ES, Getz G,
588 Mesirov JP. 2011. Integrative genomics viewer. Nat Biotechnol 29:24–26.

589 40. Thompson MG, Pearson AN, Barajas JF, Cruz-Morales P, Sedaghatian N,
590 Costello Z, Garber ME, Incha MR, Valencia LE, Baidoo EEK, Martin HG,
591 Mukhopadhyay A, Keasling JD. 2020. Identification, Characterization, and
592 Application of a Highly Sensitive Lactam Biosensor from Pseudomonas putida.
593 ACS Synth Biol 9:53–62.

594 41. Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence
595 similarity searching. Nucleic Acids Res 39:W29-37.

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

598 43. Okonechnikov K, Golosova O, Fursov M, UGENE team. 2012. Unipro UGENE: a

599 unified bioinformatics toolkit. Bioinformatics 28:1166–1167.

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

602 45. Bakan A, Dutta A, Mao W, Liu Y, Chennubhotla C, Lezon TR, Bahar I. 2014. Evol
603 and ProDy for bridging protein sequence evolution and structural dynamics.
604 Bioinformatics 30:2681–2683.

605 46. Crooks GE, Hon G, Chandonia JM, Brenner SE. 2004. WebLogo: a sequence
606 logo generator. Genome Res 14:1188–1190.

607 47. Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and
608 high throughput. Nucleic Acids Res 32:1792–1797.

609 48. Waterhouse AM, Procter JB, Martin DMA, Clamp M, Barton GJ. 2009. Jalview
610 Version 2--a multiple sequence alignment editor and analysis workbench.
611 Bioinformatics 25:1189–1191.

612 49. Nguyen L-T, Schmidt HA, von Haeseler A, Minh BQ. 2015. IQ-TREE: a fast and
613 effective stochastic algorithm for estimating maximum-likelihood phylogenies. Mol
614 Biol Evol 32:268–274.

615 50. Navarro-Muñoz JC, Selem-Mojica N, Mullowney MW, Kautsar SA, Tryon JH,
616 Parkinson EI, De Los Santos ELC, Yeong M, Cruz-Morales P, Abubucker S,
617 Roeters A, Lokhorst W, Fernandez-Guerra A, Cappelini LTD, Goering AW,
618 Thomson RJ, Metcalf WW, Kelleher NL, Barona-Gomez F, Medema MH. 2020. A
619 computational framework to explore large-scale biosynthetic diversity. Nat Chem
620 Biol 16:60–68.

621 51. Potter SC, Luciani A, Eddy SR, Park Y, Lopez R, Finn RD. 2018. HMMER web
622 server: 2018 update. Nucleic Acids Res 46:W200–W204.