

# **Comprehensive phylogenetic study of ECF sigma factors**

## **Dissertation**

zur Erlangung des Grades eines  
Doktor der Naturwissenschaften  
(Dr. rer. nat.)

des Fachbereichs Biologie der Philipps-Universität Marburg

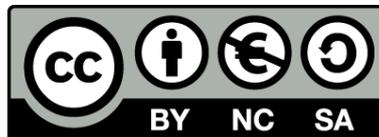
Vorgelegt von

**Delia Casas Pastor**

Aus Carbajales de Alba, Spanien

Marburg, 2020

Originaldokument gespeichert auf dem Publikationsserver der  
Philipps-Universität Marburg  
<http://archiv.ub.uni-marburg.de>



Dieses Werk bzw. Inhalt steht unter einer  
Creative Commons  
Namensnennung  
Keine kommerzielle Nutzung  
Weitergabe unter gleichen Bedingungen  
3.0 Deutschland Lizenz.

Die vollständige Lizenz finden Sie unter:  
<http://creativecommons.org/licenses/by-nc-sa/3.0/de/>

Die vorliegende Dissertation wurde von Oktober 2016 bis Januar 2020 am LOEWE-Zentrum für Synthetische Mikrobiologie (SYNMIKRO) unter Leitung von Dr. Georg Fritz angefertigt. Aus organisatorischen Gründen übernahm Frau Prof. Dr. Anke Becker ab 16.12.2019 die Betreuung.

Vom Fachbereich Biologie der Philipps-Universität Marburg (Hochschulkenziffer 1180) als  
Dissertation angenommen am \_\_\_\_\_

Erstgutachter(in): Prof. Dr. Anke Becker

Zweitgutachter(in): Dr. Simon Ringgaard

Weitere Mitglieder der Prüfungskommission:

Prof. Dr. Victor Sourjik

Prof. Dr. Lennart Randau

Tag der Disputation: \_\_\_\_\_

### **Eidesstattliche Erklärung**

Hiermit erkläre ich, dass die vorliegende Dissertation:

“Comprehensive phylogenetic study of ECF sigma factors” von mir selbstständig und ohne unerlaubte Hilfsmittel angefertigt wurde. Es wurden keine anderen als die von mir angegebenen Quellen verwendet. Zudem versichere ich, dass die Dissertation in dieser oder ähnlicher Form noch bei keiner anderen Hochschule eingereicht wurde.

.....

Delia Casas Pastor, Marburg 27 Januar 2020

## Publications

1. H. Wu\*, Q. Liu\*, **D. Casas-Pastor\***, F. Dürr\*, T. Mascher and G. Fritz. The role of C-terminal extensions in controlling ECF  $\sigma$  factor activity in the widely conserved groups ECF41 and ECF42. *Molecular Microbiology* (2019).
2. S. Chandrashekar Iyer, **D. Casas-Pastor**, D. Kraus, P. Mann, K. Schirner, T. Glatter, G. Fritz and S. Ringgaard. Transcriptional regulation by  $\sigma$  factor phosphorylation in bacteria. *Nature Microbiology*. In press.
3. **D. Casas-Pastor**, R. R. Müller, A. Becker, M. Buttner, C. Gross, T. Mascher, A. Goesmann and G. Fritz (2019). Expansion and re-classification of the extracytoplasmic function (ECF)  $\sigma$  factor family. BioRxiv preprint (2019).
4. D. Meier, **D. Casas-Pastor**, G. Fritz and A. Becker. Gene regulation by extracytoplasmic function (ECF)  $\sigma$  factors in alpha-rhizobia. *Advances in Botanical Research*. In press.

\*Equal contribution

## Table of contents

<b>Publications.....</b>	<b>4</b>
<b>Table of contents.....</b>	<b>5</b>
<b>Acknowledgments.....</b>	<b>9</b>
<b>Abstract .....</b>	<b>10</b>
<b>Zusammenfassung .....</b>	<b>11</b>
<b>List of abbreviations.....</b>	<b>13</b>
<b>1. Introduction .....</b>	<b>15</b>
1.1. DNA transcription.....	15
1.1.1. RNAP during transcription initiation across domains of life.....	15
1.1.2. Transcription elongation and termination.....	19
1.2. $\sigma^{70}$ family.....	20
1.2.1. Conserved domains in $\sigma^{70}$ family.....	22
1.3. ECF $\sigma$ factors.....	24
1.3.1. ECF-mediated responses.....	24
1.3.2. ECF diversity and classification.....	25
1.4. ECF regulators.....	25
1.4.1. Anti- $\sigma$ factors.....	26
1.4.1.1. Regulated intramembrane proteolysis.....	27
1.4.1.2. Conformational changes.....	28
1.4.1.3. Partner switching.....	29
1.4.1.4. Cell-surface signaling in FecIR-like systems.....	30
1.4.2. Hanks type kinases.....	30
1.4.3. Extensions of the ECF sequence.....	31
1.4.4. Alternative modes of regulation.....	31
1.5. Common components of signal transduction mechanisms.....	31
1.6. Bioinformatics applied to signal transduction mechanism research.....	32
<b>2. Aim and objectives.....</b>	<b>34</b>
<b>Results.....</b>	<b>35</b>
<b>3. Extracytoplasmic function <math>\sigma</math> factor (ECF) extraction and classification .....</b>	<b>36</b>
3.1. The number of identified ECFs is 50-fold larger than in the founding ECF classification ..	36
3.2. The ECF classification 2.0.....	41

3.3.	The ECF classification 2.0 refines original and identifies novel ECF groups .....	45
3.4.	ECF $\sigma$ factors feature diverse, often multi-layered, modes of regulation.....	51
3.5.	ECF hub, a public repository of the ECF classification 2.0.....	58
3.6.	Discussion and summary .....	61
<b>4.</b>	<b>Study of the binding between class I anti-<math>\sigma</math> factors and ECF <math>\sigma</math> factors.....</b>	<b>64</b>
4.1.	Class I anti- $\sigma$ domain (ASDI) retrieval and classification.....	65
4.2.	DCA predicts two main contact interfaces between ASDIs and ECFs .....	68
4.3.	SDPs confirm that two main binding surfaces determine ECF/ASDI contact.....	71
4.4.	SDPs and DCA predictions show the general binding mode of ASDIs .....	74
4.5.	The structure of ECF26/AS26 could differ from other ECF/ASDI complexes .....	77
4.6.	Testing the inactivation of anti- $\sigma$ factors by changes in membrane potential.....	80
4.7.	Discussion and summary .....	84
<b>5.</b>	<b>ECF <math>\sigma</math> factor phosphorylation.....</b>	<b>88</b>
5.1.	Members of ECF43 contain a deviant non-charged motif.....	89
5.2.	EcfP phosphorylation in T63 is required for RNA polymerase binding .....	93
5.3.	Members of ECF43 are widespread across bacteria.....	97
5.4.	ECF phosphorylation could be possible in other ECF groups .....	100
5.5.	PknT closest relatives are part of T6SS clusters in <i>Vibrio</i> spp. ....	106
5.6.	Discussion and summary .....	112
<b>6.</b>	<b>ECF regulation based on C-terminal extensions.....</b>	<b>115</b>
6.1.	C-terminal extensions have a different role in ECF41 and ECF42 .....	116
6.2.	SnoaL-like extensions differ across ECF groups .....	134
6.3.	Finding a target promoter motif for STSU_11560 from <i>Streptomyces tsukubaensis</i> .....	139
6.4.	Finding a conserved transcriptional response for members of ECF56s3.....	141
6.5.	Studies on the association of STSU_11560 to the putative anti- $\sigma$ factor STSU_11555.....	142
6.6.	Discussion and summary .....	143
<b>7.</b>	<b>Discussion and conclusion.....</b>	<b>146</b>
7.1.	Evolution of ECF $\sigma$ factors.....	146

7.2.	ECF $\sigma$ factor multiplicity.....	147
7.3.	New modes of regulation of ECFs.....	149
7.4.	Prediction of the type of regulator that targets ECFs.....	151
7.5.	Anti- $\sigma$ factor binding across the $\sigma^{70}$ family.....	151
7.6.	Vibrionales and Alteromonadales ECF43s could compose a new ECF group.....	152
7.7.	Advantages of alternative modes of regulation over anti- $\sigma$ factors.....	154
7.8.	Limitations of this study.....	155
7.8.1.	ECF retrieval pipeline.....	155
7.8.2.	ECF classification.....	157
7.8.3.	Clustering validation.....	159
7.8.4.	Challenges of the application of DCA.....	160
7.9.	Final remarks.....	162
<b>8.</b>	<b>Material and methods.....</b>	<b>163</b>
8.1.	General bioinformatic tools.....	163
8.2.	Extraction of new ECFs from NCBI.....	163
8.3.	ECF clustering.....	164
8.4.	ECF group analysis.....	166
8.5.	Classification of new ECFs against ECF clusters.....	166
8.6.	Prediction of ECF target promoter motifs.....	166
8.7.	Class I anti- $\sigma$ factor extraction.....	167
8.8.	ASDI classification.....	168
8.9.	Evaluation of ECF-ASDI co-evolution.....	168
8.10.	Specificity Determining Positions (SDPs).....	169
8.11.	Extraction of EcfP-like proteins.....	169
8.12.	Extension of STK-associated groups.....	170
8.13.	Evolution of ECF43.....	171
8.14.	Direct coupling analysis (DCA).....	172
8.15.	Promoter search for STSU_11560.....	172
8.16.	Search for a common regulon for members of ECF56s3.....	173

8.17.	Search for the putative anti- $\sigma$ factor STSU_11555 .....	173
8.18.	Bacterial strains and growth conditions.....	174
8.19.	Molecular biology techniques .....	174
8.19.1.	Construction of genetic circuits .....	174
8.19.2.	Genomic integration using CRIMoClo .....	175
8.19.3.	Tracking fluorescence emitted by GFP and optical density.....	179
<b>References</b>	.....	<b>180</b>
<b>Supplementary tables</b>	.....	<b>201</b>
<b>Curriculum vitae</b>	..... Error! Bookmark not defined.	

## Acknowledgments

I would like to thank my supervisor Dr. Georg Fritz for trusting me in so many awesome projects and for all the good scientific discussions that we had during the last 3 years, as well as for his non-stop support. This thesis would not have happened without you.

I would also like to thank the members of my thesis committee, Prof. Dr. Anke Becker, Dr. Simon Ringgaard, Prof. Dr. Victor Sourjik and Prof. Dr. Lennart Randau, who kindly accepted reviewing this thesis. I would like to especially thank Prof. Dr. Anke Becker, who accepted to be my first supervisor in Dr. Georg Fritz absence and that always provided me with useful comments and pieces of advice as my second supervisor. I would like to include in this list the rest of the people that took part in of my thesis advisory committee, Prof. Dr. Torsten Waldminghaus and Prof. Dr. Alexander Goesmann. Thank you for feedback in the project.

Several teams contributed to this work, but I would like to give a special remark to Dr. Simon Ringgaard and Dr. Shankar Chandrashekar Iyer for giving me the opportunity to collaborate with them in their amazing project. I would also like to thank Raphael Müller and Prof. Dr. Alexander Goesmann for their efforts in setting up the ECF hub platform. I appreciated critical comments and scientific discussions with Prof. Dr. Thorsten Mascher, his team and the whole ECF consortium, who also contributed to parts of this thesis. Here I would like to include Prof. Dr. Anke Becker and Dr. Doreen Meier for their feedback in the anti- $\sigma$  factor project. Lastly, I would like to thank Rute Oliveira and Dr. Marta Mendes for trusting me with the phylogenetic part of their project.

Critical reading of the thesis and scientific discussions would not have been possible without the immeasurable help of Angelika Diehl, Dr. Shankar Chandrashekar Iyer, Dr. Stefano Vecchione and Daniel Stukenberg. Thanks very much for all your comments. I would also like to thank members of the Fritz lab for the great working atmosphere during the last three years.

Although I met many great people during my PhD, I would like to especially thank the invaluable support of Adrián Izquierdo, María Esteban, Dr. Oliver Schauer, Dr. Stefano Vecchione and Dr. Andre Sim. I was really lucky to meet you all.

Finalmente, siempre estaré agradecida a mi familia. Mis padres, que con su incansable trabajo han conseguido guiarme por el buen camino y siempre serán mi ejemplo a seguir. Mi hermano, siempre con sus buenos consejos y su capacidad para relativizar los problemas. Esta tesis os la dedico a vosotros. Muchas gracias de corazón.

## Abstract

Extracytoplasmic function (ECF)  $\sigma$  factors are the most minimalistic member of the  $\sigma^{70}$  family. ECFs and their activity regulators are one of the main signal transduction mechanisms that allow bacteria to respond to extracellular changes. Aside from their natural role in bacterial homeostasis, ECFs are generally host independent and functionally orthogonal, which makes them especially attractive for constructing bacterial synthetic circuits. *In silico* identification of sets of ECFs, their target promoters and their regulators is particularly simple since ECFs and their regulators are typically encoded in the same genetic neighborhood and usually in the same operon, and ECFs usually target their own promoter. Earlier works on the phylogenetic classification of ECFs revealed that there is a correlation between ECF groups, which harbor proteins with a similar sequence, regulator type and target promoter motif elements. This showed that the phylogenetic classification of ECFs is essential to understand their modes of regulation. The large number of sequenced bacterial genomes currently deposited in databases suggests that an ECF reclassification would expand our knowledge on ECF regulation. This thesis addresses the analysis of the main modes of regulation found in the comprehensive classification of ECF  $\sigma$  factor subfamily.

For this study, I first extracted ECFs from all bacterial genomes deposited in NCBI. I identified more than 170,000 unique protein sequences that are likely to function as ECFs. This resulted in a 50-fold expansion over the original ECF library. Then, I classified the conserved  $\sigma$  domains of these proteins into more than 150 phylogenetic groups, each associated to a conserved type of regulator. I systematically described each ECF group in terms of its putative regulator, putative target promoter, taxonomic distribution and putative function. I confirmed these predictions for groups with described members. Anti- $\sigma$  factors are the main type of ECF regulator across groups, followed by C-terminal extensions of their protein and serine/threonine kinases, which have been suggested to phosphorylate ECFs. I hypothesized new alternative types of regulators for some ECF groups.

Using a combination of bioinformatic tools and collaborating with different experimental research groups, I focused on the most important regulatory elements of ECFs to shed light into their mechanism of regulation. In the case of anti- $\sigma$  factors, I focused on their most common type, class I anti- $\sigma$  factors, to reveal two shared binding interfaces between ECFs and these inhibitors. Then, I focused on the three largest ECF groups associated to C-terminal extensions, showing a different role of this additional region in the control of ECF activity in the different groups. Lastly, I focused on serine/threonine kinases to find that phosphorylation compensates for the lack of negative charges in one of the main RNA polymerase binding surfaces of ECF  $\sigma$  factors.

In summary, this thesis provides the scientific community with a comprehensive overview of ECF  $\sigma$  factor regulation, target promoter and function across phylogenetic groups, and sheds light into some of their most important regulatory mechanisms.

## Zusammenfassung

Extracytoplasmatisch wirkende (ECF)  $\sigma$ -Faktoren sind die minimalistischsten Mitglieder der  $\sigma 70$  Familie. ECFs und ihre Aktivitätsregulatoren sind eine der wichtigsten Signaltransduktionsmechanismen, die es Bakterien ermöglichen auf extrazelluläre Änderungen zu reagieren. Neben der bakteriellen Homöostase, ECFs sind im Allgemeinen wirtsunabhängig und funktionell orthogonal, wodurch sie besonders attraktiv für die Konstruktion von bakteriellen synthetischen Schaltkreisen sind. In silico Identifizierung von Paaren aus ECFs, ihren Zielpromotoren und ihren Regulatoren ist insbesondere deshalb einfach, weil ECFs und ihre Regulatoren typischerweise in derselben genetischen Nachbarschaft und üblicherweise im selben Operon kodiert sind. Außerdem erkennen ECFs üblicherweise ihre eigenen Promotoren. Frühere Arbeiten haben gezeigt, dass eine Korrelation zwischen der ECF Proteinsequenz, also der phylogenetischen Gruppe, und dem Regulationstyp, sowie dem Zielpromoter besteht, was zeigt, dass die phylogenetische Klassifizierung von ECFs essentiell für das Verständnis der Regulationsmodi ist. Die hohe Anzahl an sequenzierten bakteriellen Genomen, die aktuell in Datenbanken hinterlegt ist, legt nahe, dass eine ECF-Reklassifizierung unser Verständnis über ECF-Regulation erweitern würde.

Diese Arbeit befasst sich mit der Analyse der wichtigsten Regulationsmechanismen, die in einer umfassenden Klassifizierung der ECF- $\sigma$ -Faktor Unterfamilie gefunden wurden.

Für dieses Projekt habe ich zuerst die ECFs aus allen bakteriellen Genomen extrahiert, die in NCBI hinterlegt sind. Ich habe mehr als 170000 einzigartige Proteinsequenzen identifiziert, die wahrscheinlich als ECFs wirken. Dies resultiert in einer 50-fachen Erweiterung im Vergleich zu der vorherigen ECF-Sammlung. Anschließend habe ich die konservierten  $\sigma$ -Domänen dieser Proteine in mehr als 150 phylogenetische Gruppen gruppiert, von der jede mit einem konservierten Regulator assoziiert ist. Ich habe die ECF-Gruppen in Bezug auf ihre möglichen Regulatoren und Zielpromotoren, taxonomische Verbreitung, sowie ihrer mögliche Funktion beschrieben. In den meisten Gruppen, Anti- $\sigma$ -Faktoren sind die wichtigsten Regulatoren, gefolgt von C-terminalen Verlängerungen des Proteins und Serin/Threonin-Kinasen, die möglicherweise ECFs phosphorylieren. Ich habe Hypothesen über neue, alternative Regulationsmechanismen für einige ECF-Gruppen aufgestellt.

Durch die Nutzung von bioinformatischen Methoden und in Zusammenarbeit mit experimentell arbeitenden Gruppen, habe ich mich auf die wichtigsten regulatorischen Elemente von ECFs fokussiert, um ihre Regulationsmechanismen aufzudecken. Im Fall der Anti- $\sigma$ -Faktoren habe ich mich auf den häufigsten Typ, Klasse I Anti- $\sigma$ -Faktoren, fokussiert um zwei gemeinsame Bindungsoberflächen zwischen ECF und Anti- $\sigma$ -Faktoren zu entdecken. Anschließend habe ich mich mit den drei häufigsten ECF Gruppen befasst, die mit C-terminalen Verlängerungen assoziiert sind, wodurch ich eine verschiedenartige Rolle dieser Sequenz in der Kontrolle der ECF-Aktivität in den verschiedenen Gruppen zeigen konnte. Abschließend habe ich mich auf Serin/Threonin-Kinasen

konzentriert, um zu entdecken, dass Phosphorylierungen für den Mangel an negativen Ladungen in einer der wichtigsten Bindeoberflächen zwischen RNA-Polymerase und ECF- $\sigma$ -Faktoren kompensiert.

Zusammenfassend bietet diese Arbeit der wissenschaftlichen Gemeinschaft einen umfassenden Überblick über phylogenetische Gruppen von ECF- $\sigma$ -Faktoren ihrer Ziel promotoren und Funktionen, sowie wichtige Einblicke in einige ihrer wichtigsten Regulationsmechanismen.

**List of abbreviations**

1CS – One-component system  
2CS – Two-component system  
Å – Angstrom  
aa – Amino acids  
AS – Anti- $\sigma$   
ASD – Anti- $\sigma$  domain  
ATc – Anhydrotetracycline  
AUC – Area under the curve  
BLAST – Basic local alignment search tool  
bp - Base pairs  
BRE – B-recognition elements  
BTH – Bacteria two hybrids  
BY-kinase – Bacterial tyrosine kinases  
CAP – cAMP receptor protein  
CFU – Colony forming units  
ChIP-seq – Chromatin immunoprecipitation sequencing  
CK2 –  $\alpha$ -subunit of nucleocytosolic casein kinase 2  
CSK – Chloroplast sensor kinase  
D+E – Aspartate and glutamate  
DCA – Direct coupling analysis  
ECF – Extracytoplasmic function  $\sigma$  factor  
FHA – Forkhead-associated domain  
GFP – Green fluorescent protein  
GO – Gene Ontology  
HGT – Horizontal gene transfer  
HMM – Hidden Markov model  
HTH – Helix turn helix  
IPTG –  $\beta$ -D-1-thiogalactopyranoside  
Kdo – 3-deoxy-D-manno-octulosonic acid  
LB – Lysogeny broth  
LC-MS – Liquid chromatography-mass spectrometry  
MoClo – Modular cloning  
MS/MS – Tandem mass spectrometry  
MSA – Multiple-sequence alignment  
NCBI – Nacional center for biotechnology information

NCR – Non-conserved region  
NOG – Non-supervised orthologous group  
NTP – Nucleotide tri-phosphate  
OD<sub>600</sub> – Optical density at 600 nm  
OMP – Outer membrane protein  
ORF – Open reading frame  
PCC – Pearson correlation coefficient  
PCR – Polymerase chain reaction  
PIC – Pre-initiation complex  
PMBN – Polymyxin B nonapeptide  
PQ – Plastoquinone  
PS – Photosystem  
PTK – Plastid transcription kinase  
PUL – Polysaccharide utilization locus  
RIP – Regulated intramembrane proteolysis  
RNA-seq – RNA sequencing  
RNAP – RNA polymerase  
ROC – Receiver operating characteristic  
S+T – Serine and threonine  
SDP – Specificity-determining position  
SEM – Standard error of the mean  
sfGFP – Superfolder green fluorescent protein  
STK - Serine/threonine kinase  
Sus – Starch utilization system  
T3SS – Type three secretion system  
T6SS – Type six secretion system  
TBP – TATA box-binding protein  
TEC – Ternary elongation complex  
TM – Transmembrane  
TPR – Tetratricopeptide repeat  
TSS – Transcription start site  
UPP – Ultra-large alignments using phylogeny-aware Profiles  
X-Gal – 5-Bromo-4-chloro-3-indolyl- $\beta$ -D-galactoside

## 1. Introduction

Bacterial homeostasis is achieved through signal transduction mechanisms that connect the environment with the cytoplasm. Extracytoplasmic function  $\sigma$  factors (ECFs) are the core component of one of the main signal transduction mechanisms in bacteria in terms of their abundance and importance of the responses that they mediate (Staroń *et al.*, 2009). As members of the  $\sigma^{70}$  family, ECFs guide the RNA polymerase (RNAP) to specific promoter sequences, and thereby enable bacteria to redirect gene expression in response to deteriorating environmental conditions (Helmann, 2002; Paget and Helmann, 2003). Although ECFs-based systems are generally less prevalent than one-component systems (1CSs) and two-component systems (2CSs), previous studies revealed a large ECF abundance, with an average of seven ECFs per bacterial genome (Staroń *et al.*, 2009); a large diversity, with more than 90 phylogenetic groups (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b; Pinto and Mascher, 2016); and a diverse range of activation mechanisms (Mascher, 2013). In this thesis I study the diversity of ECF  $\sigma$  factor regulation.

### 1.1. DNA transcription

Genetic information is stored in DNA, but needs to be converted into proteins or non-coding RNA to play a role in the metabolic network of the cell. For this purpose, genetic information stored in DNA is transcribed into RNA and translated into proteins (Crick, 1958). DNA is organized in genes, discrete genomic transcribable regions that contain information for the synthesis of functional proteins or non-coding RNAs (Pesole, 2008).

DNA transcription is the initial step of gene expression and is mediated by DNA-dependent RNA polymerases (RNAP). RNAP reads DNA and synthesizes its complementary RNA. This enzyme was discovered independently in several organisms in the 60s and 70s, including *Escherichia coli* (Hurwitz *et al.*, 1961), rat liver (Weiss and Gladstone, 1959), *Micrococcus lysodeikticus* (Weiss and Nakamoto, 1961) and bacteriophage T7 (Chamberlin, Mcgrath and Waskell, 1970). Transcription requires three main steps: initiation, elongation and termination. During initiation, RNAP binds the DNA, and RNA synthesis starts. Then, the RNAP continues transcribing while it moves from 5' to 3' along the gene (elongation) (Holstege, Fiedler and Timmers, 1997; Pal, Ponticelli and Luse, 2005). Transcription stops at the end of the gene (termination) (reviewed in (Orphanides and Reinberg, 2002)). This basic process is the subject of modifications by regulatory transcription factors and varies depending of the domain of life. For instance, chromatin remodeling and histone modification exert an important regulatory role in gene transcription in eukaryotes (reviewed in (Orphanides and Reinberg, 2002)).

#### 1.1.1. RNAP during transcription initiation across domains of life

RNAP is an evolutionary-related multisubunit complex (Werner and Grohmann, 2011), although a phage-related, single-subunit RNAP exists in plant chloroplasts and mitochondria (McAllister, 1993;

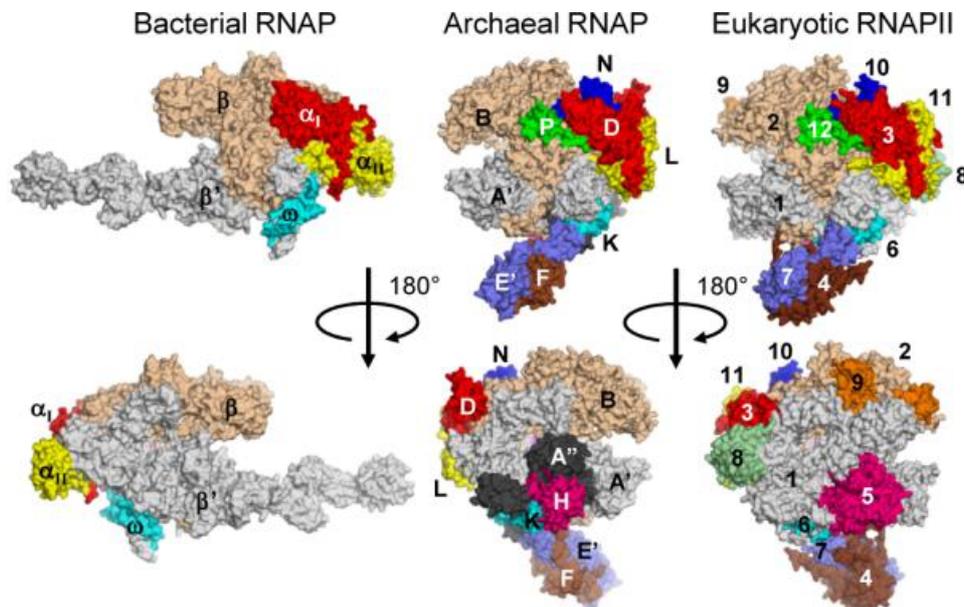
Cermakian *et al.*, 1997; Forrest *et al.*, 2017). Bacterial multi-subunit RNAP is composed of five core subunits ( $\beta\beta'\alpha_2\omega$ ) and a dissociable subunit ( $\sigma$ ). The largest subunits,  $\beta$  and  $\beta'$ , form the catalytic center where RNA is synthesized from NTPs using ssDNA as template (reviewed in (Murakami, 2015)). The cleft between both subunits is occupied by DNA, RNA and the DNA-RNA hybrid. Even though  $\beta$  and  $\beta'$  subunits are conserved, insertions in these subunits describe certain bacteria lineages (Lane and Darst, 2010a). Moreover, cyanobacterial  $\beta'$  is split into two peptides,  $\beta'$  and  $\gamma$  subunits (Schneider and Hasekorn, 1988). Core RNAP needs the transient binding of the  $\sigma$  subunit to recognize suitable promoter sequences.  $\sigma$  binds to the core RNAP, forming the RNAP holoenzyme, and drives it to promoter sequences, where it binds to -35 and -10 promoter elements reviewed in (Murakami, 2015) (Section 1.2).

Eukaryotic and archaeal RNAPs are more similar between themselves than to bacterial RNAP (Fig. 1.1). One of the most noticeable differences between archaeal/eukaryotic RNAP and bacterial RNAP is the presence of a stalk in the former (Fig. 1.1). This structure, composed of two proteins (Rpb4/RpoF and Rpb7/RpoE'), protrudes in the periphery of the RNAP clamp structure and plays multiple roles, including transition to the open complex (Hirtreiter, Grohmann and Werner, 2010), promotion of processivity and a more efficient termination (Naji, Grünberg and Thomm, 2007) (reviewed in (Werner and Grohmann, 2011)).

Eukaryotes contain three types of RNAPs for the transcription of different sets of genes: 1) RNAP I (Pol I), in charge of the transcription of rRNA; 2) RNAP II (Pol II) transcribes the mRNA, miRNA, snRNA and SnoRNA; and 3) RNAP III (Pol III), which transcribes tRNA and 5S rRNA (Roeder and Rutter, 1969; Carter and Drouin, 2009). On top of the largest two subunits (Rpb1 and Rpb2), homologs of eight extra subunits are present in all the three types of eukaryotic (reviewed in (Cramer *et al.*, 2008)). In contrast, DNA transcription in Archaea is carried out by a single RNAP, very similar in structure and domain composition to eukaryotic Pol II (Hirata, Klein and Murakami, 2008) (Fig. 1.1).

Eukaryotic core RNAPs are composed of 5 to 17 subunits, of which four are homologs of bacterial  $\beta$ ,  $\beta'$ ,  $\alpha$  and  $\omega$  subunits. These four subunits are common to all domains of life (Cramer, 2002; Hirata, Klein and Murakami, 2008). This core RNAP establishes contacts with several transcription factors with different degrees of stability, making difficult to give a definition to core RNAP subunits in eukaryotes (reviewed in (Werner and Grohmann, 2011)). Similarly to bacterial  $\beta$  and  $\beta'$  subunits, the largest and second largest subunits of the eukaryotic and archaeal RNAPs (Rpb1 and Rpb2) compose the RNAP active cleft (Cramer *et al.*, 2000). All multisubunit RNAPs share a conserved catalytic core, including two double-psi  $\beta$  barrels, one in  $\beta$ /Rsb2 and another in  $\beta'$ /Rsb1 (Lane and Darst, 2010b). These two double-psi  $\beta$  barrels are also present in the catalytic center of RNA-dependent RNA polymerases involved in RNAi synthesis, but in this case both are located in the same peptide (Salgado *et al.*, 2006). Other conserved regions shared between RNAPs and RNA-dependent RNA

polymerases are the trigger loop helices and the bridge helix from  $\beta'$  (Salgado *et al.*, 2006; Lane and Darst, 2010b).



**Figure 1.1. Multisubunit RNA polymerase** across the three domains of life. Source: (Hirata, Klein and Murakami, 2008), under license number 4702561480935 from Springer Nature.

Pol II is the most studied RNAP and is composed of 12 subunits (Cramer *et al.*, 2008). Pol II cooperates with the so-called “general transcription factors” (TFIIB, TFIID, TFIIIE, TFIIF and TFIIH) to bind to promoter DNA in the pre-initiation complex (PIC), open its double strand and stimulate promoter escape in the transition to elongation phase (reviewed in (Sainsbury, Bernecky and Cramer, 2015)). The most important general transcription factors required for transcription are TFIID and TFIIB (reviewed in (Sainsbury, Bernecky and Cramer, 2015)). TFIIB has structural and functional homology to the dissociable  $\sigma$  factor subunit of the bacterial RNAP (Liu *et al.*, 2010) and is essential for transcription in Archaea (reviewed in (Werner and Grohmann, 2011)), indicating its importance across domains of life. Nevertheless, a common origin of TFIIB and  $\sigma$  factors is controversial (Werner and Grohmann, 2011). During canonical PIC assembly, TFIIB makes sequence-specific contacts with the B recognition elements or BREs, DNA elements upstream and downstream from the TATA box. The TATA box is an element present in the promoter ~20% of the Pol II-regulated yeast genes and 10%-20% human genes with consensus TATWAWR (Basehoar, Zanton and Pugh, 2004; Gershenzon and Ioshikhes, 2005; Cooper *et al.*, 2006); however, the amount genes with TATA box might be larger than previously thought (Rhee and Pugh, 2012). TATA boxes are located ~30bp upstream of the transcription start site (TSS) in humans (Sainsbury, Bernecky and Cramer, 2015). TFIID is a protein complex composed the TATA box-binding protein (TBP), essential for transcription, and 13-14 TBP-associated factors that control promoter specificity (reviewed in (Sainsbury, Bernecky and Cramer, 2015)). TBP binds to the minor groove of the TATA box and bends it 90 degrees (Kim *et al.*, 1993). TFIIB and TBP bind to the promoter, together with the auxiliary factor TFIIA (Geiger *et al.*, 1996), and recruit the Pol II-TFIIF complex (Sainsbury, Niesser

and Cramer, 2013). In this complex, TFIIF prevents unspecific binding of Pol II to DNA (Conaway *et al.*, 1991) and stabilizes TFIIB within the PIC (Čabart *et al.*, 2011). Then, TFIIE and TFIIH join this core initiation complex to form the complete PIC in its closed conformation. DNA in this complex is melted in the presence of NTPs, forming the open complex (reviewed in (Sainsbury, Bernecky and Cramer, 2015)). DNA melting and RNA synthesis initiation is facilitated by TFIIB.

Bacteria, as well as Archaea, contain a single RNAP that carries out all the transcription. Bacterial RNAP is composed of a 5-subunit core ( $\beta'\beta\alpha_2\omega$ ) and a dissociable  $\sigma$  subunit, involved in DNA binding and melting. The five core subunits have homologs in the eukaryotic and archaeal systems –  $\beta'$  is a homolog of Rsb1 (Jokerst *et al.*, 1989),  $\beta$  is a homolog of Rsb2 (Sweetser, Nonet and Young, 1987), the  $\alpha$  subunits are homologs of Rpb3 and Rpb11 (Zhang and Darst, 1998) and the  $\omega$  subunit is a homolog of Rpb6 (Minakhin *et al.*, 2001). The dissociable  $\sigma$  subunit of the bacterial RNAP, first discovered in 1969 (Burgess *et al.*, 1969), has functional and structural homology to eukaryotic/archaeal TFIIB (Liu *et al.*, 2010). TFIIB contains two core regions, the B-ribbon and the B-core, connected by a flexible linker. The B-ribbon domain, in TFIIB N-terminus, binds to the Pol II dock domain, whereas its B-core domain, in TFIIB C-terminus, binds to TATA-TBP complex, BREs and Pol II clamp helices, conserved across the three domains of life (Kostrewa *et al.*, 2009; Werner and Grohmann, 2011). This arrangement is reminiscent of  $\sigma_2$  and  $\sigma_4$  domains of bacterial type IV  $\sigma$  factors, where  $\sigma_2$  domain, in  $\sigma$  N-terminus, also binds to the clamp helices of the RNAP and  $\sigma_4$  domain binds to the flap tip helix of the RNAP and to the -35 element of the promoter (Campbell, Muzzin, *et al.*, 2002; Liu *et al.*, 2010). Similarly to  $\sigma$  factor 3.2 domain, TFIIB flexible linker plays a role in the DNA opening (B-linker region) and positioning DNA in the active cleft so as to start transcription (B-reader region) (Kostrewa *et al.*, 2009; Sainsbury, Niesser and Cramer, 2013). TFIIB and  $\sigma_{3.2}$  run along the RNA exit channel and the active cleft (Campbell, Muzzin, *et al.*, 2002; Kostrewa *et al.*, 2009), hampering transcription. Consequently, TFIIB clashes with the nascent RNA after 7bp have been synthesized and needs to be released during promoter escape (Tran and Gralla, 2008). In Pol II, the release of TFIIB is delayed to 12-13bp RNA by TFIIF stabilization (Čabart *et al.*, 2011). Bacterial  $\sigma$  factor is released from the core RNAP after ~9bp of RNA have been synthesized (Hansen and McClure, 1980; Metzger *et al.*, 1993). However, data on two-component  $\sigma$  factors revealed that  $\sigma_2$  domain can retain its interaction with RNAP through elongation phase and be released in a stochastic manner (Sengupta, Prajapati and Mukhopadhyay, 2015).

After the assembly of the RNAP in the promoter region, the RNAP enters in a phase of abortive transcription, the abortive cycling, during which short RNA transcripts (3-9bp) are released without the RNAP complex disengaging from the promoter (reviewed in (Werner and Grohmann, 2011)(Goldman, Ebright and Nickels, 2009). This cycle is probably caused by the clash between the nascent RNA chain and the  $\sigma_{3.2}$ /TFIIB occupying RNA exit channel and the need to release  $\sigma$ /TFIIB from its binding to the RNAP and promoter to continue transcription (Werner and Grohmann, 2011; Luse, 2013). Pol II has a smaller tendency to abortive cycling due to the assistance of TFIIH (Dvir,

Conaway and Conaway, 1997). Once that RNA enters in the exit channel (9-10bp), the downstream transcription bubble size increases together with RNA. After the transcription bubble reaches the critical size of 17-18bp, its upstream part collapses leaving a bubble of ~10bp, typical of the mature elongation complex (Holstege, Fiedler and Timmers, 1997; Pal, Ponticelli and Luse, 2005). At this point, TFIIF is not required any longer in Eukaryotic RNAPs (Pal, Ponticelli and Luse, 2005) and other transcription initiation factors, including TFIIB, H and TBP are released (reviewed in (Luse, 2013)).

As RNAP transitions into mature elongation, transcription progresses in 5' to 3' direction and nascent RNA fills the RNA exit channel. During this transition, transcript slippage is common when Pol II complexes with 15-21bp RNA read repetitive DNA (Pal and Luse, 2003). Moreover, complexes paused at 17-32bp have tendency to backtrack or arrest depending on the template sequence (Pal, McKean and Luse, 2001; Luse, 2013). After 30bp of RNAP have been synthesized, the RNAP complex exhibits all the characteristics of an elongation complex, it has less tendency to slip and backtrack, and no interaction with initiation factors is left (reviewed in (Luse, 2013)).

#### 1.1.2. Transcription elongation and termination

In productive transcription, RNAP complex escapes from the promoter (promoter escape) and moves in 5' to 3' direction along the gene. This ternary elongation complex (TEC) is composed of RNAP, DNA and RNA (reviewed in (Werner and Grohmann, 2011)). Elongation is a discontinuous process modulated by DNA and RNA sequences, as well as elongation factors and gene-specific transcription factors (reviewed in (Werner and Grohmann, 2011)). Transcription elongation requires the bridge and trigger helices of the active cleft (Brueckner, Ortiz and Cramer, 2009). The archaeal/eukaryotic-specific RNAP stalk interacts with the nascent RNA and repositions RNAP clamp helices (Armache *et al.*, 2005; Werner and Grohmann, 2011). This stalk is required for full RNAP processivity (defined as the number of nucleotides polymerized per initiation event (Werner and Grohmann, 2011)) (Hirtreiter, Grohmann and Werner, 2010) and mRNA 3'-end processing (Runner, Podolny and Buratowski, 2008). Bacterial RNAP lacks of stalk and uses the flap domain of  $\beta$  subunit to bind to the nascent transcript (reviewed in (Werner and Grohmann, 2011)).

During promoter escape and elongation, transcription pausing is a common event. These pauses are the basis of several transcription regulatory mechanisms, including attenuation (Kingston and Chamberlin, 1981), antitermination and termination (reviewed in (Ait-Bara *et al.*, 2017)), promoter-proximal pausing (reviewed in (Jonkers and Lis, 2015)), slippage, transcription arrest, coupling of transcription and translation in Bacteria or mRNA splicing in Eukaryotes (reviewed in (Zhang and Landick, 2016) and (Saba *et al.*, 2019)). Transcription pause occurs due to a several reasons, including sequence specific contacts between RNAP and DNA or RNA, and RNA secondary structures (Saba *et al.*, 2019). The initial paused elongation complex has the tendency to rearrange into a long-lived paused elongation complex through different mechanisms, of which the most of

important is backtracking, or reverse translocation of DNA and RNA (Saba *et al.*, 2019). Transcription cannot progress from backtracking complexes and RNAPs need to cleave the transcript internally, releasing the blockage and realigning the 3'-OH of the nascent RNA in the active site (Deighan and Hochschild, 2006; Sigurdsson, Dirac-Svejstrup and Svejstrup, 2010; Werner and Grohmann, 2011). The cleavage is carried out by TFIIIS in Eukaryotes, TFS in Archaea and GreB in Bacteria (reviewed in (Werner and Grohmann, 2011)). An important enhancer of elongation and modulator of termination is NusG, which is the only universally conserved transcription factor across all the domains of life (SPT5/Spt5/NusG in Eukaryotes, Archaea and Bacteria, respectively), pinpointing elongation as an essential phase across evolution (Werner and Grohmann, 2011; Lawson *et al.*, 2018).

Transcription termination occurs when the elongation complex becomes unstable and dissociates from DNA and RNA. This destabilization can be caused by DNA sequence (intrinsic termination) or by the binding of a termination factor (factor-dependent termination) (reviewed in (Peters, Vangeloff and Landick, 2011)). Intrinsic termination is based on a weaker RNA-DNA hybrid. This type of termination is less common in eukaryotic Pol I and II, where the stability of the RNA-DNA hybrid usually determines transcription pausing and arrest (reviewed in (Peters, Vangeloff and Landick, 2011; Gehring, Walker and Santangelo, 2016)). In archaea and Pol III intrinsic termination is based on a poly-adenine or trails in the templated DNA (reviewed in (Gehring, Walker and Santangelo, 2016)). In bacteria, intrinsic terminators contain a G-C rich stem loop structure followed by a U-tract (Carafa, Brody and Thermes, 1990).

Factor-dependent termination is generally mediated by Rho in Bacteria. Rho is a homo-hexamer helicase with RNA-dependent ATPase activity that binds to *rut* sequences (Rho utilization) of the RNA in the elongation complex. After binding, Rho moves along the RNA towards the TEC and triggers the release of the RNA from the RNAP and DNA (reviewed in (Peters, Vangeloff and Landick, 2011)). Aside from Rho, Mfd dissociates transcription elongation complexes blocked by DNA lesions in *E. coli* (Park, Marr and Roberts, 2002) and the RNase J1 degrades RNA from stalled TEC and induces the disassembly of RNAP from the DNA in *Bacillus subtilis* (Šiková *et al.*, 2019). Eukaryotic Pol I terminates transcription with the binding of a polymerase-specific factor to the DNA downstream from the elongation complex (Uzman *et al.*, 2000), whereas Pol II termination is triggered by the recognition of the polyadenylation site and it is coupled with the cleavage and polyadenylation of the 3' end of the RNA (Fong *et al.*, 2015).

### 1.2. $\sigma^{70}$ family

Bacterial RNAP  $\sigma$  factors are divided into two families,  $\sigma^{54}$  and  $\sigma^{70}$ , with no sequence similarity (Merrick, 1993). Members of  $\sigma^{54}$  require an enhancer-binding protein of the AAA+ family to hydrolyze ATP and induce the formation of the open complex (Rombel *et al.*, 1998; Studholme and Buck, 2000; Werner and Grohmann, 2011). Instead, members of  $\sigma^{70}$  family can start transcription

without any other protein.  $\sigma^{70}$ s are modular proteins composed by up to four conserved  $\sigma$  domains (named  $\sigma_1$  to  $\sigma_4$ ) with a distinct function. Of those, only  $\sigma_2$  and  $\sigma_4$  domains are essential for  $\sigma$  function. Even though a single member of  $\sigma^{54}$  is contained in each bacterial genome, several members of  $\sigma^{70}$  with distinct features can be found per genome (reviewed in (Gruber and Gross, 2003)).

As a consequence of their abundance and diversity,  $\sigma^{70}$ s are subdivided into groups according to phylogenetic relatedness (Lonetto, Gribskov and Gross, 1992). Group 1 contains primary or housekeeping  $\sigma$  factors, which harbor the four  $\sigma$  domains and in some cases an extension of their sequence of variable length, called non-conserved region (NCR) (Lonetto, Gribskov and Gross, 1992; Leibman and Hochschild, 2007). Housekeeping  $\sigma$  factors, such as RpoD in *E. coli*, are usually unique in the genome and are essential for cell viability (Lonetto, Gribskov and Gross, 1992). The remaining  $\sigma^{70}$  groups, or alternative  $\sigma$  factors, are truncated respect to the housekeeping  $\sigma$  factors. Often, alternative  $\sigma$  factors are not essential and play a role in stress resistance, such as survival in stationary phase, heat shock, and antimicrobial molecules (reviewed in (Helmann, 2002)).

There are three main classes of alternative  $\sigma$  factors depending on their domain structure. Group 2  $\sigma$  factors are non-essential  $\sigma$  factors that lack the first part of the  $\sigma_1$  domain ( $\sigma_{1.1}$  region), whereas group 3  $\sigma$  factors lack  $\sigma_1$  region and group 4  $\sigma$  factors lack both  $\sigma_1$  and  $\sigma_3$  domains (Lonetto, Gribskov and Gross, 1992; Lonetto *et al.*, 1994; Helmann, 2002). Members of group 2, represented by RpoS in *E. coli*, are usually general stress and stationary phase  $\sigma$  factors (reviewed in (Paget, 2015)), but they are responsible of oxidative stress acclimation in Cyanobacteria (Hakkila *et al.*, 2019). In Bacteroidetes,  $\sigma^{\text{ABfr}}$ -like group 2  $\sigma$  factors function as essential primary  $\sigma$  factors instead of group 1  $\sigma$  factors (Vingadassalom *et al.*, 2005). The same type of housekeeping group 2  $\sigma$  factors may appear in *Chlorobium* spp. (Iyer, Koonin and Aravind, 2004). Members of group 3 have more specialized functions related to flagellum synthesis (FliA from *E. coli*), sporulation ( $\sigma^{\text{WhiG}}$  from *Streptomyces* and  $\sigma^{\text{F}}$ ,  $\sigma^{\text{E}}$ ,  $\sigma^{\text{G}}$  and  $\sigma^{\text{K}}$  from *B. subtilis*) and heat shock response and general stress response in Gram-positives ( $\sigma^{\text{B}}$  from *B. subtilis*) (reviewed in (Paget, 2015)).

Members of group 4, also called extracytoplasmic function  $\sigma$  factors (ECFs), are the most diverse and abundant  $\sigma$  factors and are part of one the most important signal transduction mechanism in bacteria (Helmann, 2002; Staroń *et al.*, 2009). ECFs, first identified by Lonetto and colleagues (Lonetto *et al.*, 1994), usually play a role in the defense against periplasmic and extracellular stresses, including antimicrobial resistance, but also have a role in metal resistance, iron acquisition, biofilm formation, carbohydrate degradation and oxidative stress resistance, among others (reviewed in (Paget, 2015)).

Helmann defined an extra  $\sigma^{70}$  group, group 5, composed of proteins related to TxeR from *Clostridium difficile* that are generally involved in the transcription of toxin and bacteriocin genes (Helmann, 2002). Members of group 5 have little sequence similarity to other members of  $\sigma^{70}$  (Helmann, 2002). Lastly, a new  $\sigma^{70}$  group, represented by  $\sigma^{\text{I}}$  from *Clostridium thermocellum*, contains  $\sigma_2$  domain and a divergent domain in C-terminus ( $\sigma_{\text{I-C}}$ ) that is thought to have a similar function as conserved  $\sigma_4$

domain (Wei *et al.*, 2019).  $\sigma^1$ s regulate the expression of cellulosomes, extracellular multi-enzyme complexes in charge of cellulose degradation (Nataf *et al.*, 2010).

### 1.2.1. Conserved domains in $\sigma^{70}$ family

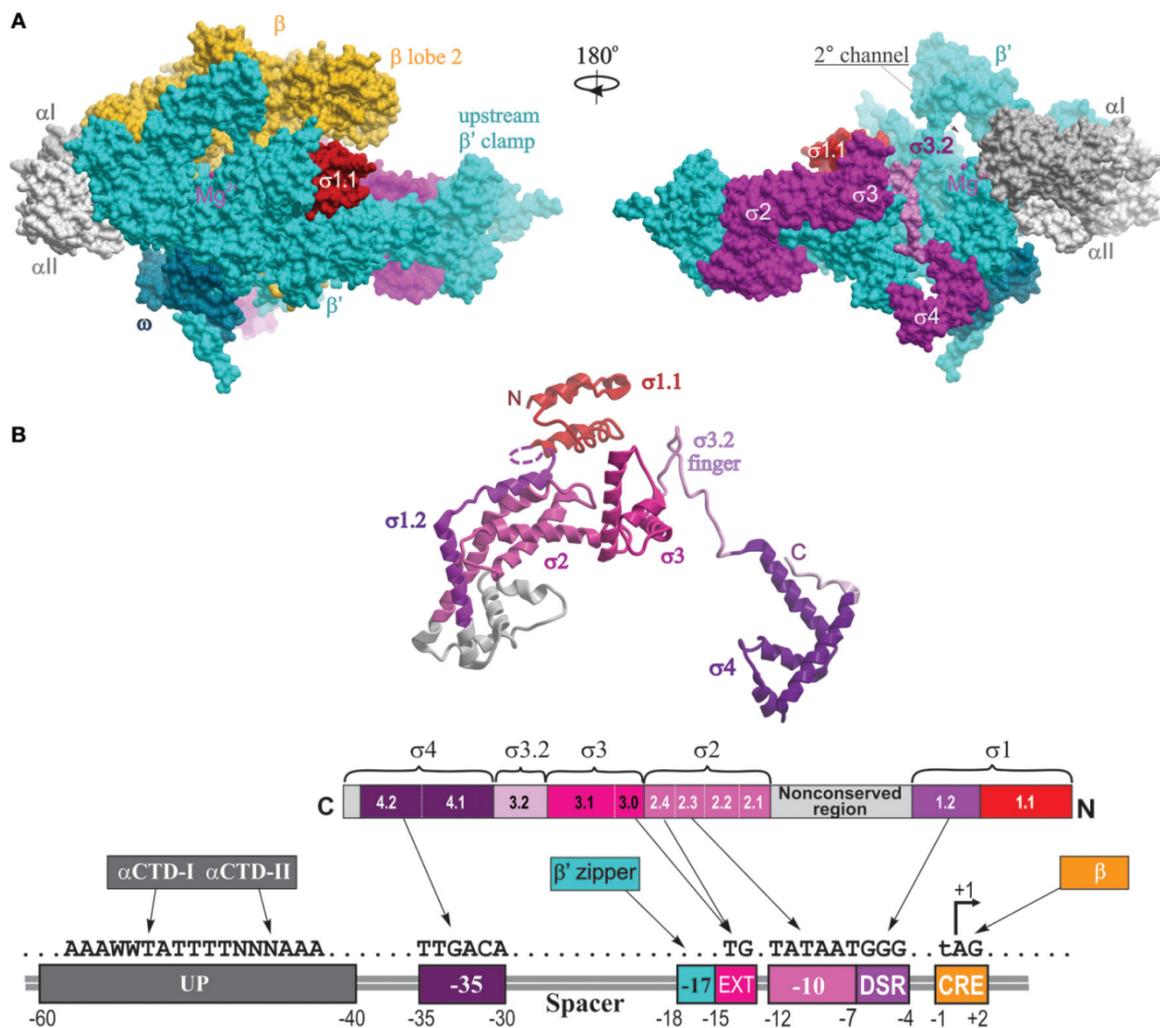
The modular structure of members of the  $\sigma^{70}$  family makes it possible to assign a function to each domain and even to regions within domains. It is important to notice that, aside from the core  $\sigma$  domains,  $\sigma$  factors may also contain additional domains that are not related to the core  $\sigma$  function but to their regulation (Pinto, Liu and Mascher, 2019). Next paragraphs will describe the most common  $\sigma^{70}$  domains.

Domain  $\sigma_1$  is subdivided into regions  $\sigma_{1.1}$  and  $\sigma_{1.2}$  (Fig 1.2). Of these,  $\sigma_{1.1}$ , also known as gatekeeper, prevents  $\sigma$  binding to DNA without the core RNAP (Dombroski, Walter and Gross, 1993). Upon RNAP holoenzyme formation, an acidic area contained in region 1.1 mimics DNA and occupies the RNAP active cleft in a way that only promoters similar enough to the consensus sequence that the  $\sigma$  factor recognizes can displace it (Vuthoori *et al.*, 2001; Murakami, 2013). Region 1.1 is only present in group 1  $\sigma^{70}$ s (Paget, 2015). Plastid  $\sigma$  factors contain an unconserved N-terminal region, absent in bacterial  $\sigma$  factors (Schweer, Türkeri, Kolpack, *et al.*, 2010). This region contains a N-terminal acidic area, similar to region 1.1, and a C-terminal basic area (Puthiyaveetil, Ibrahim and Allen, 2013). It is thought that acidic and basic patches interact, preventing this region from entering in the RNAP catalytic cleft and allowing  $\sigma$  factors to be less stringent with the promoter motifs from which they can start transcription (Puthiyaveetil, Ibrahim and Allen, 2013). Region  $\sigma_{1.2}$ , present in group 1 and group 2  $\sigma^{70}$ s, binds to the non-template strand of the promoter “discriminator”, a promoter element with consensus sequence 5’-GGG-3’ located between bases -6 to -4 from the TSS (Zhang *et al.*, 2012) (Fig 1.2). A binding pocket formed by  $\sigma_{1.2}$  and  $\sigma_2$  binds base -6 and it flips out (Zhang *et al.*, 2012).

Some group 1  $\sigma^{70}$ s, such as RpoD in *E. coli*, contain a non-conserved region (NCR), of variable length and sequence, between domains  $\sigma_1$  and  $\sigma_2$ . NCR is present in group 1  $\sigma$  factors from Proteobacteria, *Aquifex*, Spirochaetes and *Chlamydia* (Iyer, Koonin and Aravind, 2004). This region has been implicated in open complex formation in RpoD from *E. coli* since it binds to the DNA upstream of the -10 element (Narayanan *et al.*, 2018). Other functions of the NCR include the inhibition of transcription pausing during early elongation and the promotion of promoter escape (Leibman and Hochschild, 2007).

Domain  $\sigma_2$  makes specific contacts with the single-stranded non-template strand of the -10 element of the promoter, which leads to DNA unwinding in this region (reviewed in (Paget, 2015)) (Fig 1.2). The -10 element is a 6bp stretch of DNA centered at position -10 from the TSS with a consensus sequence of TATAAT in *B. subtilis* and *E. coli* (Liu, Brutlag and Liu, 2001; Feklistov and Darst, 2011; Zhang *et al.*, 2012) (Fig 1.2).  $\sigma_2$  domain can be divided in four regions,  $\sigma_{2.1}$  to  $\sigma_{2.4}$ . Regions  $\sigma_{2.1}$  and  $\sigma_{2.2}$  bind to the clamp helices of the  $\beta'$  subunit of the RNAP, which is considered to be the main attachment interface between  $\sigma$  factors and RNAP apoenzyme (Murakami, Masuda and Darst, 2002). Regions  $\sigma_{2.1}$

and  $\sigma_{2.2}$  form two anti-parallel  $\alpha$  helices with a negatively charged surface that binds to the clamp helices to the  $\beta'$  subunit (Murakami, Masuda and Darst, 2002). Region  $\sigma_{2.3}$  and  $\sigma_{2.4}$  bind to the non-template strand of the -10 element.  $\sigma_{2.3}$  is a flexible loop, whereas  $\sigma_{2.4}$  is an  $\alpha$  helix. These regions are involved in flipping out bases -11 and -7, which starts unwinding in the -10 element (Zhang *et al.*, 2012; Paget, 2015; L. Li *et al.*, 2019). Base -7 is inserted in a pocket created by  $\sigma_2$  and  $\sigma_{1.2}$  regions. However, in ECF  $\sigma$  factors, unwinding at base -7 is performed by the  $\sigma_2$  region and the  $\beta$  subunit gate loop (L. Li *et al.*, 2019). Additionally, ECF  $\sigma$  factors flip out base -12 using  $\sigma_2$  domain (L. Li *et al.*, 2019).



**Figure 1.2. General structure of a group 1  $\sigma^{70}$  proteins.** A. Overview of the structure of *Thermus thermophilus* RNA polymerase in complex with  $\sigma^A$  (PDB: 11W7 (Yokoyama *et al.*, 2002)). Different proteins in the complex and different  $\sigma$  regions are depicted in different colors and labelled.  $\sigma_{1.1}$  region (red) was modelled from the structure of *E. coli* holoenzyme (PDB: 4YG2 (Murakami, 2013)). The  $\beta$  subunit is removed in the second panel to show  $\sigma_{3.2}$  and part of  $\sigma_4$  occupying the RNA exit channel. The second channel, indicated in panel two, is required for the entry of NTPs. Catalytic  $Mg^{2+}$  is indicated by a magenta sphere. B. Overview of the functions of  $\sigma^{70}$  conserved regions.  $\sigma$  factor structure is based on  $\sigma^A$  from *T. thermophilus* (PDB: 11W7 (Yokoyama *et al.*, 2002)). A canonical bacterial promoter with its different elements is depicted. RNA polymerase holoenzyme regions in charge of bind each promoter element are shown.  $\sigma$  factors bind in four regions:  $\sigma_{1.2}$  binds to the discriminator (“DSR”),  $\sigma_{2.3}$  binds to the -10 element,  $\sigma_{3.0}$  binds to the extended -10 element (“EXT”) and  $\sigma_{4.2}$  binds to the -35 element. The C-terminal domain of the  $\alpha$  subunits ( $\alpha$ CTD-I and -II) bind to the UP element.  $\beta$  binds to the core recognition element (“CRE”), located around the transcription start site.  $\beta'$  zipper binds to the Z-element (“-17”), located around the -17 nucleotide. Figure distributed under Creative Commons Attribution License (CC BY) from (Lee and Borukhov, 2016).

Domain  $\sigma_3$  can be divided in three functional regions,  $\sigma_{3.0}$  (also known as  $\sigma_{2.5}$ ,  $\sigma_{3.1}$  and  $\sigma_{3.2}$ . Regions  $\sigma_{3.0}$ , composed by a single  $\alpha$  helix, and  $\sigma_{3.1}$ , with a helix-turn-helix structure, interact with the mayor grove of the extended -10 element in the RNAP open complex (Liu *et al.*, 2017) (Fig 1.2). The extended -10 element occupies bases -15 and -14 from the TSS, contains a consensus 5'-TG-3', and appears in ~20% of the promoters from *E. coli* (Burr *et al.*, 2000). Extended -10 element allow for transcription initiation in the absence of -35 element (Kumar *et al.*, 1993). Instead,  $\sigma_{3.2}$  region expands along the RNAP active cleft and occupies the RNA exit channel, where it reorganizes ssDNA simulating RNA (Zhang *et al.*, 2012). When the nascent RNA is larger than 4bp, it crashes with the  $\sigma_{3.2}$  domain, forcing its expulsion for the completion of promoter escape (Cashel, Hsu and Hernandez, 2003). In ECF  $\sigma$  factors,  $\sigma_3$  is missing and the non-conserved linker between  $\sigma_2$  and  $\sigma_4$  domain fulfills  $\sigma_{3.2}$  function (Fang *et al.*, 2019).

Domain  $\sigma_4$  is composed of four  $\alpha$  helices and can be split into two regions,  $\sigma_{4.1}$  (first and second helices) and  $\sigma_{4.2}$  (third and fourth helices) (Fig 1.2). Region  $\sigma_{4.1}$  binds to the  $\beta$  flap-tip helix, connecting to the active cleft of the RNAP, and  $\sigma_{4.2}$  bins to the -35 element (L. Li *et al.*, 2019). The -35 element is a 6bp stretch of DNA centered at position -35 from the TSS which contains a consensus 5'-GTGACA-3' in *E. coli* (Burr *et al.*, 2000) (Fig 1.2).

Along with the description of  $\sigma^{70}$  regions, I have referred to the most important bacterial promoter regions, including the -10, extended -10 and -35 elements, and the discriminator. Other promoter elements that are important for transcription efficiency are the UP elements – a TA rich motif located 40-60bp from the TSS and bound by the C-terminal domain of the RNAP  $\alpha$  subunit (Estrem *et al.*, 1998) – and the spacer between -35 and -10 elements, which has a consensus length of 17bp (Burr *et al.*, 2000). Deviations from this length favor the recognition of promoters by alternative  $\sigma$  factors (Typas and Hengge, 2006).

### 1.3. ECF $\sigma$ factors

ECFs the most abundant and diverse member of the  $\sigma^{70}$  family. In contrast to other  $\sigma$  factors, ECFs are activated in response to extracytoplasmic or intracellular signals, usually triggered by different types of stress. Indeed, ECFs are the core components of one of the main signal transduction mechanisms in bacteria, only outnumbered by one and two component systems (Staroń *et al.*, 2009). Aside from their importance for bacterial gene expression regulation, ECF  $\sigma$  factors have been used as components of synthetic genetic circuits due to their orthogonality and host-independent behavior (Rhodius *et al.*, 2013; Pinto *et al.*, 2018).

#### 1.3.1. ECF-mediated responses

ECFs mediate the response to different types of stressors. While ECFs activity is often activated in response to external stimuli involved in cell envelope homeostasis and stress adaptation (Lonetto *et al.*, 1994; Grosse, Friedrich and Nies, 2007; Mascher, Hachmann and Helmann, 2007; Paget, 2015),

some ECFs also detect cytoplasmic stimuli and regulate functions such as detoxification of reactive oxidative species, stationary phase survival (Francez-Charlot *et al.*, 2009; Staroń and Mascher, 2010), metal resistance and homeostasis (Grosse, Friedrich and Nies, 2007), morphological changes throughout the cell-cycle (Staroń *et al.*, 2009; Paget, 2015), virulence (Kazmierczak, Wiedmann and Boor, 2005; Llamas *et al.*, 2009) and iron acquisition (Braun, Mahren and Ogierman, 2003). Moreover, ECFs are the main general stress and stationary phase  $\sigma$  factor in Alphaproteobacteria (Francez-Charlot *et al.*, 2009).

### 1.3.2. ECF diversity and classification

ECFs are the most diverse type of  $\sigma$  factors (Staroń *et al.*, 2009). Their first classification grouped ECFs from less than 400 genomes into 67 phylogenetic groups based on sequence similarity, and revealed that conservation at protein level is often accompanied by conservation of the target promoter motif and a genomic neighborhood (Staroń *et al.*, 2009). Altogether, this work proposed that it is possible to predict ECF target promoter, its regulatory mechanism and its target genes from sequence information alone (Staroń *et al.*, 2009). Following studies expanded the number of phylogenetic groups by focusing on nine planctomycetal (Jogler *et al.*, 2012) and 100 actinobacterial (Huang *et al.*, 2015b) genomes, again identifying correlations between protein sequence and function.

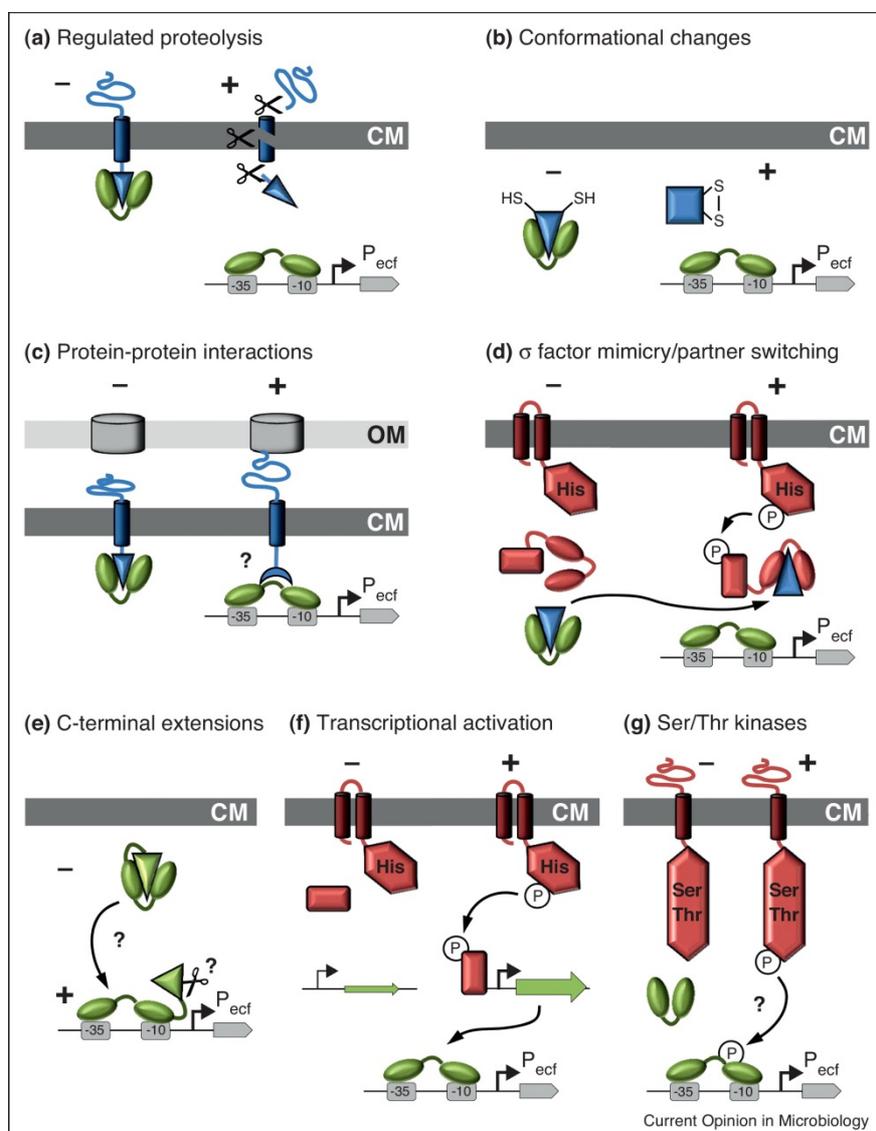
### 1.4. ECF regulators

Given that bacteria typically contain several  $\sigma$  factors that compete for the binding to the RNAP core complex, it is key to regulate the activity of ECFs in response to changing conditions. ECFs are often regulated by anti- $\sigma$  factors, proteins that bind and sequester ECFs in an inactive conformation (Paget and Helmann, 2003). Anti- $\sigma$  factors are typically bound to the membrane and contain a cytoplasmic anti- $\sigma$  domain (ASD) that binds to their cognate ECF. Upon the onset of the inducing signal, anti- $\sigma$  factors undergo conformational changes or get degraded, thereby releasing their cognate ECF, which can then guide RNAP to its specific target promoter to initiate the expression of coding sequences that would respond to the triggering stress (Helmann, 2002).

Besides the regulation via anti- $\sigma$  factors, ECFs may also be regulated by several other mechanisms, such as conformational changes (Li *et al.*, 2002; Campbell *et al.*, 2007), two-component systems (Nizan-Koren *et al.*, 2003; Francez-Charlot *et al.*, 2009) and C-terminal extensions (Wecke *et al.*, 2012; Liu, Pinto and Mascher, 2018) many of which were first predicted by genomics approaches (reviewed in (Mascher, 2013)). A functional role of serine/threonine kinases (STKs) over ECF activity has been proposed from comparative genomic studies (Staroń *et al.*, 2009; Mascher, 2013) and suggested by *in vivo* experiments (Bayer-Santos *et al.*, 2018).

### 1.4.1. Anti- $\sigma$ factors

Anti- $\sigma$  factors exist for all the groups of  $\sigma^{70}$  proteins. Even though anti- $\sigma$  factors often completely inhibit  $\sigma$  factor function, in some cases they are modulators of transcription initiation. For instance, the bacteriophage T4 anti- $\sigma$  factor AsiA binds to the  $\sigma_4$  domain of RpoD in *E. coli* and blocks the expression of genes that require -35 binding for expression (Orsini *et al.*, 1993; Lambert *et al.*, 2004). As a consequence, the RNAP is redirected to the transcription of phage genes (Ouhammouch *et al.*, 1995). Anti- $\sigma$  factors are neither conserved at a sequence nor at a structural level (Campbell, Masuda, *et al.*, 2002; Lambert *et al.*, 2004; Sorenson, Ray and Darst, 2004; Campbell *et al.*, 2007; Maillard *et al.*, 2014; Schumacher *et al.*, 2018; Wei *et al.*, 2019). However, all the anti- $\sigma$  factors that regulate ECFs have been found to have a related structure (Schumacher *et al.*, 2018). So far, three classes of ECF anti- $\sigma$  factors have been described according to the structure of their ASD.



**Figure 1.3. Main regulatory mechanisms of ECF  $\sigma$  factors.** ECF  $\sigma$  factors are colored in green, anti- $\sigma$  factors are colored in blue and other regulators that catalyze phosphorylation are colored in red. This figure was taken from (Mascher, 2013) is licensed to be reused in this thesis by Elsevier under number 4751271218118.

Class I ASDs (ASDIs) regulate ~33% of the ECF  $\sigma$  factors and are the most common type of ECF anti- $\sigma$  factor (Campbell *et al.*, 2007). About 70% of the class I anti- $\sigma$  factors are bound to the membrane (Campbell *et al.*, 2007). This is likely related to the common extracytoplasmic nature of the signals they transmit. ASDIs contain four  $\alpha$  helices, of which the first three form a helix bundle connected to the fourth helix by a flexible linker (Campbell *et al.*, 2003; Anthony, Newman and Donohue, 2004; Shukla *et al.*, 2014; Devkota *et al.*, 2017). In zinc-binding anti- $\sigma$  factors (ZAS), the first three  $\alpha$  helices may be stabilized by a zinc-finger, which in some cases fulfills an active role in the regulation of the anti- $\sigma$  factor. ZAS are divided in two groups, Hx<sub>23-26</sub>Hx<sub>3</sub>Cx<sub>2</sub>C (in short, HHCC) and Cx<sub>23-26</sub>Hx<sub>3</sub>Cx<sub>2</sub>C (CHCC), depending on the fourth residue that coordinates zinc (Rajasekar *et al.*, 2016). CHCC anti- $\sigma$  factors have been suggested to be regulated by the state of zinc coordination, whereas the in the HHCC type zinc fulfills only an structural role (Rajasekar *et al.*, 2016). Nevertheless, this vision has been rejected since RsiW from *B. subtilis*, which holds a CHCC motif, is insensitive to oxidative stress (Devkota *et al.*, 2017), while ChrR in *Rhodobacter sphaeroides* is activated by oxidative stress and bears a HHCC motif (Anthony, Warczak and Donohue, 2005).

Class II ASDs (ASDIIs) are a simpler version of ASDIs. ASDIIs only contain two  $\alpha$  helices, the first of which binds to the  $\sigma_4$  domain and the second to the  $\sigma_2$  domain (Herrou *et al.*, 2012; Maillard *et al.*, 2014). Two member of this group, CnrY from *Cupriavidus metallidurans* and PhyR from *Caulobacter crescentus*, have been crystalized (Herrou *et al.*, 2012; Maillard *et al.*, 2014). CnrY is a transmembrane protein that regulates cobalt and nickel resistance (Maillard *et al.*, 2014), whereas PhyR is a soluble protein that regulates the general stress response  $\sigma$  factor EcfG in Alphaproteobacteria. Aside from EcfG, PhyR also binds to an anti-anti- $\sigma$  factor, NepR, that becomes active in response to stress (Francez-Charlot *et al.*, 2009) (see Section 1.4.1.3 and Fig. 1.3d).

Class III ASDs (ASDIIs) contain three  $\alpha$  helices, where the first and second bind to  $\sigma_4$  domain and the third to  $\sigma_2$  domain (Schumacher *et al.*, 2018). The structure of only one member of this group, BldN from *Streptomyces venezuelae*, has been resolved (Schumacher *et al.*, 2018). This protein is involved in the regulation of the production of aerial hyphae (Schumacher *et al.*, 2018).

Inactivation of anti- $\sigma$  factors is required for the release of the ECFs to start transcription and happens through different means, which are generally conserved for anti- $\sigma$  factors that regulate ECFs from the same group. In most of the cases anti- $\sigma$  factors are not the sensors of the signal that triggers their inactivation, but other proteins that in turn regulate anti- $\sigma$  factors. Mechanisms of ECF inactivation have been the focus of several great reviews (Mascher, 2013; Treviño-Quintanilla, Freyre-González and Martínez-Flores, 2013; Sineva, Savkina and Ades, 2017).

#### 1.4.1.1. Regulated intramembrane proteolysis

Most of the transmembrane anti- $\sigma$  factors are inhibited by regulated intramembrane proteolysis (RIP) (Fig. 1.3a). This system requires of three proteases for anti- $\sigma$  factor degradation. Site-1 protease, which may be the sensor of the system, acts in the periplasm and creates the substrate for site-2

protease, a metalloprotease that cleaves the anti- $\sigma$  factor within the transmembrane helix. Then, the anti- $\sigma$ /ECF complex is released into the cytoplasm and a soluble unspecific protease finishes anti- $\sigma$  factor degradation (reviewed in (Heinrich and Wiegert, 2009)).

The most studied example of RIP occurs in the context of RseA, the anti- $\sigma$  factor of the essential ECF RpoE (group ECF02) in *E. coli*. RpoE is activated in response to unfolded outer membrane proteins (OMPs) in the periplasm (Mecenas *et al.*, 1993). The site-1 protease, DegS, senses the presence of C-terminal regions of unfolded OMP with its PDZ domain and starts the degradation of RseA (Alba *et al.*, 2002). RseB, encoded in the same operon as RseA, needs to be removed from the surface of the periplasmic part of RseA for the site-1 proteolysis to happen (Chaba *et al.*, 2011). RseB removal from RseA surface is achieved by the binding of intermediates of lipopolysaccharide (LPS) synthesis, or LPSs with a modified structure (Tam and Missiakas, 2005; Lima *et al.*, 2013). In this way, DegS and RseB function as two inputs of an AND gate that results in the degradation of RseA (Chaba *et al.*, 2011). Then, the site-2 protease RseP cleaves RseA and the degradation of the remaining cytoplasmic peptide is finished by ClpXP (Alba *et al.*, 2002). RseP harbors two PDZ domains, involved in sensing RseA site-1 degradation intermediate (Li *et al.*, 2009). Some details of RpoE regulation, such as the role of RseC, encoded in the same operon and found to positively modulate RpoE activity remain unsolved (Missiakas *et al.*, 1997).

Another example of RIP-regulated anti- $\sigma$  factor is RsiW from *B. subtilis*, which sequesters SigW (group ECF01 in (Staroń *et al.*, 2009)). In response to cell envelope stress caused by antimicrobial peptides, RsiW is cleaved by the site-1 metalloprotease PrsW (reviewed in (Ho and Ellermeier, 2012)). This is not enough for the site-2 protease, RasP, to perform its activity, and hence it is thought that there are more players in RsiW degradation (Heinrich, Hein and Wiegert, 2009). After RasP has performed its function, the cytosolic protease ClpXP finishes RsiW degradation (reviewed in (Ho and Ellermeier, 2012)).

Yet, another RIP-regulated anti- $\sigma$  factor, RsiV from *B. subtilis*, has been described. RsiV inhibits SigV (group ECF30 in (Staroń *et al.*, 2009)), which is in charge of responding to lysozyme (Ho *et al.*, 2011). RsiV contains a two amphipathic, periplasmic  $\alpha$  helices that hinder the site-1 protease cleavage site when they lay on the cell membrane (Lewerke *et al.*, 2018). RsiV C-terminus binds to lysozyme, which forces the exposure of the site-1 protease, signal peptidase, cleavage site (Castro *et al.*, 2018; Lewerke *et al.*, 2018). Then, the site-2 protease RasP and other cytosolic proteases finish RsiV degradation (Lewerke *et al.*, 2018).

#### 1.4.1.2. Conformational changes

Soluble anti- $\sigma$  factors are typically regulated by conformational changes (Fig. 1.3b). The prototype of this system is RsrA from *Streptomyces coelicolor*, which regulates the ECF SigR (group ECF12). RsrA is a CHCC-type ZAS that coordinates zinc under reducing conditions (Li *et al.*, 2003). Upon the onset of oxidative stress, disulfide bridges form between pairs of cysteine residues either within or

across proteins, causing disulfide stress. RsrA senses this disulfide stress by expelling its zinc atom and forming a disulfide bridge between the first and the last cysteine of its CHCC zinc-binding motif (Li *et al.*, 2003). As a consequence of this conformational change, RsrA is incapable of binding to SigR (Li *et al.*, 2003).

ChrR from *Rhodobacter sphaeroides*, the anti- $\sigma$  factor of RpoE (group ECF11), is also regulated by conformational changes. ChrR is soluble and contains a C-terminal, zinc-binding, cupin-like domain, conserved in anti- $\sigma$  factors of members of group ECF11. The cupin-like domain of ChrR is required for singlet oxygen response (Campbell *et al.*, 2007). Singlet oxygen prevents binding of ChrR to RpoE, although this seems to be unrelated to the ability of the cupin domain to bind zinc (Greenwell, Nam and Donohue, 2011). It has been suggested that ChrR response to singlet oxygen is dependent upon oxidation of the cysteines in the HHCC-type zinc-binding domain of ChrR ASD (Chabert *et al.*, 2019). Their oxidation would destabilize the zinc-finger fold and lead to its unfolding (Chabert *et al.*, 2019).

Interestingly, ECFs can directly sense redox state by the formation of disulfide bridges between conserved cysteine residues in their  $\sigma_4$  domain (Shukla *et al.*, 2014). This mechanism can be observed in SigK from *Mycobacterium tuberculosis* (Shukla *et al.*, 2014), which is also regulated by the membrane-bound anti- $\sigma$  factor RskA, subjected to RIP. SigK disulfide bridge stabilizes its binding to the RskA's ASD (Shukla *et al.*, 2014).

#### 1.4.1.3. Partner switching

Partner switching is a mechanism of anti- $\sigma$  factor regulation that has been described for the soluble anti- $\sigma$  factors that inhibit the general stress response ECFs in Alphaproteobacteria, which are part of group ECF15 (Fig. 1.3d). In this mechanism, an anti-anti- $\sigma$  factor sequesters the cognate anti- $\sigma$  factor, thereby releasing the ECF (Francez-Charlot *et al.*, 2009). The anti-anti- $\sigma$  factor is fused to the response regulator of a 2CS, generally encoded in the same genetic neighborhood. This anti-anti- $\sigma$  factor is able to bind the anti- $\sigma$  factor when the response regulator domain is phosphorylated (Francez-Charlot *et al.*, 2009). In this way, the anti- $\sigma$  factor would bind the anti-anti- $\sigma$  factor instead of the ECF, which is then free to start transcription from the appropriate promoter regions (Francez-Charlot *et al.*, 2009).

The hallmark of this system regulates SigT in *Caulobacter crescentus*. The soluble class II anti- $\sigma$  factor NepR inhibits SigT activity (reviewed in (Francez-Charlot *et al.*, 2015)). However, upon phosphorylation of the response regulator domain of the anti-anti- $\sigma$  factor, PhyR, SigT is released and NepR binds PhyR (Lourenço, Kohler and Gomes, 2011). The anti-anti- $\sigma$  factor domain of PhyR is an ECF-like protein, similar enough to SigT to bind NepR, but divergent enough to not be able to induce transcription initiation. The inducing signal of this system is sensed by the histidine kinase PhyK, and probably other histidine kinases (Foreman, Fiebig and Crosson, 2012), which transmits the signal to the response regulator domain of PhyR (Lourenço, Kohler and Gomes, 2011).

#### 1.4.1.4. Cell-surface signaling in FecIR-like systems

FecR-like anti- $\sigma$  factors regulate FecI-like ECFs depending on the signal sensed by a FecA-like outer membrane transporter (Fig. 1.3c). FecIR system from *E. coli* is the best studied example of this type of regulation, although similar systems are often found across Proteobacteria and Bacteroidetes (Staroń *et al.*, 2009). FecI promotes the uptake of ferric citrate via the transcriptional activation of its transport system, encoded in the *fecABCDE* operon (Van Hove, Staudenmaier and Braun, 1990). The first protein encoded in this operon, FecA, is an outer-membrane TonB-dependent receptor that transports citrate loaded with  $\text{Fe}^{3+}$ . FecA transmits the presence of ferric citrate to FecR. As a consequence, the complex formed by FecI and the N-terminal part of FecR directs RNAP to the transcription of the transport system *fecABCDE* (reviewed in (Braun, Mahren and Ogierman, 2003)). A special case is the FecI-like ECF PrhI from *Ralstonia solanacea*, since it participates in plant infection (Braun and Mahren, 2005). Moreover, when cells are depleted of  $\text{Fe}^{2+}$ , Fur releases its repression over the transcription of *fecIR* operon, so that *fecABCDE* transport system can be synthesized (Braun, 1997).

FecR-like anti- $\sigma$  factor are considered as a regulator rather an inhibitor of ECF activity since they have pro- $\sigma$  activity, this is, FecR-like anti- $\sigma$  factors enhance FecI binding to the RNAP (Mahren and Braun, 2003). However, this pro- $\sigma$  activity is not a general feature of FecR-like anti- $\sigma$  factors, since it has not been found in some FecIR-like systems (Quesada *et al.*, 2016). Work by Bastiaansen *et al.* found that FecR-like anti- $\sigma$  factors are proteolytically cleaved as part of their mechanism of ECF inhibition (Bastiaansen *et al.*, 2014, 2015).

#### 1.4.2. Hanks type kinases

ECFs from certain groups, including ECF43 and ECF59-62 show microsynteny with Hanks-type serine/threonine kinases (STKs) (reviewed in (Mascher, 2013), Fig. 1.3g). Hanks-type kinases transfer the  $\gamma$ -phosphate of ATP or GTP to the hydroxyl group of serine, threonine or tyrosine, forming a phosphate monoester (Hanks and Hunter, 1995). Hanks-type kinases are described by a common two-lobed structure, where the smaller lobe, mainly composed of antiparallel  $\beta$ -sheets, binds and orientates the donating nucleotide, whereas the larger  $\alpha$ -helical lobe binds to the acceptor peptide and initiates phosphotransfer (Hanks and Hunter, 1995). Hanks-type kinases were first discovered in Eukaryotes, but were later found in Bacteria and Archaea, suggesting that they have a common evolutionary origin (reviewed in (Stancik *et al.*, 2018)). In contrast to Eukaryotes, tyrosine residues are not usually targeted by Hanks kinases in Bacteria (Janczarek *et al.*, 2018). STK-mediated gene expression has been proven to be essential for various bacterial cellular processes, such as growth, iron transport, secondary metabolite production, antibiotic resistance and virulence (reviewed in (Janczarek *et al.*, 2018)). In these processes STKs phosphorylate response regulators of 2CSs and key components of the transcription and translation machinery such as transcriptional regulators (reviewed in (Janczarek

*et al.*, 2018)). Phosphorylation might activate or inactivate the activity of the target proteins (Janczarek *et al.*, 2018). At the same time, cross-phosphorylation between STKs is possible (Cousin *et al.*, 2013; Janczarek *et al.*, 2018). Given the importance of STKs in bacterial metabolism, they may also control ECF activity. A functional role of a the STK PknS in the regulation of a ECF EcfK has been found in *X. citri* (Bayer-Santos *et al.*, 2018); however, the direct phosphorylation of an ECF  $\sigma$  factor has never been proven.

#### 1.4.3. Extensions of the ECF sequence

Some ECFs contain additional domains aside from the core ECF regions ( $\sigma_2$  and  $\sigma_4$  domains connected by a non-conserved linker) (Fig. 1.3e). The most common type of extensions of the ECF sequence occur in the C-terminus, which involve different types of domains depending on the ECF group (Staroń *et al.*, 2009; Pinto, Liu and Mascher, 2019). A functional role of the C-terminal extension of groups ECF41, ECF42 and ECF44 has been described (Gómez-Santos *et al.*, 2011; Wecke *et al.*, 2012; Liu, Pinto and Mascher, 2018; Wu *et al.*, 2019).

#### 1.4.4. Alternative modes of regulation

Transcriptional regulation of ECF expression occurs in members of group ECF32, whose expression is controlled by the transcription factors HrpSR, which are in turn regulated by a 2CS (Merighi *et al.*, 2003; Nizan-Koren *et al.*, 2003) (Fig. 1.3f). Members of ECF39 are directly regulated at a transcriptional level by a 2CS (Luo *et al.*, 2014; Tran *et al.*, 2019). Non-ECF  $\sigma$  factors are also regulated at the level of translation and through molecules such as ppGpp. For instance, the translation of SigH in *E. coli* is active under exposure to high temperature due to the disruption of the secondary structure of its mRNA (Nagai, Yuzawa and Yura, 1991), and ppGpp increases the amount of RNAP bound to the stationary phase  $\sigma$  factor RpoS (Jishage *et al.*, 2002). However, to date neither of these types of regulation has been described for ECF  $\sigma$  factors.

#### 1.5. Common components of signal transduction mechanisms

Aside from ECF  $\sigma$  factors, bacteria contain several systems that allow them to sense and transduce signals. These systems are essential for environmental fitness and provide an adaptive advantage when several bacterial species are competing for survival in the same niche. Signal transduction mechanisms sense extracellular signals and transmit them to the cytoplasm. However, bacteria also sense cytoplasmic concentrations of certain molecules, such as reactive oxygen species or the alarmone (p)ppGpp, in order to keep homeostasis. Signal transduction is typically carried out by a modular protein with an extracytoplasmic sensing N-terminus and a cytoplasmic signal output C-terminal domain, although this protein does not need to be transmembrane (Parkinson and Kofoed, 1992; Galperin, 2004). These two domains could be in the same protein – for instance, in 1CSs with a DNA-binding output domain – or in different proteins – for instance in 2CSs, which are composed of

a sensory transmembrane histidine kinase that phosphorylates a cytoplasmic receiver domain generally fused to a DNA-binding response regulator (reviewed in (Hoch, 2000)). Some signal transduction mechanisms, such as phosphorelay systems, contain intermediate signal transducers between the sensing and the output domain (reviewed in (Hoch, 2000)).

Extracellular sensing domains are largely uncharacterized. One of the most common sensing domains is PASTA, which binds  $\beta$ -lactams (Yeats, Finn and Bateman, 2002). The most common cytoplasmic sensory domains are PAS and GAF domains (reviewed in (Galperin, 2004)). PAS and GAF domains appear in a variety of sensing proteins, including histidine kinases from 2CSs, and are able to accommodate a wide variety of small-molecule ligands (reviewed in (Galperin, 2004)). Some integral membrane segments may also have sensing functions, for instance, ethylene receptors in plants and bacteria (Mount and Chang, 2002).

Once the signal is sensed from the periplasm or the environment, it must be transmitted to the intracellular signal transduction domain. The mechanism of this process is largely unknown, but it involves rotation and/or piston-like movements of the transmembrane helices of histidine kinase dimers in 2CSs (Casino, Rubio and Marina, 2010; Zhang and Hendrickson, 2010). This movements are translated in the cytoplasm into autophosphorylation in the case of histidine kinase dimers of 2CSs and STKs (reviewed in (Casino, Rubio and Marina, 2010; Pereira, Goss and Dworkin, 2011)). Aside from histidine kinase domains in 2CSs, and STKs, other common transducer domains are methyl-accepting elements in chemotaxis proteins, phosphatases, and enzymes in charge of the synthesis of the secondary messengers cAMP (adenylate cyclases) and c-di-GMP (diguanylate cyclases), and their degradation (phosphodiesterases) (reviewed in (Galperin, 2004)).

From this point, the signal takes the shape of a transferable phosphoryl group or the secondary messengers cAMP and c-di-GMP. Phosphoryl groups are transmitted either to another intermediate protein or directly to the protein that provides the output of the transduction, which usually contains a DNA binding domain. However, cAMP binds directly the cAMP receptor protein (CAP) (Martínez-Antonio and Collado-Vides, 2003). CAP is an activator of gene expression found to regulated >50% of *E. coli* coding sequences (Martínez-Antonio and Collado-Vides, 2003). Aside from DNA-binding domains, output domains can have enzymatic activity, regulated by the signal transduction pathway.

#### **1.6. Bioinformatics applied to signal transduction mechanism research**

Comparative genome analyses are key elements of the research in signal transduction mechanisms since 1) the domains that take part in these systems are often conserved, and 2) large-scale *in silico* analyses provide a way to compare the ability of different bacteria to respond to environmental changes (Galperin, 2004). Similarly, ECF research has been heavily influenced by comparative genomics since the first description of these subfamily of  $\sigma$  factors (Lonetto *et al.*, 1994). Common methods used in comparative genomics rely in sequence similarity searches to find more members of a certain family of proteins. Sequence information contained in large families of homologous proteins

can be utilized to define subfamilies with specialized functions. Moreover, phylogenetic analyses that focus on the similarity between proteins can be used to assess the co-evolution between families of proteins, and co-variation-based computational tools can be harnessed to predict the network of amino acid interactions that define the three-dimensional structure of a family of proteins (de Juan, Pazos and Valencia, 2013).

One of the latter methods is direct-coupling analysis (DCA) (Martin Weigt *et al.*, 2009). DCA scores the co-variation between pairs of residues, which indicates their likelihood of interaction (Martin Weigt *et al.*, 2009). When two residues interact, deleterious mutations in one can be translated into a compensatory mutation in the second residue that would restore the contact (Martin Weigt *et al.*, 2009). Local co-variation approaches cannot distinguish between couplings that come from direct or indirect interactions (Martin Weigt *et al.*, 2009). For instance, if residue  $y$  interacts with  $x$ , and  $x$  with  $z$ , local co-variation approaches could show a coupling between  $x$  and  $z$ . However, DCA uses global inference implemented through a message passing approach to determine indirect interactions and discard them from the coupling score (Martin Weigt *et al.*, 2009). Recent implementations of DCA, such as Gaussian DCA, have managed to reduce the execution time, while keeping accuracy (Baldassi *et al.*, 2014).

## 2. Aim and objectives

ECF  $\sigma$  factors are the core element of one the main signal transduction mechanism in bacteria. Phylogenetic analyses of ECF  $\sigma$  factors revealed their great abundance and diversity, both in sequence and regulation, with more than 90 phylogenetic groups that usually feature a conserved type of regulator, target promoter motif and cellular response (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Mascher, 2013; Huang *et al.*, 2015a). Despite their great diversity, only ~3,500 ECFs from ~500 sequenced organisms had been phylogenetically analyzed at the start of this work. Given this low number, the original ECF classification could have overlooked alternative, rarer regulatory mechanisms, only evident when looking at a larger dataset. Furthermore, an increase in size and diversity of current ECF groups would allow for a group-wise application of co-variation analyses. These analyses could guide the description of the most important contacts that mediate ECF regulation by elements such as anti- $\sigma$  factors and C-terminal extensions in specific phylogenetic groups.

Considering the importance of the ECF subfamily, its analysis in a limited number of organisms, and the lack of an in-depth understanding of its regulation, the first aim of this work is to test whether the ECF subfamily could be expanded, in diversity and number of proteins, using the proteomes available in public databases. If this is possible, this work would aim at studying the diversity and regulation of ECF  $\sigma$  factors. For this purpose, I will:

1. Systematically find ECFs in bacterial genomes, classify them according to protein sequence similarity, and perform a comprehensive analysis of the natural diversity of ECFs-based signal transduction systems.
2. Analyze the mechanisms that govern the interaction between ECFs and their most common regulators, anti- $\sigma$  factors.
3. Establish the functional role of STKs and C-terminal extensions in ECF activity in different ECF groups.

The general goal of this work is to provide the scientific community with a comprehensive guide on the regulation, target promoter and function of members of the ECF  $\sigma$  factor subfamily according to their phylogenetic group.

## **Results**

The following four sections describe the results of this thesis (Sections 3-6). Each section is followed by a short discussion and summary of the main findings. A general discussion can be found in Section 7.

### 3. Extracytoplasmic function $\sigma$ factor (ECF) extraction and classification

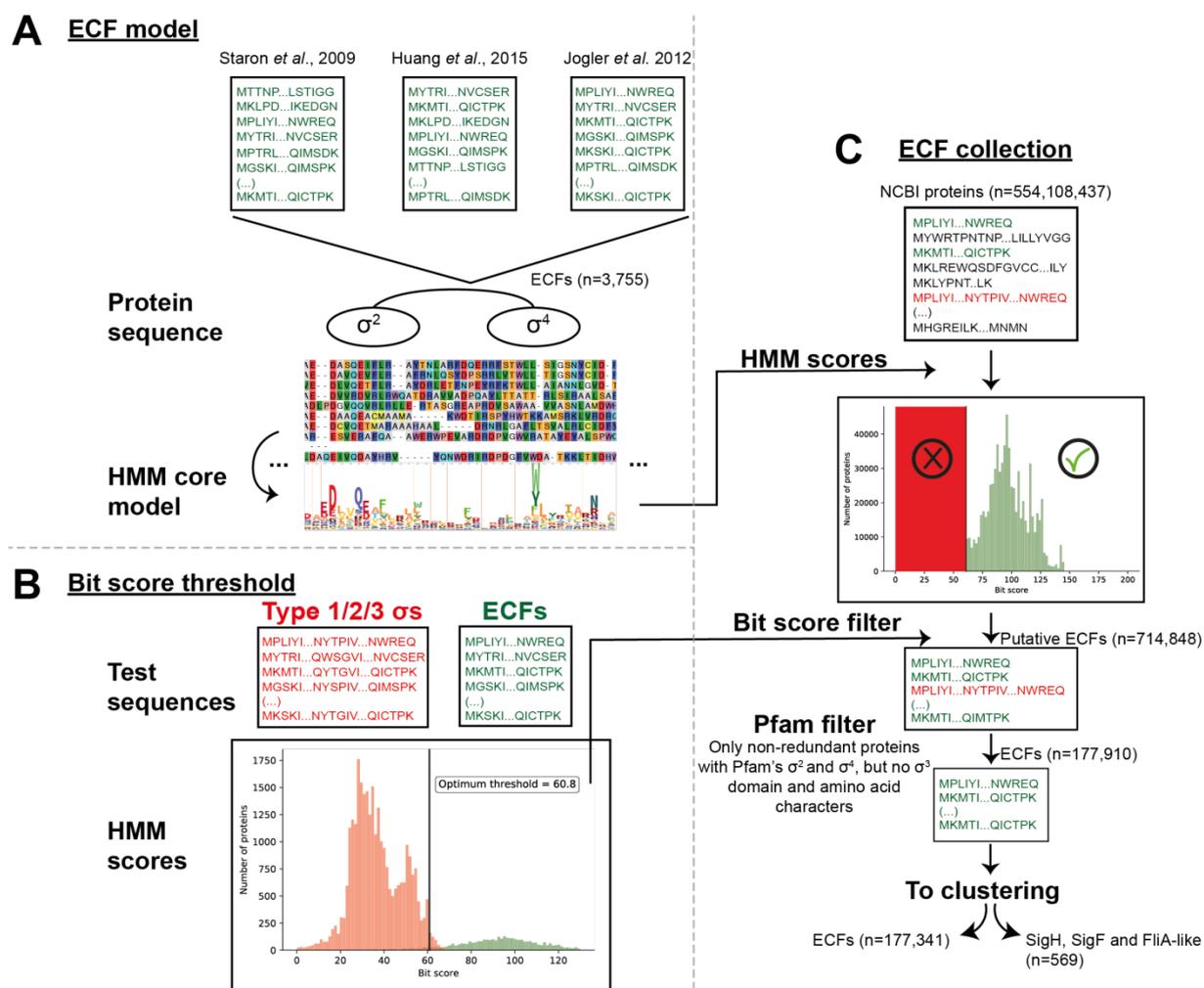
Although the initial ECF classification studies helped to understand the large diversity of ECFs, they addressed ECFs from a limited number of genomes and/or focused on specific phyla (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b). Based on the relatively sparse sequence space, some of the original ECF groups contain only few (less than 10) proteins – the so-called “minor” groups (ECF100, ECF102, etc.) – or cluster divergent sequences into a single group (e.g. ECF01, ECF10, ECF20) (Staroń *et al.*, 2009; Jogler *et al.*, 2012). In light of these limitations, a comprehensive and robust expansion of the ECF classification that reflects the massive increase in sequenced bacterial genomes is essential for the study of ECF  $\sigma$  factors.

In this section, I will address the search for ECF  $\sigma$  factors in all available genomes and metagenomes deposited in the National Center for Biotechnology Information (NCBI) as per February 2017. As a result of this search, I increased the number of ECF proteins 50-fold. I defined 157 ECF groups, of which 22 had sequence similarity to previously described ECF groups. Furthermore, 62 original groups were preserved and expanded with previously unclassified sequences. The data I collected on the features of the resulting 157 groups (Table S3.1) will be publicly accessible in the database “ECF hub”, which will also provide tools for ECF analysis. The possibilities offered by this hierarchical, comprehensive classification will be illustrated throughout the following sections, including the application of co-variation-based methods, such as DCA, for the prediction of contacts between ECFs and their regulators. This wealth of data represents a valuable resource to both computational and experimental researchers for guiding the characterization of ECF  $\sigma$  factors of unknown function. Part of the content of this section has been published as a preprint in bioRxiv (Casas-Pastor *et al.*, 2019).

#### 3.1. The number of identified ECFs is 50-fold larger than in the founding ECF classification

The number of protein sequences in public databases has expanded extensively since the founding ECF classification efforts (Staroń *et al.*, 2009), suggesting a proportional increase in the number of ECFs. To identify novel ECFs, I first extracted the sequences from all previous ECF classification efforts (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b), aligned them and created a general ECF Hidden Markov Model (HMM) for the ECF core region, including the linker between  $\sigma_2$  and  $\sigma_4$ , but excluding any potential protein domains fused N- or C-terminally to the ECF (Fig. 3.1A). To distinguish ECFs from other  $\sigma$  factors, I first scored this general ECF HMM against two sets of training sequences, 1) true ECFs from the founding classification and 2) a negative control set of  $\sigma^{70}$  factors from groups 1, 2 and 3 that additionally contain domains  $\sigma_3$ , and  $\sigma_1$  in some cases. This allowed for the definition of a threshold score that maximizes true positive ECFs – with a score higher than the threshold (Fig. 3.1B; *green*) – while minimizing the number of false positive  $\sigma$  factors – with a score lower than the threshold (Fig. 3.1B; *red*). Then, I selected the non-redundant protein

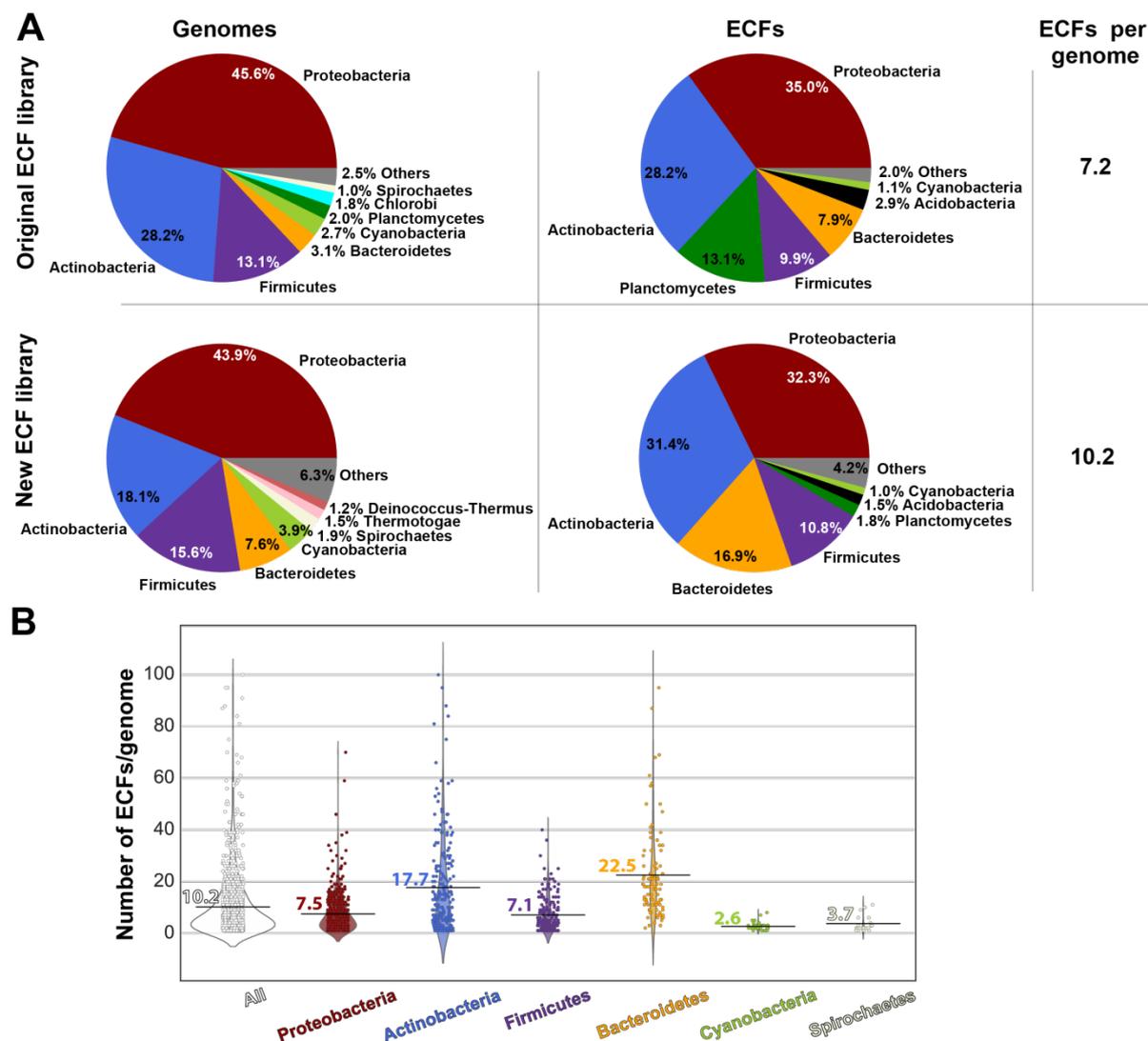
sequences from the NCBI database for which the generic ECF HMM yielded scores higher than this threshold (Fig. 3.1C). As further quality controls, I filtered for sequences containing the Pfam domains  $\sigma_2$  and  $\sigma_4$  but lacking the  $\sigma_3$  domain, and discarded proteins with non-amino acidic characters, such as X or J. This resulted in a library of 177,910 non-redundant candidate ECF sequences. Some of the candidate ECF  $\sigma$  factors included in this list clustered together with group 3  $\sigma^{70}$ s, indicating the presence of a cryptic  $\sigma_3$  domain, which prompted me to remove them from the list of ECF  $\sigma$  factors. This left me with 177,341 non-redundant ECFs, accounting for a ~50-fold expansion over the number of identified ECFs in the founding ECF classification (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b) (Fig. 3.1C).



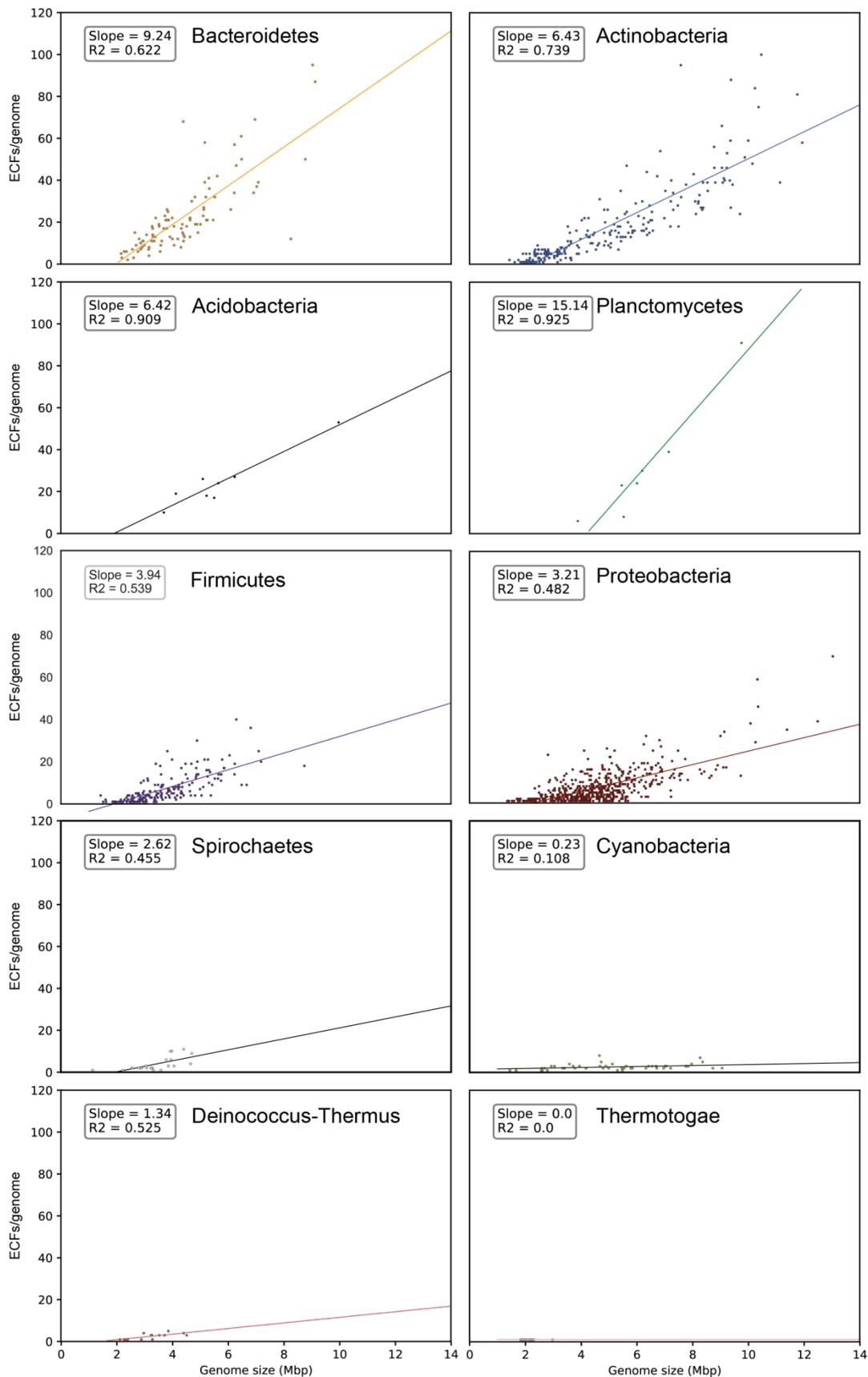
**Figure 3.1. ECF retrieval pipeline.** **A:** ECF sequences from previous classification efforts (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b) were collected and aligned. An HMM was built from the area containing  $\sigma_2$ , linker and  $\sigma_4$  regions. **B:** ECFs from (A) were used as positives, while the  $\sigma$  factors containing a  $\sigma_3$  domain in the Pfam database were used as negatives for the definition of an optimal bit score threshold for ECF extraction. Positives and negatives were scored using the HMM model from (A). The bit score threshold that produced the largest specificity and sensitivity in the classification process was derived with a ROC curve. **C:** The HMM model from (A) was used to score all proteins from NCBI as per February 2017, using as threshold the bit score defined in (B). Putative ECFs without  $\sigma_2$  or  $\sigma_4$  domain, or with  $\sigma_3$  domain, or proteins with characters that do not denote natural amino acids, were discarded. The final set of non-redundant ECFs included 177,910 proteins. Source: (Casas-Pastor *et al.*, 2019).

Next, I analyzed the taxonomic origin of this expanded ECF library to determine the typical number of ECFs found in individual bacterial phyla. To enable such statistics, I focused on the subset of

complete genomes of non-metagenomic origin, labeled by NCBI as “reference” or “representative” genomes, thereby mitigating bias towards heavily sequenced species. Analysis of the 12,539 ECFs extracted from 1,234 of these genomes showed that the taxonomic distribution of the input genomes became more diverse than in the original classification efforts (Fig. 3.2A; *Genomes*). In particular, the fraction of the three most abundant phyla – Proteobacteria, Actinobacteria and Firmicutes – was reduced from 86.9% in the original to 77.6% in the new classification. This reduction was accompanied by an increase in the number of species from underrepresented phyla, such as Bacteroidetes and Cyanobacteria (Fig. 3.2A; *Genomes*). In addition, 19 new ECF-containing phyla emerged (Table S3.2). Yet, these 19 phyla have a limited contribution to the overall ECF database, given their low number of sequenced genomes. This difference in the taxonomic origin of the species included in original and new classifications naturally changes the taxonomic origin of ECFs gathered in each library. For instance, the fraction of ECFs from underrepresented genomes, such as Bacteroidetes and Planctomycetes, is larger in the new ECF library (Fig. 3.2A; *ECFs*). This is not the case for Cyanobacteria and Acidobacteria, which contribute a smaller percentage of ECFs than in the original library (Fig. 3.2A; *ECFs*). These differences in taxonomic composition in the ECF library are reflected in the average number of ECFs per genome, which increases from approx. seven ECFs per genome in the original ECF library (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b) to about ten ECFs per genome in the new library (Fig. 3.2B). Confirming the findings of previous reports (Staroń *et al.*, 2009; Huang *et al.*, 2015b), the number of ECFs per genome is directly proportional to genome size (Fig. 3.3), although the average number of ECFs per genome depends on the phyla of origin (Fig. 3.2B). Bacteroidetes and Actinobacteria have the greatest abundance of ECFs, with an average of 22.5 and 17.7 ECFs per genome, respectively (Fig. 3.2B). Phyla with a lower abundance of ECFs include Cyanobacteria and Spirochaetes, with an average of 2.7 and 3.7 ECFs per genome, respectively (Fig. 3.2B). Firmicutes and Proteobacteria contain an intermediate number of ECFs, 7.1 and 7.5, respectively (Fig. 3.2B). These differences might indicate different dependence on ECFs as signal-transduction system in different phyla, as previously noticed for Actinobacteria, which are particularly rich in ECFs, but also in 1CS and 2CS (Huang *et al.*, 2015b).



**Figure 3.2. Taxonomic analysis of the ECF library.** A: Taxonomic composition of the input genomes, ECFs and average number of ECFs per genome in the original ECF classification (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b) and in this work. For the data of this work, I only included ECFs and genomes from complete and non-metagenomic assemblies tagged as “representative” or “reference” by NCBI, selecting RefSeq assemblies when both RefSeq and GenBank assemblies are available for the same genome. B: Number of ECFs per genome for phyla with more than 20 complete genomes available. Average number of ECFs per genome is shown. Source: (Casas-Pastor *et al.*, 2019).

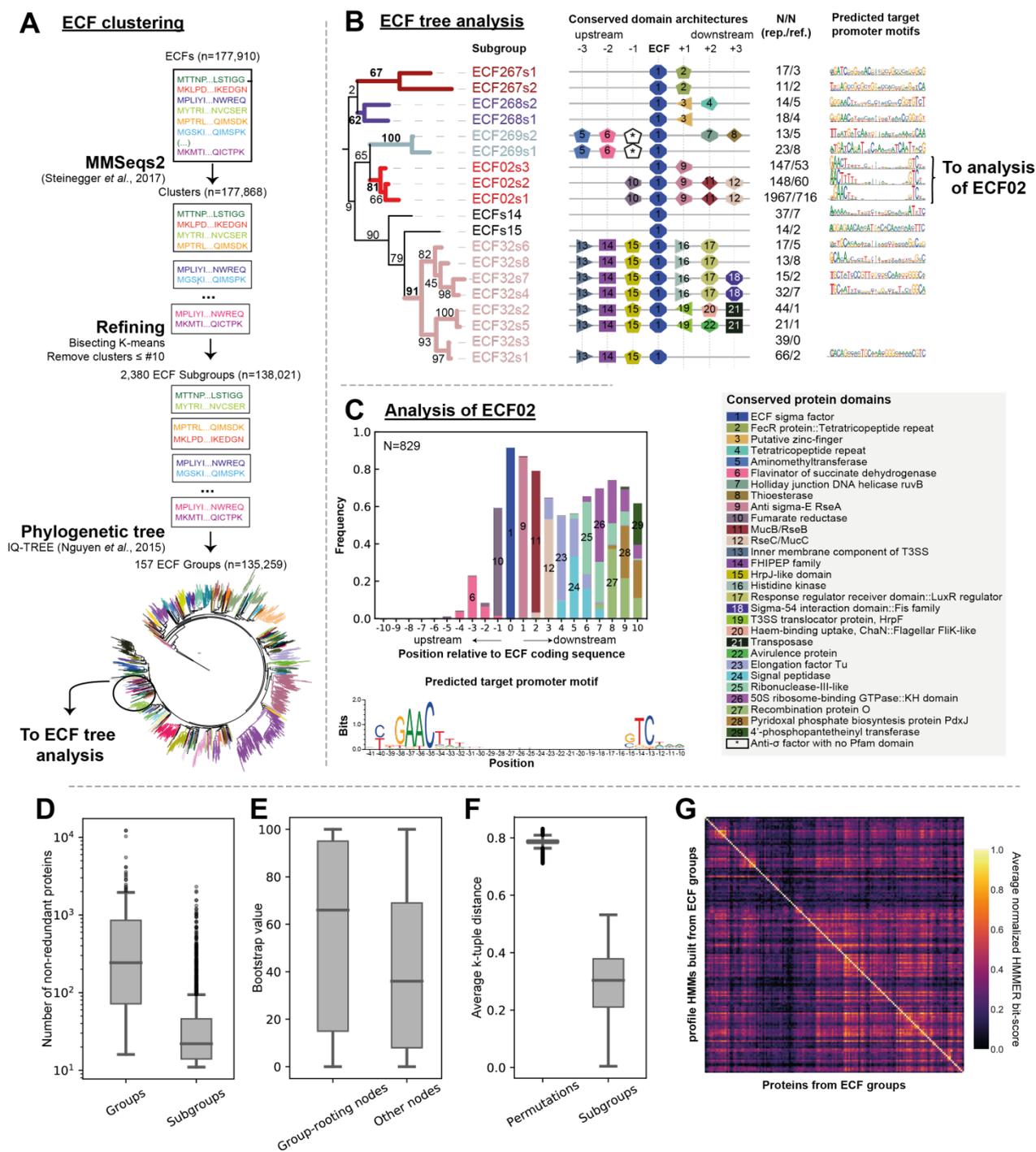


**Figure 3.3. Number of ECFs per genome according to genome size for different phyla.** A linear regression with its fitting parameters is shown for each graph as a guide to the eye. Larger genomes, associated to more complex life styles, tend to have a larger number of ECFs. Some phyla are rich in ECFs, while others do not contain any (Thermotogae). Source: (Casas-Pastor *et al.*, 2019).

### 3.2. The ECF classification 2.0

The wealth of new proteins identified in the library expansion prompted the reclassification of ECF  $\sigma$  factors into groups with common characteristics. To this end, I first subjected the 177,910 protein sequences of the new ECF library to the rapid MMSeqs2 clustering algorithm (Steinegger and Söding, 2017), followed by a quality step that bisects the resulting clusters until the maximum pairwise k-tuple distance between sequences was  $\leq 0.60$  (Fig. 3.4A). Clusters with  $\leq 10$  proteins were discarded to ensure high sequence coherence within clusters, while preventing an explosion of small clusters with limited statistical relevance (Fig. 3.4A). This procedure yields a total of 2,380 ECF clusters (referred to as “subgroups”) with a median of 22 non-redundant proteins per subgroup (Fig. 3.4D). Subgroups capture 77.3% of the proteins, while 22.7% of the proteins remain unclassified, similar to the statistics in the original classification (Staroń *et al.*, 2009). Permutation tests on subgroups showed that the average k-tuple distance is significantly lower (two-tailed Student’s t-test; p-value  $< 1e-16$ ) in ECF subgroups as compared to random clusters of the same size distribution, indicating that subgroups are a well-defined entity (Fig. 3.4F).

Then, I computed a phylogenetic tree based on the consensus sequence of each subgroup. This tree helps to identify the evolutionary relationship between the ECF subgroups (Fig. 3.5). As outgroups I included sequences with a low-scoring  $\sigma_3$  domain, as well as the consensus sequence of all  $\sigma_3$ -containing proteins in Pfam, the latter of which I used as root of the tree. Not surprisingly, proteins with a low-scoring  $\sigma_3$  domain clustered at the base of the tree (Fig. 3.5) and formed three groups with sequence similarity to the sporulation  $\sigma$  factor SigF from Firmicutes and Actinobacteria, the flagellum biosynthesis  $\sigma$  factor FliA and the stationary phase  $\sigma$  factor SigH from *Bacillus* spp. Although they are not part of the ECF classification, these groups constitute the link between the group 3 and group 4  $\sigma^{70}$ s and account for the quality of this clustering approach. Other sequences with  $\sigma_3$  domain remained unclassified (0.18% of the unclassified sequences).

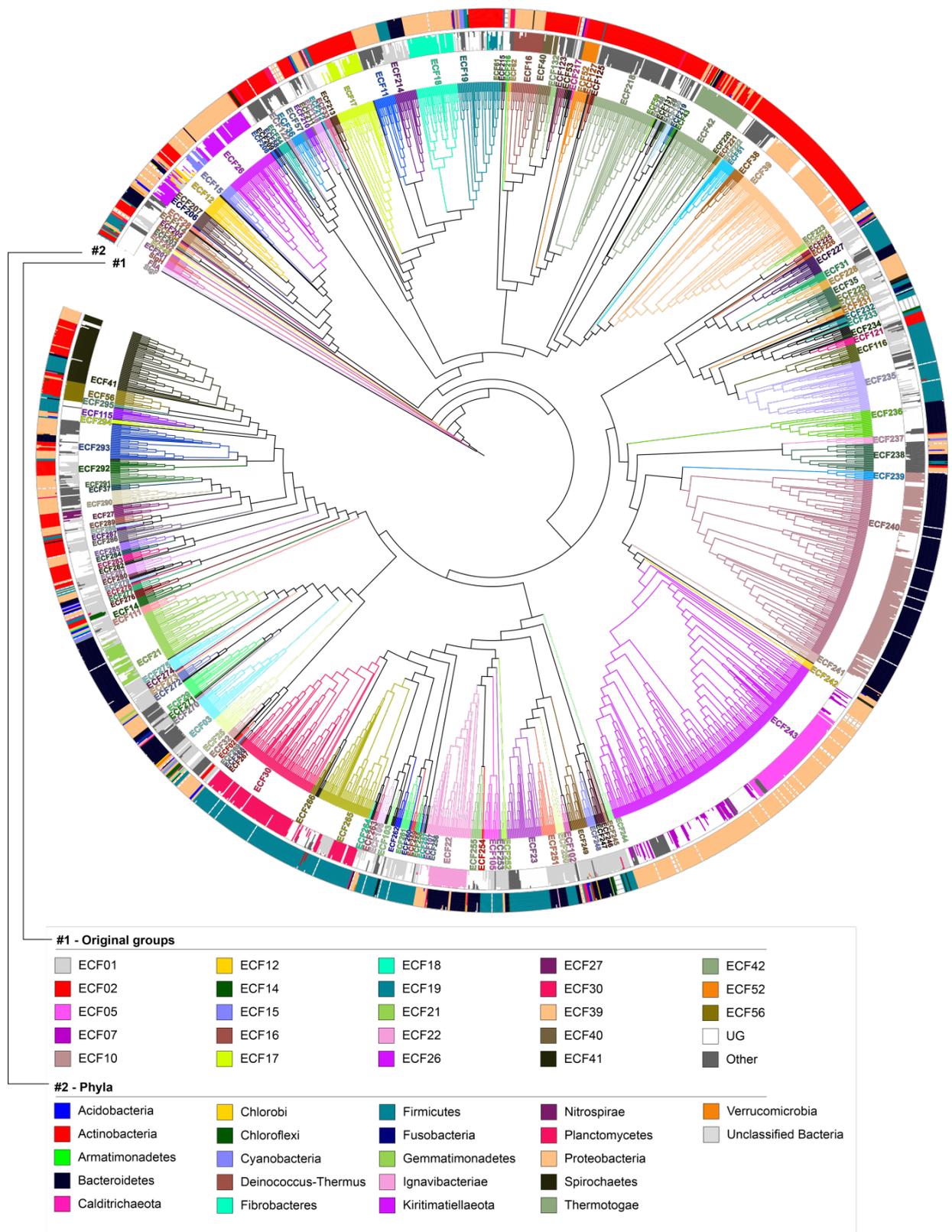


**Figure 3.4. ECF clustering pipeline.** **A:** The ECF clustering pipeline starts with non-redundant ECF  $\sigma$  factor sequences stripped to their  $\sigma_2$  and  $\sigma_4$  domains. These were clustered using MMSeqs2 and the resulting clusters were refined using bisecting K-means until the maximum intra-cluster distance was  $\leq 0.6$ . Subgroups with less than 10 sequences were not further considered. The consensus sequences of the resulting subgroups were hierarchically clustered, resulting in the ECF  $\sigma$  factor phylogenetic tree, which was used as the basis for the ECF group definition. **B:** Example of the resulting ECF tree for the clade composed of groups ECF267, ECF268, ECF269, ECF02 and ECF32. Leaves of the phylogenetic tree represent the consensus sequence of a subgroup. Every branch is associated to a bootstrap value. High bootstrap values are usually present in branches that define groups. The presence of shared conserved protein domain architectures in the genetic neighborhoods of subgroups that form monophyletic clades was used as a criterion for the ECF group definition. In this figure, domain architectures were considered conserved when they were present in more than 50% of the proteins of a certain position. The description of the conserved domain architectures can be found as legend, where “:” indicates that two domains are present in the same protein. Only genetic neighborhoods from organisms labeled as “representative” or “reference” by NCBI were considered for the calculation of genetic neighborhood conservation. The number of non-redundant ECFs and ECFs from “representative” and “reference” genomes is included as a column (N/N(rep./ref.)). Target promoter motifs were predicted for subgroups as explained in Section 8.6. Subgroups with non-self-regulated ECFs do not feature a conserved promoter motif. **C:** Example analysis of group ECF02. The bar plot shows the position-dependent frequency of domain architectures in the genetic context of members of ECF02 from “representative” or “reference” organisms (N=832). Only domain architectures that appear in more than 20% of the proteins encoded in a certain position are shown. Note that this architecture frequency might be underestimated due to the presence of higher scoring overlapping domains that interfere with the automatic

domain identification. The predicted target promoter motif for ECF02 is shown and has been confirmed for several members of ECF02 (Rhodius and Mutalik, 2010; Barchinger *et al.*, 2016). **D**: ECF group and subgroup size distribution, represented as box-plot. Size is expressed as the number of non-redundant proteins. **E**: Bootstrap value distribution in branches that define groups compared to branches that do not define groups. Bootstrap values tend to be larger in the former. **F**: Permutation validation of ECF subgroups. Average k-tuple distance for ECF subgroups and 100 sets of randomly generated clusters with the same size distribution as ECF subgroups. The difference in score distribution is statistically significant (Student's t-test p-value < 1e-16). **G**: Thumbnail of the average normalized bit-score of each ECF group (x-axis) against each HMM (y-axis). HMMs yield the highest HMMER bit score against their own ECF group (diagonal). Some cross-talk may occur between neighboring groups. Source: (Casas-Pastor *et al.*, 2019).

To identify subgroups with common characteristics, I performed an in-depth analysis of the genomic context of ECFs in each subgroup and aggregated subgroups into a total of 157 ECF groups. For the definition of these ECF groups, the phylogenetic tree was manually split into monophyletic clades, unless clades shared a similar genetic context and putative anti- $\sigma$  factor type (Fig. 3.4B). Genetic neighborhood and anti- $\sigma$  factor type were evaluated in the subset of ECFs from “representative” or “reference” genomes, as defined by NCBI (<https://www.ncbi.nlm.nih.gov>), including only RefSeq genomes when both RefSeq and GenBank assemblies exist. This helps to mitigate the bias towards commonly sequenced bacteria. As a result, 76.0% of the ECFs were captured in groups, displaying a median group size of 243 non-redundant proteins (Fig. 3.4D). As an example, figure 3B shows a close-up view on 19 ECF subgroups within the ECF tree, together with the proteins in their genetic neighborhood that feature >50% domain architecture conservation (i.e., a combination of their Pfam domains). Here, it is evident that ECFs in subgroups ECF02s1, ECF02s2 and ECF02s3 share a conserved genomic context with the anti- $\sigma$  factor RseA, and the regulators RseB and RseC for the former two subgroups, suggesting that ECFs in these subgroups feature the same mode of regulation as RpoE from *E. coli* (belonging to ECF02s1) (see Section 1.4.1.1). Likewise, the subgroups aggregated into group ECF32 display strong conservation with a two-component system (2CS) and a large number of genes encoding a type III secretion system (T3SS) (Fig. 3.4B). These results underline the previous notion that ECFs with close phylogenetic distance often share a conserved genomic context, the gene products of which are typically involved in the regulation of ECF activity and/or direct transcriptional targets of the ECF (Staroń *et al.*, 2009). This not only provides the basis for the definition of an ECF group, but also helps to predict putative functions and regulatory mechanisms to ECF groups with no experimentally described members (Table S3.1).

To provide a systematic overview on the conserved genomic context in each ECF group, I analyzed the frequencies of genes with a conserved protein domain architecture encoded up- and downstream of the ECF (Fig. 3.4C). For group ECF02, for instance, this revealed that downstream of the regulators RseA-C there is an enrichment of genes encoding translation regulators (e.g. EF-Tu), even though the specific position of individual genes is less conserved (Fig. 3.4C). However, despite the overall conservation of the genomic context within an ECF group, I often find subgroup-specific traits with respect to the positioning and the specific type of conserved genes, indicating that the definition of ECF subgroups is highly relevant to the biological function of an ECF  $\sigma$  factor.



**Figure 3.5. ECF  $\sigma$  factor tree.** Phylogenetic tree of the consensus sequences of ECF subgroups. Clades are colored and named according to their group. Ring #1 shows the assignment of proteins from each subgroup into the ECF groups from the original classification. Original ECF groups with less than 1% sequences are shown under "Other". Ring #2 shows the phylogenetic origin of the ECFs in a given subgroup. Source: (Casas-Pastor *et al.*, 2019).

In addition to the conserved genomic context, ECFs often auto-regulate the expression of their own genes, allowing bioinformatic prediction of their putative (sub)group-specific target promoters from conserved bi-partite DNA elements upstream of the ECF-encoding operon (Staroń *et al.*, 2009; Rhodius *et al.*, 2013). When applying a similar analysis to ECF subgroups and groups (Fig. 3.4B and C), I found overrepresented promoter motifs in many groups, e.g. ECF02, while others did not show significant motifs, e.g. ECF32, consistent with observations that the latter are not auto-regulated (Nizan-Koren *et al.*, 2003) (Table S3.1). Interestingly, even though predicted target promoter motifs were not used in the definition of the ECF groups, split points that define ECF groups (based on conserved genomic context) usually agree well with similar promoter elements (Fig. 3.4B). However, as for the conservation of the genomic context, I sometimes find subgroup-specific putative target promoters (e.g. in group ECF30 (information not shown)), highlighting the added value of the fine-grained clustering approach taken here.

The definition of ECF groups based on genomic context conservation is further supported by the statistical properties of the ECF subgroup tree, which typically displays high bootstrap support scores at the rooting branches of ECF groups (Fig. 3.4B and E), indicating that these groups are robust with respect to re-sampling of the original data set. To further check the performance of the new classification approach, I created profile Hidden Markov models (HMMs) from the conserved  $\sigma_2$  and  $\sigma_4$  domains of all sequences at the ECF group and subgroup level and tested whether these models were capable of faithfully classifying ECF sequences from their own groups (Fig. 3.4G). This analysis showed that sequences were assigned to the correct ECF group in 99.3% of cases, while assignment to the correct subgroup was successful in 94% of the cases. The lower performance of subgroup assignment was not surprising, given that neighboring ECF subgroups share higher sequence similarity than neighboring ECF groups. These results confirm that the definition of ECF groups and subgroups is based on a rational statistical approach and my tests support that their HMMs are specific and sensitive to allow for the classification of novel ECF  $\sigma$  factors, as later discussed in Section 3.5.

### 3.3. The ECF classification 2.0 refines original and identifies novel ECF groups

As a proof of concept, I compared the original ECF classification and the classification introduced in this work. To this end, I created HMMs from the original ECF groups and used them to classify the new ECFs gathered here. I saw that many of the sequences that classified into a particular new group were also classified into the same original group (Fig. 3.5, ring #1), indicating that there is a broad degree of correlation between the different classification approaches. Accordingly, for these groups of high coherence I maintained the original group names to label ECF groups presented in this work. Further in-depth analysis of the composition of the new groups revealed that 62 out of the 94 original groups are preserved, 21 are merged into larger groups, five remain mainly ungrouped, three are

scattered across several subgroups, and three are present only in small percentages in some groups (Table 3.1 and Fig. 3.6).

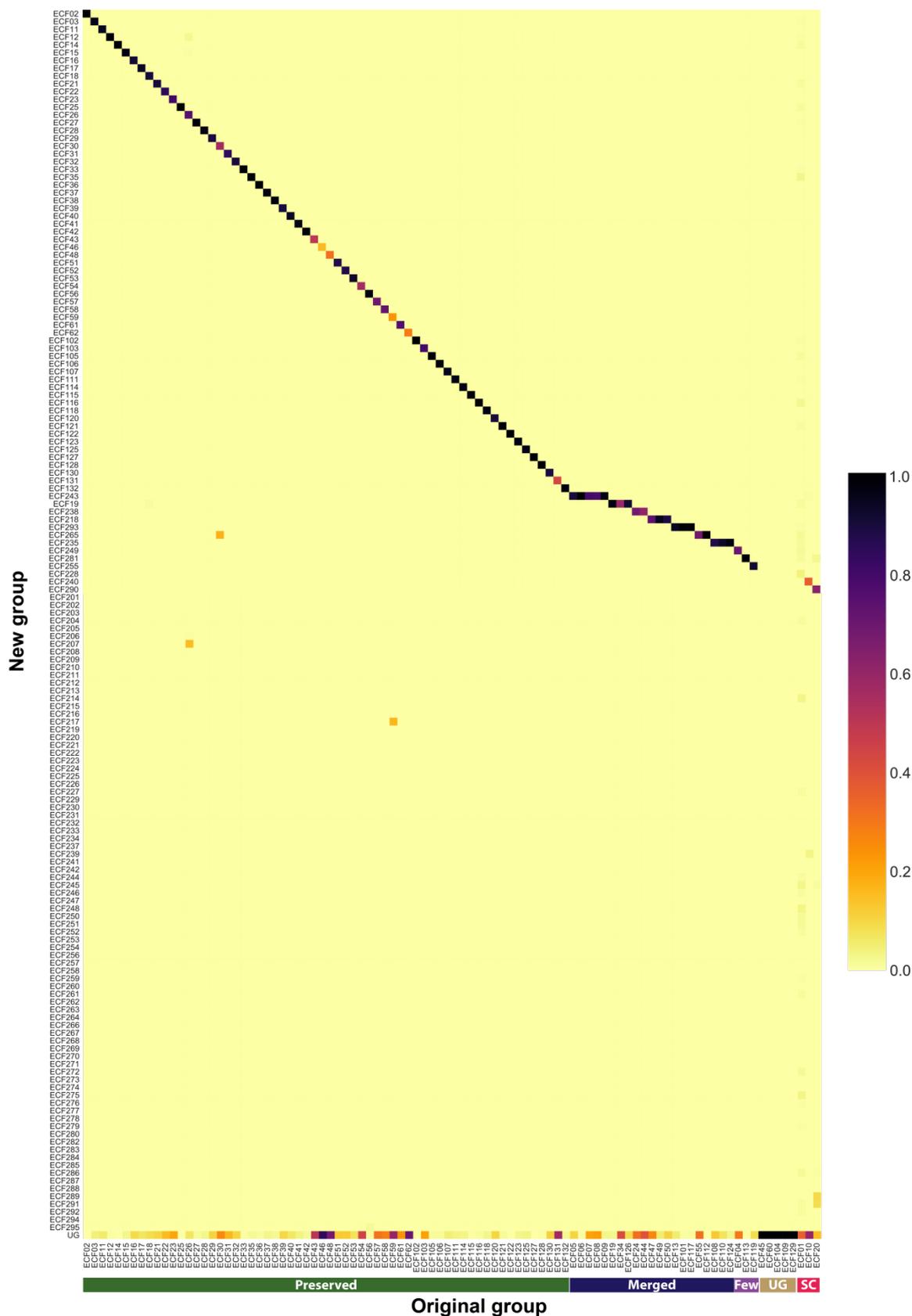
**Table 3.1. Rearrangements of original ECF groups.** Equivalence between original and new groups. Further information supporting this table can be found in figure 3.6. Original ECF groups can either be preserved (not shown in this table), merged, present in the new classification but composing a small percentage of the destination group, ungrouped in the new classification, or scattered across different new ECF groups. Source: (Casas-Pastor *et al.*, 2019).

Original ECF groups	New ECF groups
<b>Merged (21)</b>	
ECF05:ECF06:ECF07:ECF08:ECF09	ECF243
ECF13:ECF101:ECF117	ECF293
ECF19:ECF34:ECF126	ECF19
ECF24:ECF44	ECF238
ECF55:ECF112	ECF265
ECF47:ECF49:ECF50	ECF218
ECF108:ECF110:ECF124	ECF235
<b>Small percentages (3)</b>	
ECF04	ECF249
ECF113	ECF281
ECF119	ECF255
<b>Ungrouped (5)</b>	
ECF45	None
ECF60	None
ECF104	None
ECF109	None
ECF129	None
<b>Scattered (3)</b>	
ECF01	many (Fig. 3.6)
ECF10	many (Fig. 3.6)
ECF20	many (Fig. 3.6)

One case of an extremely scattered original groups is ECF01 (Fig. 3.6). This group was already considered highly diverse in the first ECF classification (Staroń *et al.*, 2009) and, based on the relatively unspecific HMM model of this group, it acquired more sequences in subsequent classification efforts (Jogler *et al.*, 2012; Huang *et al.*, 2015b). As a result, I did not consider the proteins from ECF01 for the nomenclature of the ECF groups in this work. Another highly scattered original group is ECF20 (Fig. 3.6). ECF20 is present in four main groups of this classification: ECF281, ECF289, ECF290 and ECF291 (Table S3.1). ECF281, ECF290 and ECF291 seem to be related to heavy-metal stress, since their genetic neighborhoods contain a conserved heavy-metal resistance protein in position +2 downstream of the ECF-encoding sequence in ECF281 and ECF290, and the full operon of a metal efflux pump in ECF291. This function of ECF291 has been experimentally confirmed for CnrH in *Cupriavidus metallidurans* (ECF291s9) (Grass, Fricke and Nies, 2005). Nevertheless, the anti- $\sigma$  factors encoded in the genetic context of members of these groups differ. ECF281 features a zinc finger-containing anti- $\sigma$  factor in position +1, while in the case

of ECF289 this protein contains a DUF3520 domain fused to a von Willebrand factor domain; ECF290 contains a RskA-like anti- $\sigma$  factor, and, lastly, ECF291 contains a CnrY-like anti- $\sigma$  factor in position -2 (Table S3.1). Based on this anti- $\sigma$  factor diversity, it seems likely that their cognate ECFs are regulated in response to different input stimuli, thereby warranting the definition of different ECF groups.

The last scattered group is ECF10. Even though minor parts of the original group ECF10 appear across the new ECF classification, groups ECF239 and ECF240 receive most of the proteins of the original ECF10. Although these two groups are both located within the large clade of FecI-like ECFs (Fig. 3.5), members of ECF239 do not contain genes with a conserved carbohydrate-binding domain in their neighborhood, a characteristic described for members of the original ECF10 (Staroń *et al.*, 2009). These conserved elements are part of polysaccharide utilization loci (PULs). PULs encode “starch utilization system” (Sus)-like systems, associated to utilization of the host glycans for the synthesis of capsular polysaccharides in saccharolytic Bacteroidetes from the gut microbiota, such as *Bacteroides thetaiotaomicron* (Martens, Roth, *et al.*, 2009).



**Figure 3.6. Agreement between original and new ECF groups.** This heatmap shows the frequency of proteins in each original group (x-axis) associated to each new group (y-axis). Original groups could be: preserved (“Preserved”), if there is a new group that contains most of their elements; merged (“Merged”): when several original groups are merged into a single new group; providing a low percentage of the proteins in their new group (“Few”); ungrouped (“UG”); or scattered (“SC”) across several groups of the new classification. New groups are named after their original group when their characteristics are preserved. Source: (Casas-Pastor *et al.*, 2019).

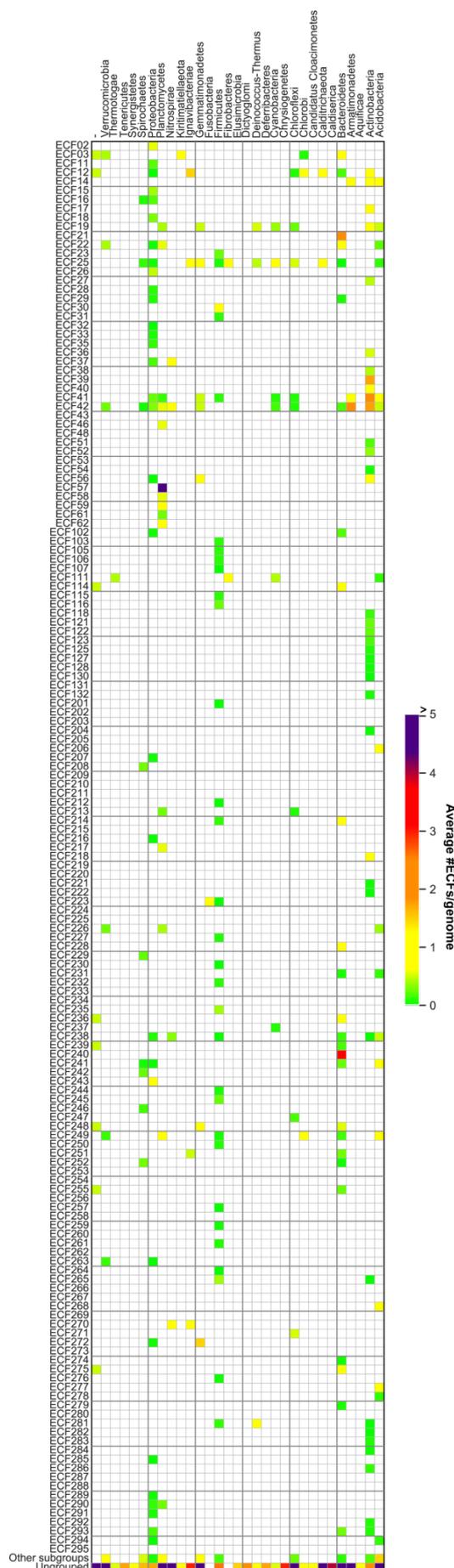
Even though scattered original groups are interesting, group merging events are more common. Their high occurrence is probably due to the incorporation of new protein sequences that bridge previously isolated ECF groups. Indeed, this possibility was considered in the founding ECF classification (Staroń *et al.*, 2009). One example of a merged group is ECF243, which constitutes the largest group of the new classification and contains proteins previously associated to original FecI-like groups ECF05 to ECF09 (Fig. 3.6). The reasons for merging these original groups were that 1) members of these original groups form a monophyletic clade in the ECF tree (Fig. 3.5) and 2) they contain a common genetic neighborhood with a FecI-like anti- $\sigma$  factor typically in position +1 from their coding sequence and a TonB-dependent receptor in position +2 (Table S3.1). Another example of new group product of merging is ECF238, which contains sequences from the original groups ECF24 and ECF44 (Fig. 3.6, Table S3.1). Members of ECF238 contain a cysteine-rich C-terminal extension of approximately 20 amino acids (Table S3.1), which is likely required for the activation of members of ECF238 when the appropriate metal in the right redox state is present in the cytoplasm, as found for CorE2 from *Myxococcus xanthus* (ECF238s15) (Marcos-Torres *et al.*, 2016).

**Table 3.2. Description of 22 new groups that have 0% sequences with similarity to any original group.** The table shows the number of non-redundant ECFs (N column), the number of ECFs from organisms tagged as “representative” or “reference” in NCBI (<https://www.ncbi.nlm.nih.gov>), excluding GenBank assemblies when an equivalent RefSeq assembly exists (N (rep/ref) column), their taxonomic origin, their putative regulator and other special traits. The taxonomic origin of groups where no representative/reference members are available is marked with ‘-’. Groups where regulators were not identified are marked with ‘-’. Source: (Casas-Pastor *et al.*, 2019).

ECF group	N	N (rep/ref)	Taxonomic origin	Regulator	Special traits
ECF201	69	13	Firmicutes (100%)	-	Closest ECF group to type III $\sigma$ factors
ECF202	35	0	-	-	
ECF208	79	25	Spirochaetes (100%)	Putative anti- $\sigma$ factor	Associated to glycosyl transferases fused to IDEAL domains
ECF210	139	5	Proteobacteria (100%)	-	
ECF215	49	0	-	-	
ECF216	46	13	Proteobacteria (100%)	Putative anti- $\sigma$ factor	
ECF219	88	20	Actinobacteria (100%)	Putative anti- $\sigma$ factor	Lack of $\sigma_{2.1}$ region in some subgroups
ECF220	55	11	Proteobacteria (100%)	C-terminal extension	Transmembrane proteins in +1 and -1
ECF221	243	50	Actinobacteria (100%)	Putative anti- $\sigma$ factor	
ECF222	46	14	Actinobacteria (100%)	Putative anti- $\sigma$ factor	
ECF229	102	19	Spirochaetes (100%)	Putative anti- $\sigma$ factor	Associated to proton-conducting membrane transporters
ECF234	43	4	Firmicutes (100%)	-	
ECF241	855	144	Bacteroidetes (68.28%), Proteobacteria (24.14%), Acidobacteria (6.21%), Spirochaetes (0.69%)	Putative FecR-like anti- $\sigma$ factor located C-terminally from a heavy-metal resistance protein	Located in the FecR clade
ECF242	147	42	Proteobacteria (44.19%) and Spirochaetes (55.81%)	Putative FecR-like anti- $\sigma$ factor	Associated to TonB-dependent receptors, except in proteins from Spirochaetes. Located in the FecR clade
ECF254	31	9	Firmicutes (100%)	-	-

ECF258	77	25	Firmicutes (100%)	DUF4179 -containing anti- $\sigma$ factor	Associated to ABC transporters
ECF267	28	6	Proteobacteria (100%)	-	
ECF280	44	14	Proteobacteria (100%)	Putative anti- $\sigma$ factor	Broad genetic context conservation
ECF282	128	28	Actinobacteria (100%)	Transcriptional regulation and perhaps ClpXP proteolysis (Seipke <i>et al.</i> , 2014)	
ECF287	55	18	Actinobacteria (100%)	Cys-rich C-terminal extension	
ECF288	74	32	Firmicutes (100%)	Cys-rich C-terminal extension	Associated to DUF2461 in +1
ECF294	300	52	Proteobacteria (96.15%), Acidobacteria (3.95%)	SnoaL-like C-terminal extension	

What is likely the most interesting contribution of the new classification are the entirely new groups. I found 22 new groups that could not be assigned to any original group (Table 3.2). Six of these groups contain less than ten proteins from representative/reference organisms and, therefore, their genetic neighborhood conservation was not further analyzed. From the remaining groups, 10 share a conserved genetic neighborhood with putative anti- $\sigma$  factors. A special case of these is ECF241, which is part of the FecI-like clade and represents an evolutionary intermediate between members ECF240, derived from original ECF10 and related to glycan utilization, and the iron uptake FecI-like group ECF242 and ECF243. Nevertheless, ECF241 shows no sequence similarity to any original FecI-like group. Instead of the canonical FecR-like anti- $\sigma$  factor from FecI-like groups, members of ECF241 contain a conserved two-transmembrane helix protein in position +1 or -1 from their coding sequence that in some cases hits the Pfam model for heavy-metal resistance proteins (Pfam: PF13801). Given the lack of an anti- $\sigma$  factor in a group within the FecI-like clade, I further analyzed this conserved protein. Its N-terminus, the region that typically contains the anti- $\sigma$  domain (ASD) in anti- $\sigma$  factors, is not long enough (usually less than 20 amino acids) to feature a typical ASD. However, a multiple-sequence alignment of these proteins, including the ASDs of canonical FecR-like anti- $\sigma$  factors, revealed that a putative, divergent ASD might be located in the C-terminal cytoplasmic part of the conserved protein. To my knowledge, this is the first time an anti- $\sigma$  domain has been predicted C-terminally from transmembrane helices. The second most common regulators of ECF activity in these new ECF groups are C-terminal extensions (four out of 22), with groups ECF287 and ECF288, from Actinobacteria and Firmicutes, respectively, containing cysteine-rich C-terminal extensions, and group ECF294 with a SnoaL-like extension (Table S3.1). A potential regulator was not found for members of ECF201 and ECF282. In the case of ECF282, the regulation could be carried out by a novel mechanism that involves transcriptional regulation and ClpXP proteolysis, as explained in Section 3.4.



Taken together, the ECF groups presented in this work preserve many of the original groups, expanding them with more proteins, and splitting or merging them in some cases. Here, I described the new findings concerning the 22 new ECF groups with no sequence similarity to any original group. However, a full overview of all the ECF groups and their occurrence in different bacterial phyla is shown in Fig. 3.7 and the summarized description of the groups is available in Table S3.1.

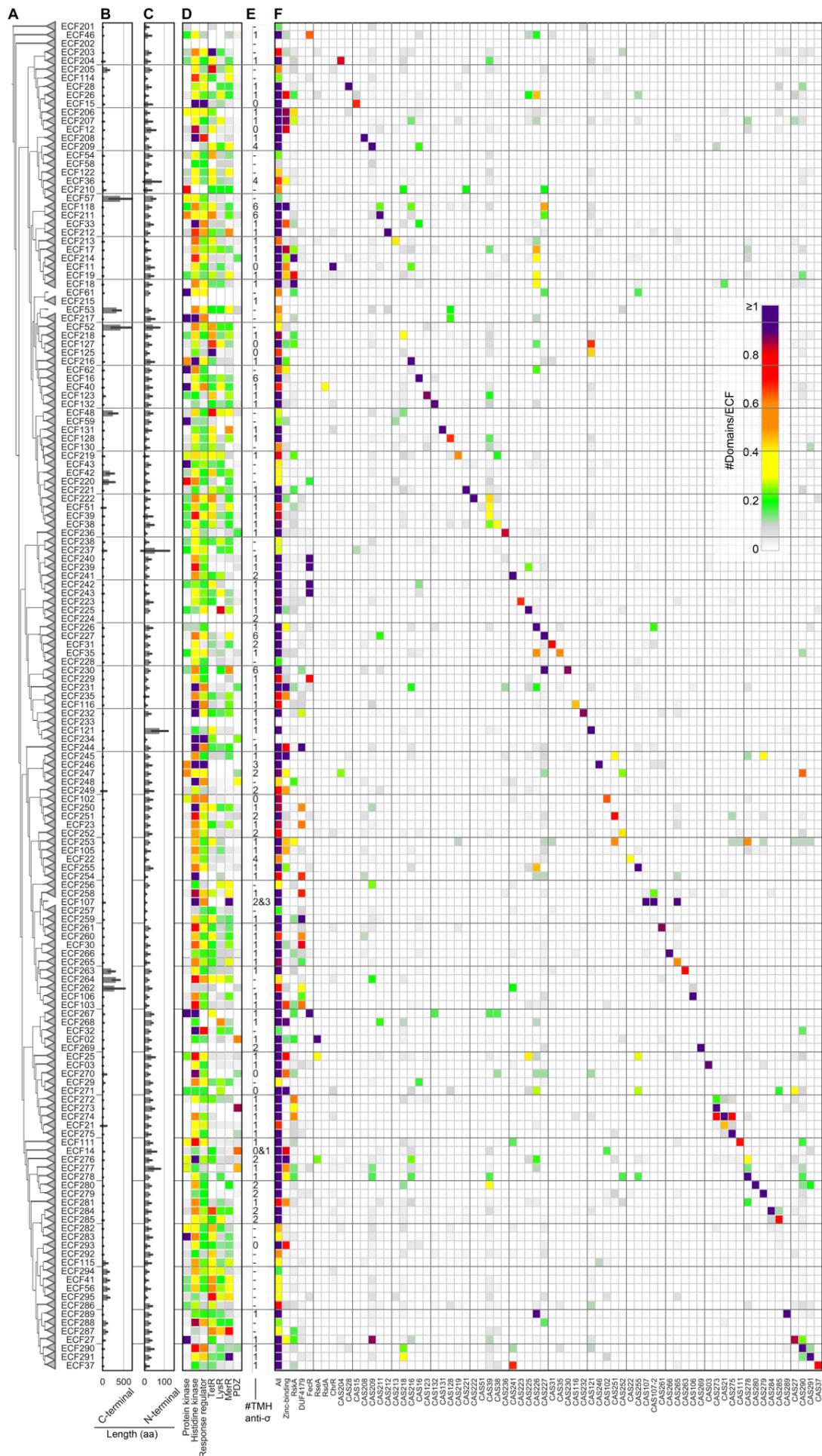
**Figure 3.7. ECF abundance in different phyla.** The heatmap shows the average number of ECFs from a certain ECF group in a certain phylum. ECFs grouped against subgroups that are not part of any group (“Other subgroups”) and ECFs that remain ungrouped (“Ungrouped”) are also shown. Underrepresented phyla are rich in the latter category. These values were calculated using the set of ECFs present in complete, non-metagenomic genomes from “reference” and “representative” organisms, selecting only RefSeq assemblies when both RefSeq and GenBank are available for the same organism. Organisms not assigned to any phyla are represented by “-“. Source: (Casas-Pastor *et al.*, 2019).

### 3.4. ECF $\sigma$ factors feature diverse, often multi-layered, modes of regulation

Given the large diversity of the ECF  $\sigma$  factor family, it is essential to focus on individual groups to extract hypotheses concerning their biological function, regulation and DNA binding site. Genetic neighborhoods of ECF  $\sigma$  factors typically contain an anti- $\sigma$  factor with a single transmembrane helix, encoded in position +1 downstream of the ECF coding sequence. However, it is well known that other regulatory elements might be substituting it, ranging from fused C-terminal extensions, to two-component systems and STKs (Staroń *et al.*, 2009; Mascher, 2013) (Section 1.4). Here, I provide an overview of the different modes of regulation present across the groups in the present ECF classification. A comprehensive description of all ECF groups can be found in Table S3.1.

Most of the ECF groups (114 out of 157) contain a putative anti- $\sigma$  factor, as defined by 1) the presence of Pfam domains of known ASDs, 2) sequence similarity to

anti- $\sigma$  factors of the founding classification (Staroń *et al.*, 2009) and 3) presence of transmembrane helices. Anti- $\sigma$  factors are typically encoded in position +1 from the ECF coding sequence. In most of the cases, the putative anti- $\sigma$  factor does not match any Pfam domain of experimentally addressed anti- $\sigma$  factors. In order to decipher common types of anti- $\sigma$  factors present across the ECF tree, I built HMMs from the cytoplasmic area (when it harbors a conserved region that could be a putative ASD) of the extracted putative anti- $\sigma$  factors that did not fit any Pfam domain. With these models, I searched the proteins encoded by 10 genes up- and downstream of the ECF coding sequence in all ECF groups (Fig. 3.8F). Interestingly, I found that most of the putative anti- $\sigma$  factors are ECF group-specific, in agreement with previous experimental observations showing orthogonality between anti- $\sigma$  factors of different groups (Rhodius *et al.*, 2013). However, exceptions are the clade that contains ECF222, ECF51, ECF38 and ECF39, which share the same type of one-transmembrane helix anti- $\sigma$  factor, groups regulated by FecR-like anti- $\sigma$  factors, mostly located in the FecI-like clade of the ECF tree (ECF239, ECF240, ECF242 and ECF243), groups regulated by RskA-like anti- $\sigma$  factors, anti- $\sigma$  factors with a putative zinc-finger, and anti- $\sigma$  factors with a DUF4179, almost exclusively present in Firmicutes (Fig. 3.8F). The number of transmembrane helices on putative anti- $\sigma$  factors is usually one (82 ECF groups), followed by two (14 groups), six (five groups), four (three groups) and three (one group) (Fig. 3.8E). Soluble anti- $\sigma$  factors, as defined by the absence of a predicted TM helix, are present in ten ECF groups (Fig. 3.8E). However, since this analysis can only identify soluble anti- $\sigma$  factors with sequence similarity to existing anti- $\sigma$  factors, it is likely that other soluble anti- $\sigma$  factor variants exist. Additionally, even though I evaluated the similarity of new ECF  $\sigma$  factors to known ASDs, it is not guaranteed that all the new putative anti- $\sigma$  factors function as such, given the vast diversity, lack of sequence conservation and lack of studies confirming their function.



**Figure 3.8. Genetic context analysis of ECF groups.** **A:** Schematic representation of the ECF  $\sigma$  factor tree. **B, C:** Bar plot with the average number of amino acids after the end of  $\sigma_4$  domain (C-terminus) or before  $\sigma_2$  domain (N-terminus), respectively. Error bars indicate standard deviation. **D:** Average number of regulatory domains in genetic neighborhoods per ECF. **E:** Number of predicted transmembrane helices of the putative anti- $\sigma$  factor encoded in the genetic neighborhood of groups. **F:** Average number of anti- $\sigma$  factor domains per ECF, predicted in the genetic neighborhood of members of an ECF group. Source: (Casas-Pastor *et al.*, 2019).

ECF107 contains two putative anti- $\sigma$  factors, which could be part of the same protein complex or compete for binding the ECF (Pinto and Mascher, 2016), thereby illustrating the complexity and diversity of anti- $\sigma$  factor mediated regulation. A second example is ECF102, whose only described member, SigX from *P. aeruginosa*, has been suggested to have a role in mechanosensing (Chevalier *et al.*, 2019). SigX is part of a seven-gene operon which includes a mechanosensitive ion channel (CmpX) encoded in position -1, a putative anti- $\sigma$  factor (CfrX) encoded in position -2 and an outer membrane porin (OprF) encoded in position +1 (Chevalier *et al.*, 2019). Even though original reports hypothesized that the regulation of SigX is carried out by the putative anti- $\sigma$  factor CfrX (Pinto and Mascher, 2016), new reports suggest that its regulatory mechanism is more complex and involves also CmpX and OprF (Chevalier *et al.*, 2019). I observed that these proteins are conserved in ECF102s1. Moreover, the mechanosensitive ion channel is conserved in subgroups 2 and 5, which indicates a similar regulation as members of subgroup 1. A similar case, in which proteins in addition to the anti- $\sigma$  factor are required for ECF regulation, is ECF31. The only characterized member of ECF31, SigY from *B. subtilis* (subgroup 1), requires both the anti- $\sigma$  factor YxlC, encoded in +1, and YxlD, encoded in +2, for its regulation, presumably forming a protein complex with the ECF (Yoshimura *et al.*, 2004). YxlCD homologs are conserved across ECF31.

The second most common regulatory mode of ECF  $\sigma$  factors is the presence of C-terminal protein extensions to the ECF (19 groups) (Fig. 3.8B), which is typically correlated with the lack of putative anti- $\sigma$  factors (Fig. 3.8F). This agrees with the idea that the extension is substituting the anti- $\sigma$  factor in the regulation of the ECF (Pinto, Liu and Mascher, 2019). ECFs with the same type of C-terminal extension cluster together in the same group (i.e., members of ECF42, with tetratricopeptide repeats in their extension), or in neighboring groups (i.e. members of ECF41, ECF56, ECF294, and ECF295, with SnoaL-like C-terminal extensions). Given that only the core ECF domains were inputs of the ECF classification process, this supports the notion that the extension interacts with the core ECF regions in a unique manner depending on the type of domain that it bears (Wu *et al.*, 2019). An interesting exception is ECF205, which also has a SnoaL-like extension but is in proximity to the root of the ECF tree (Fig. 3.5), indicating that more factors, in addition to its C-terminal extension, determine the sequence conservation of this group. Aside from the Pfam domains identified in C-terminal extensions of the founding ECF classification (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b), I identified a domain of unknown function (DUF1835) in ECF264, an extension with five or seven transmembrane helices in ECF263, and a CGxxGxGxCxC motif in ECF288.

Canonical C-terminal extensions are usually longer than 50aa, but I found that some groups contain short C-terminal extensions difficult to identify when only looking at protein length. These groups

usually lack any other discernable means of regulation, which points towards the short extension as a modulator of ECF activity. One of these groups is ECF238, which merges the original groups ECF44 and ECF24. Members of ECF238 contain conserved cysteine residues in their short (~20aa) C-terminal extension and in the linker in some instances. One of the described members of ECF238, CorE2 from *Mycococcus xanthus* (subgroup 15), is known to be activated by  $\text{Cd}^{2+}$  and  $\text{Zn}^{2+}$  via this cysteine-rich C-terminal domain (Marcos-Torres *et al.*, 2016; Pérez, Muñoz-Dorado and Moraleda-Muñoz, 2018). Another group with a short C-terminal extension is ECF29, which contains a conserved RCE/D motif in its ~30 extra amino acids. Unfortunately, no member of ECF29 has been experimentally addressed, but the absence of a putative anti- $\sigma$  factor similarly suggests a regulatory role of this short extension.

N-terminal extensions of the ECF core regions occur less often, they are generally shorter than canonical C-terminal extensions and they are prone to be overlooked whenever the ECF is translated from non-canonical start codons. The only well-described N-terminal extension appears in ECF121 (Fig. 3.8C). This extension has been studied in BldN from *Streptomyces coelicolor* (subgroup 1), where it has to be proteolytically degraded to process the proprotein to its mature ECF, which then is subject to anti- $\sigma$  factor regulation (Bibb and Buttner, 2003). Nonetheless, subgroups from several groups contain N-terminal extensions (Table S3.1). For instance, in ECF36s4, represented by SigC from *M. tuberculosis*, the N-terminal extension has been proposed to inhibit the DNA contact in the uninduced state since members of this subgroup lack an obvious anti- $\sigma$  factor (Thakur, Joshi and Gopal, 2007). Alternatively, the N-terminal extension of two members of ECF12s1,  $\sigma^R$  from *S. coelicolor* and SigH from *Mycobacterium smegmatis*, generates an isoform translated from an earlier start codon, which is unstable upon exposure to thiol oxidants (Kim *et al.*, 2009). This makes  $\sigma^R$  susceptible to  $\sigma^R$ -activated ClpP1/P2 proteases and thus implements a negative feedback loop that contributes to turning off the stress response (Kim *et al.*, 2009).

Other putative regulators of  $\sigma$  factor activity often found in the conserved genetic neighborhood of ECFs were STKs (Fig. 3.8D). ECF  $\sigma$  factors of five original groups have been hypothesized to be directly phosphorylated by a protein kinase (ECF43 and ECF59-ECF62 (Staroń *et al.*, 2009; Jogler *et al.*, 2012)). This is also likely the case for EcfK from *Xanthomonas citri*, another member of ECF43 (Bayer-Santos *et al.*, 2018). I add to the list of protein kinase-associated groups ECF217, ECF267 and ECF283 (Fig. 3.8D). Other groups such as ECF40, ECF27 or ECF210 contain protein kinases only in certain subgroups. Proteins from original group ECF60 were not classified by the pipeline since only eight members of ECF60 were extracted. Moreover, some proteins that should belong to group ECF43 fail to be identified as ECF  $\sigma$  factor, reducing the size of this group. One reason could be the divergent  $\sigma_2$  domain observed in members of ECF60 and ECF43. Consequently, Section 5.4 will address the expansion of STK-associated groups. Protein kinase-related ECF groups typically lack co-encoded anti- $\sigma$  factors (Fig. 3.8D). The only exception is group ECF267, which contains a putative FecR-like anti- $\sigma$  factor with tetratricopeptide repeats in position +1. Given that ECFs from group

ECF267 are very distant from the FecI-like clade (ECF239-ECF243) (Fig. 3.5), it seems possible that this anti- $\sigma$  factor does not target members of ECF267, but other FecI-like ECFs. However, none of the organisms that contain members of ECF267 contain any FecI-like ECF  $\sigma$  factor. Whether the anti- $\sigma$  factor and/or the STK regulate the activity of members of ECF267 is unclear.

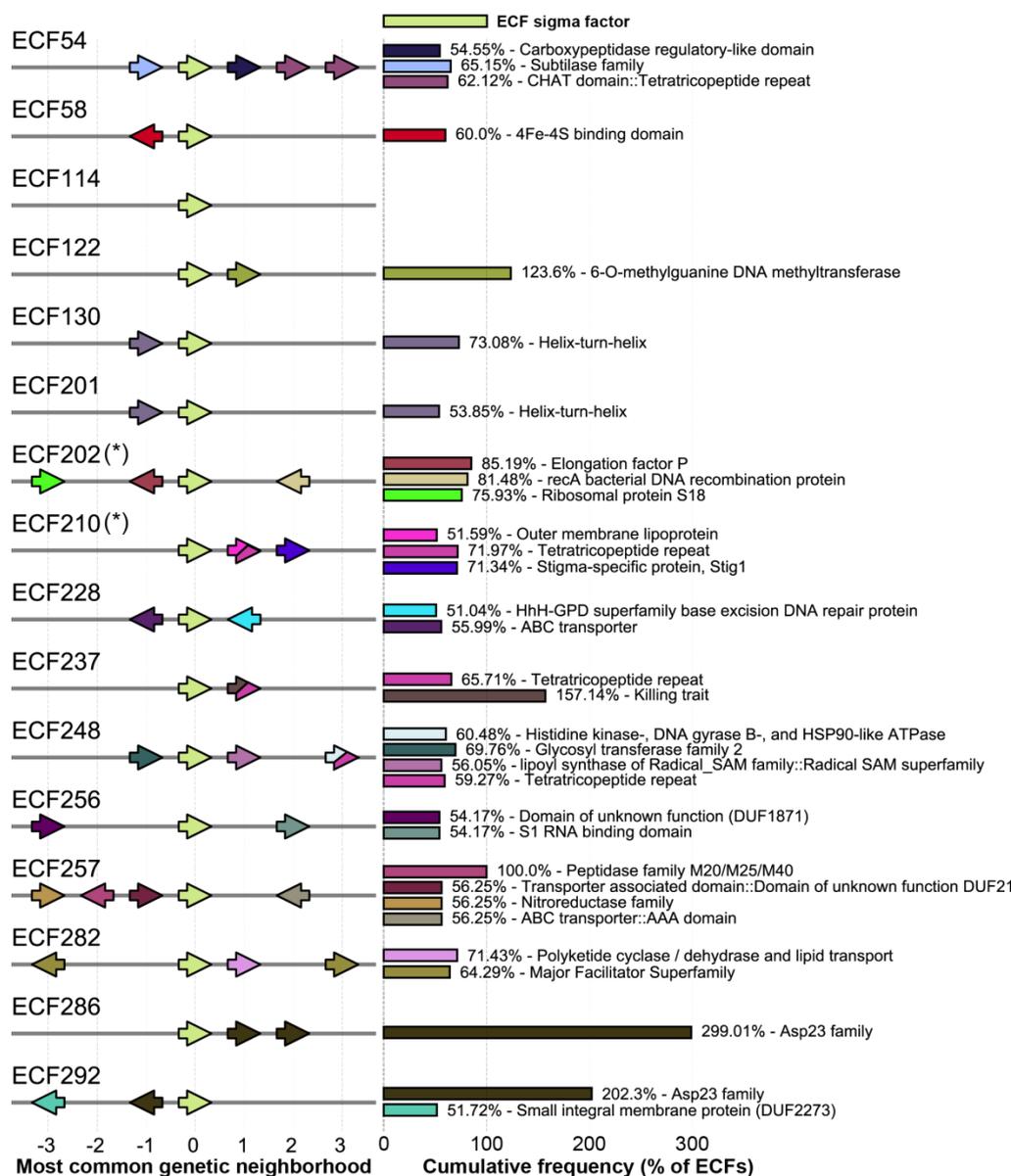
Four groups contain two-component systems in their genetic neighborhood. These regulators can co-occur in combination with anti- $\sigma$  factors, as in the case of ECF15 and ECF246, or not, as in ECF32, ECF234 and subgroups 1, 2 and 3 of ECF39. These possibilities reflect the different regulatory mechanisms exerted by two-component systems. On one hand, members of ECF15, the main general stress response  $\sigma$  factors in Alphaproteobacteria, are activated by a partner-switching mechanism, as described in Section 1.4.1.3. This is unlikely the case for members of ECF246, since their response regulator is fused to a transcriptional regulator instead of an ECF-like domain. Future analysis of members ECF246 could determine whether the putative anti- $\sigma$  factor and/or the two-component system encoded in their genetic context regulates ECF activity. For members of ECF32, it was indeed shown that the two-component system indirectly regulates transcription of the ECF  $\sigma$  factor by inducing the expression of the transcription factors HrpSR (Merighi *et al.*, 2003; Nizan-Koren *et al.*, 2003; Lan *et al.*, 2006). Members of ECF32 in turn activate the synthesis of the type III secretion system *hrp*, required for plant infection (Merighi *et al.*, 2003; Nizan-Koren *et al.*, 2003). In the case of ECF39, 2CSs are directly regulating the transcription of the ECF, as described for SigE from *S. coelicolor* and  $\sigma^{25}$  from *Streptomyces avermitilis* (Luo *et al.*, 2014; Tran *et al.*, 2019). Members of this group are involved in antibiotic synthesis and cell-wall stress resistance (Luo *et al.*, 2014; Tran *et al.*, 2019). This direct regulation of the 2CS over the ECF expression could also occur in members of ECF234, given the absence of a putative anti- $\sigma$  factor and the fusion of the response regulator to a transcriptional regulator. The physiological function of ECF234 seems to be related to an ABC transporter present in its genetic context, but the substrate of the transporter is unknown.

On top of these elements, I found that some ECF groups contain conserved transcriptional regulators in their genetic contexts, such as TetR-like repressors, which appear in groups with anti- $\sigma$  factors (ECF125) and, remarkably, in ECF203, which lacks any obvious regulator (Fig. 3.8D). Given the lack of characterized members of ECF203, it is unclear whether this TetR repressor regulates the expression of members of ECF203 or is part of their response. In favor of the former, members of ECF203 do not seem to be auto-regulated, since they lack a conserved (predicted) target promoter motif (Table S3.1). Other transcriptional regulators include LysR- and MerR-like repressors, which appear in several ECF groups associated with anti- $\sigma$  factors (Fig. 3.8D).

A total of 16 ECF groups are not linked to any of the above-mentioned regulators (Fig. 3.9), inspiring the prediction of novel, putative regulators of ECF activity. So far, only three of the 16 groups have experimentally addressed members, namely ECF228, ECF282 and ECF114. SigP from *Porphyromonas gingivalis* (ECF228s7) is only present in measurable concentrations when stabilized by direct interaction with the response regulator PorX from the two-component system PorXY

(Kadowaki *et al.*, 2016). Even though response regulators are not conserved in the genetic context of members of ECF228 (Fig. 3.9), it is possible that members of ECF228 are unstable and require other proteins, such as chaperons. In the case of the novel group ECF282,  $\sigma^{\text{AntA}}$  from *Streptomyces albus* (subgroup 2) is regulated at the level of transcription and might be target of ClpXP proteolysis (Seipke, Patrick and Hutchings, 2014). Indeed, homologs of  $\sigma^{\text{AntA}}$  have been considered a new group of ECF  $\sigma$  factors that control the expression of antimycins (Seipke, Patrick and Hutchings, 2014). Even though the C-terminal AA dipeptide, suggested as target of ClpXP proteolysis (Seipke, Patrick and Hutchings, 2014), is only present in members of subgroup 2, members of other ECF282 subgroups could be regulated in a similar manner, since different ClpXP proteases have different binding specificities (Balogh *et al.*, 2017). In ECF114, SigH from *Porphyromonas gingivalis* (subgroup 4) plays a role in aerotolerance. SigH it is induced upon exposure to O<sub>2</sub> and promotes oxidative stress protection and hemin uptake (Yanamandra *et al.*, 2012). Although it is speculated that SigH is transcriptionally activated (Yanamandra *et al.*, 2012), no transcription factor in charge of this task has been identified.

The lack of canonical ECF regulators, but the presence of other conserved elements (Fig. 3.9) in the genetic neighborhood of the remaining 13 groups prompted the generation of speculative hypotheses about the regulation of these groups. However, a general issue of this analysis is that it is hard to discriminate whether these elements are regulators and/or targets of ECF activity, suggesting that both options should be considered in downstream experimental analyses. Interestingly, I found new putative regulators/targets of regulation of the original groups ECF54 and ECF130. ECF54 is encoded near a protein with a 4Fe-4S cluster, whereas ECF130 is encoded in proximity to a helix-turn-helix (HTH) containing protein (Fig. 3.9). Since HTH motifs are usually related to DNA binding, this protein could be a new type of transcriptional factor involved in transcriptional control of members of ECF130. A similar case is found in the new group ECF201, which is usually co-encoded with HTH proteins in position -1 (Fig. 3.9). Another interesting case is found in members of ECF237, which contain several “killing trait” proteins (Pfam: PF11757) in their vicinity (Fig. 3.9). These domains were described for RebB, one of the three proteins necessary for the assembly of R-bodies in the Paramecium endosymbiont *Caedibacter taeniospiralis* (Heruth *et al.*, 1994). Given the absence of conservation for the rest of the proteins from the R-body and the presence of several copies of the killer domain in members of ECF237 (which are mainly present in Bacteroidetes unrelated to the Alphaproteobacterium *C. taeniospiralis*), it is possible that proteins with this domain have an alternative function, not related to R-body assembly, but potentially involved in controlling ECF activity. Lastly, members of ECF286 and ECF292 share genetic neighborhood with several copies of Asp23 proteins (Pfam: PF03780) (Fig. 3.9). Asp23 is one of the most abundant proteins of *Staphylococcus aureus* and its deletion leads to upregulation of the cell wall stress response (Müller *et al.*, 2014). Therefore, Asp23 proteins could be acting as a new type of anti- $\sigma$  factor that regulate the activity of members of ECF286 and ECF292.



(\*) Genetic neighborhood conservation extracted from the full dataset, including non-representative/reference organisms

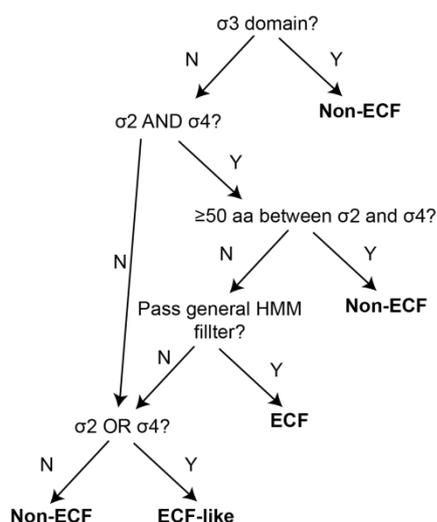
**Figure 3.9. Genetic neighborhood of ECF groups that lack a canonical regulator.** The left side shows the typical positions of genes encoding a certain protein domain architecture (present in > 50% of the genetic contexts). Only positions  $\pm 3$  from the ECF coding sequence are displayed. The direction of the arrow indicates the most common orientation of the coding sequence. The cumulative percentage of proteins with a certain domain architecture is shown on the right. Only proteins from reference and representative organisms, taking only RefSeq proteins when both RefSeq and GenBank assemblies exist for the same genome, are considered. Exceptions (marked with stars) are made for groups with less than 10 proteins present in these organisms, this is ECF202 and ECF210. Source: (Casas-Pastor *et al.*, 2019).

### 3.5. ECF hub, a public repository of the ECF classification 2.0

In collaboration with R. Müller and Prof. A. Goessmann from the Justus Universität Gießen, we set up a database that contains the information of the ECF classification 2.0, the so-called ECF hub. This web resource will be publicly available and will allow users to 1) browse the ECF classification and search for ECFs in any bacterial taxonomic level, 2) visualize the conservation of certain domains in groups and subgroups and their predicted target promoter motif, 3) download the HMM and raw sequences of any group and subgroup, as well as their predicted target promoter motif and any genetic

neighborhood figure, 4) access an overview of the information gathered about every ECF group, including protein characteristics, regulation and function of studied members, when available, or suggested by the conserved genetic neighborhood composition, and references to relevant literature. In addition to this functions, ECF hub provides researchers with tools for the analysis of their protein sequences, allowing them to decide whether their proteins are ECFs, ECF-like or non-ECF proteins, and classifying ECFs into ECF groups and subgroups.

For the decision on whether a protein is an ECF, I designed a pipeline based on the work by Staroń and colleges (Staroń *et al.*, 2009) in which proteins that do not contain neither  $\sigma_2$  nor  $\sigma_4$  domain are not considered ECFs. Therefore, only canonical  $\sigma^{70}$  proteins would pass this filter. Moreover, proteins with  $\sigma_3$  domain or with linkers of 50 amino acids or longer which can, hence, contain a cryptic  $\sigma_3$  domain, are not considered ECFs. This filter discards group 1, 2 and 3  $\sigma^{70}$  proteins. Instead, proteins with no  $\sigma_3$  domain that fail to contain both  $\sigma_2$  or  $\sigma_4$  domains are considered “ECF-like” proteins. This category includes proteins with ECF  $\sigma$  factor function but with divergent core domains, such as: 1) EcfP from *Vibrio parahaemolyticus*, which lacks of a canonical  $\sigma_2$  domain as discussed in Section 5, 2) SigI from *B. subtilis* and ComX from *Streptococcus pneumoniae*, which lack a canonical  $\sigma_4$  domain (Wei *et al.*, 2019), and 3)  $\sigma^1$ -like ECFs, which contain a  $\sigma_{1-C}$  domain instead of a canonical  $\sigma_4$  domain and are involved in the synthesis of components of the cellulosome in cellulolytic clostridia (Ortiz de Ora *et al.*, 2018). Aside from ECF  $\sigma$  factors, anti-anti- $\sigma$  factors that regulate members of group ECF15, such as PhyR from *Caulobacter crescentus*, fall also in the “ECF-like” category. Proteins that contain both  $\sigma_2$  and  $\sigma_4$  domain, lack  $\sigma_3$  domain (either canonical or cryptic) and score higher than the ROC-optimized threshold against the general ECF HMM (explained in Sections 8.2 and 3.1) are considered ECF  $\sigma$  factors (Fig. 3.10).



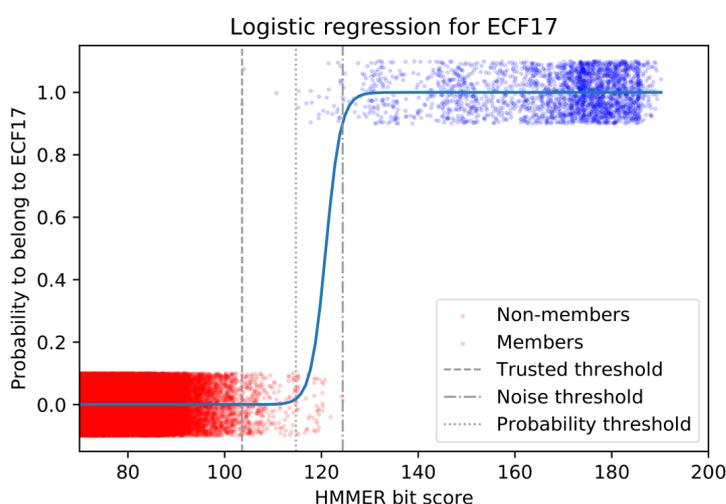
**Figure 3.10. Decision tree on whether a protein of interest is an ECF, ECF-like or a non-ECF.** The decision is taken based on the lack of  $\sigma_3$  domain or any cryptic  $\sigma_3$  domain, the presence of  $\sigma_2$  and  $\sigma_4$  domains, and a score against the general ECF HMM above the ROC-optimal threshold of 60.8 (see Section 8.2).

Another tool of ECF hub that I designed was the classification tool, where ECF  $\sigma$  factors are classified against groups and subgroups. For this, I derived two bit score cut-offs for each group and subgroup HMM, trusted and noise cut-offs. Trusted cut-offs are defined as the minimum bit score achieved by any true member of the group/subgroup under evaluation, whereas noise cut-offs are the maximum bit score achieved by any ECF that is not members of that group/subgroup. Trusted and noise cut-offs provide the first step in the evaluation of the membership of a protein against a group/subgroup. Only clusters with a score higher than the trusted cut-off, or the noise cut-off in cases where no cluster scores above its trusted cut-off, pass to the next step. Then, I obtained the probability

that a protein belongs to a cluster using a logistic fit, as described in (Brown, Krishnamurthy and Sjölander, 2007). First, I represented in the x-axis the bit score and in the y-axis the probability that a bit score is produced by a real member of the group/subgroup under evaluation, this is 1 for members and 0 for non-members. I fitted the data to a logistic regression describing the transition of the bit scores from non-members to members of a cluster (Fig. 3.11). The logistic function describing the probability  $P(q, i)$  of protein sequence  $q$  belonging to group/subgroup  $i$  reads

$$P(q, i) = \frac{1}{1 + e^{-k_i \times \frac{B_{qi} - \bar{B}_i}{\sigma(B_i)} - c_i}} \quad (\text{Eq. 3.1})$$

where  $B_{qi}$  is the bit score obtained for the protein sequence  $q$  scored against the HMM model of group/subgroup  $i$ , and  $\bar{B}_i$  and  $\sigma(B_i)$  being the mean and standard deviation of all bit scores obtained for members of group/subgroup  $i$ , respectively. The parameters  $k_i$  and  $c_i$  were fitted using non-linear least squares fit. The selection of a unique probability threshold for ECF groups/subgroups was carried out with a ROC curve, using the members of groups/subgroups as true positives and members of all other groups/subgroups as true negatives. This probability threshold, together with trusted and noise cut-offs, and fitting parameters of the logistic curve are available for groups, subgroups and original groups under request.



**Figure 3.11. Example logistic regression.** Logistic regression fitting the distribution of scores of members ( $y=1$ , blue) and non-members ( $y=0$ , red) of group ECF17. Proteins (represented as points) have been plotted  $\pm 0.1$  in the y-axis for clarity. X-axis has been capped at bit score=75 for clarity. Three thresholds are used for protein classification, namely, trusted threshold, i.e. bit score of the lowest scoring member, noise threshold, i.e. largest score of any non-member, and probability threshold, i.e. ROC-optimized probability that a protein belongs to a group. Source: (Casas-Pastor *et al.*, 2019).

This probability threshold is used in the decision of whether a protein belongs to a cluster. The probability that a protein belongs to a cluster is calculated only for clusters that pass the trusted/noise cut-off filter. The cluster with the largest probability is the one to which the protein is assigned, as long as its probability value is above the probability cut-off. Some small subgroups could not be fitted by the logistic regression. In that case, the probability is considered 1 if the protein has a bit score larger than the trusted cut-off and it does not match any other subgroup, or 0 otherwise.

Only protein sequences and HMMs stripped to their  $\sigma_2$  and  $\sigma_4$  domains were used for classification so as to avoid contribution of non-conserved regions such as N- and C-terminal extensions and linker. This strategy provided a slightly better accuracy than the classification of full-length proteins against full-length HMMs in the self-classification of subgroups, with 94.02% proteins correctly classified for stripped sequences versus 93.93% for full-length sequences.

### 3.6. Discussion and summary

The ECF retrieval and classification introduced in this section refines and greatly expands previous ECF classification efforts. Thanks to its two-tiered clustering approach, it provides a high-resolution view of the ECF family. ECF subgroups, composed of closely related proteins, were further hierarchically clustered into 157 ECF groups, defined based on a common genetic neighborhood, which indicates a similar mode of regulation. As part of the *in silico* characterization of ECF groups, I predicted their putative regulators, their target promoter motifs and their most likely function (Table S3.1). As already observed for the previous classification, these predictions are biologically meaningful in that they correctly reflect results of experimentally studied members, whenever available. The comprehensive description of the ECF groups serves as a source of testable hypotheses that will support the experimental description of new ECFs, which will lead, in turn, to more precise and detailed group descriptions. A comprehensive description of the ECF groups will be available on the web resource ECF hub.

The new ECF classification presented in this work has changes with respect to the original classification. Even though 62 of the 94 original groups were preserved, 21 were merged, five were ungrouped, and three each were scattered or present in the new classification but composing only small parts of their new group (Table 3.1). The new ECF groups are monophyletic clades of the ECF phylogenetic tree, which can be subdivided into hierarchically-distributed ECF subgroups. This high-resolution, comprehensive classification provides advantages with respect to partial updates. One example comes from ECF54 and ECF58, identified in two different works and in two phyla, Actinobacteria and Planctomycetes, respectively (Jogler *et al.*, 2012; Huang *et al.*, 2015b). Within the ECF tree, these two groups are direct neighbors with a bootstrap support value of 17, indicating a large protein similarity between them. None of them has a putative anti- $\sigma$  factor or any other clear regulator of their activity, and they contain different elements in their genetic context (Fig. 3.9). These results suggest that ECF58 and ECF54 have the same origin, but they evolved independently in Actinobacteria and Planctomycetes, acquiring different genes in their genetic neighborhood. What remains unclear is whether the regulation of members of ECF54 and ECF58 has common features, as expected for ECFs with a common origin.

As part of the description of ECF groups, I analyzed their most likely regulators and the types of putative anti- $\sigma$  factors encoded in their genetic neighborhood (Fig. 3.8). Most of the predicted anti- $\sigma$  factors are highly specific for their own groups (Fig. 3.8F). Exceptions occur in neighboring ECF

groups, e.g. in the FecI-like clade (ECF239 to ECF243) or in the clade formed by groups ECF214, ECF18 and ECF19, indicating co-evolution between ECF and anti- $\sigma$  factor sequences. However, the general lack of the same type of anti- $\sigma$  factors in neighboring groups reflects their large diversity and their specificity, which has been exploited for the construction of orthogonal genetic circuits (Rhodius *et al.*, 2013). Anti- $\sigma$  factors are not the only genes conserved in the genetic context of ECF  $\sigma$  factors. In this study, I identified the ECF groups associated to other known ECF regulators such as C-terminal and N-terminal extensions, two-component systems, STKs (Mascher, 2013), and other regulators such as TetR repressors (Fig. 3.8).

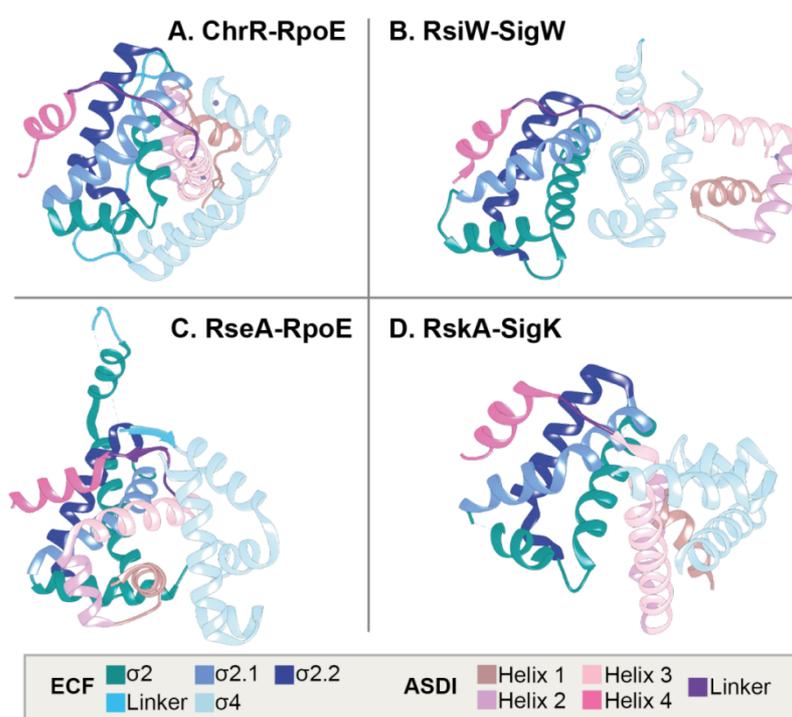
With an average of approx. 10 ECFs per genome, ECFs are more abundant than previously thought (Staroń *et al.*, 2009). Confirming previous reports (Staroń *et al.*, 2009; Han *et al.*, 2013; Huang *et al.*, 2015b), the number of ECFs is proportional to genome size (Fig. 3.3), with species thriving in diverse environments typically featuring larger genomes that provide them with the ability to sense and respond to a large variety of external signals. One example is the bacterium *Sorangium cellulosum* So0157-2, which features a genome that is more than 1Mbp larger than its close relative *S. cellulosum* So ce56, allowing the former to adapt to alkaline conditions (Han *et al.*, 2013). Accordingly, the number of ECFs in *S. cellulosum* So0157-2 (82 ECFs) is significantly larger than in *S. cellulosum* So ce56 (70 ECFs), emphasizing the increased regulatory capacity incurred by genome expansion. Among the ECFs acquired exclusively in *S. cellulosum* So0157-2, I found an additional member of ECF03, an extra member of ECF26, two additional members of ECF41 and one extra member of ECF56. ECF03 and ECF26 are novel acquisitions present in *S. cellulosum* So0157-2 but not in *S. cellulosum* So ce56. Indeed, members of ECF03 are mainly present in Bacteroidetes (Table S3.1), and could have been acquired by horizontal gene transfer. However, this protein is not overexpressed under alkaline conditions (Han *et al.*, 2013), indicating that this ECF is either not autoregulated, or not responsible for alkaline resistance in *S. cellulosum* So0157-2. In contrast, the additional member of ECF26 contained in *S. cellulosum* So0157-2 is overexpressed at pH 10 (Han *et al.*, 2013) and could therefore be part of the alkaline resistance observed for *S. cellulosum* So0157-2. This ECF belongs to ECF26s1, which shares a conserved genetic neighborhood with a catalase (-1 from the ECF coding sequence) and a cytochrome b561 (position -2). Whether ECF26 or any other of these ECFs provides *S. cellulosum* So0157-2 with alkaline resistance needs further investigation.

In summary, the updated ECF classification presented in this section serves as a source of testable hypotheses to guide the experimental characterization of this important class of bacterial regulators. The ECF classification comes together with a full description of ECF groups, including the putative group-specific ECF regulators, conserved proteins encoded in the same genetic neighborhood, and predicted target promoter motifs (Table S3.1). Collectively, this information allows for the prediction of the potential function of the members of the group, which is verified by experimentally described members, when available. Moreover, the two levels of this hierarchical classification provide a broad sequence collection with an appropriate degree of similarity (or variability) required for *in silico*

prediction tools that employ sequence variation-based algorithms, described later in this thesis. The ECF classification and tools for its analysis will be available online as ECF hub, where researchers in the field can classify their own ECFs and check the ECFs of their organism.

#### 4. Study of the binding between class I anti- $\sigma$ factors and ECF $\sigma$ factors

The most common regulators of ECF  $\sigma$  factors are anti- $\sigma$  factors, which sequester their cognate ECF under un-induced conditions. Like two-component systems, where a transmembrane histidine kinase is regulating the function of a response regulator with an output domain, anti- $\sigma$  factors and ECFs form signaling modules that transfer input cues from the extracellular space to the modification of gene expression. ECF anti- $\sigma$  factors have been classified into three families according to the secondary structure of their anti- $\sigma$  domain (ASD). Class I anti- $\sigma$  domains (ASDI), first described as a common fold of ECF anti- $\sigma$  factors by Campbell and colleagues (Campbell *et al.*, 2007), are the most abundant and also the most complex, with four  $\alpha$  helices that bind the ECF  $\sigma$  factor (Campbell *et al.*, 2007; Schumacher *et al.*, 2018) (see Section 1.4.1 for a description of anti- $\sigma$  factor classes).



**Figure 4.1. Structures of ECF  $\sigma$  factors in complex with class I anti- $\sigma$  factors.** ECFs are shown in pink colors, whereas anti- $\sigma$  factors appear in blue colors. Different areas of the protein are differentially colored (see legend). Different ASDIs inhibit ECF  $\sigma$  factor activity utilizing a different binding mode.

Given their abundance, the structures of four ASDI-ECF complexes have been solved (Campbell *et al.*, 2003, 2007; Shukla *et al.*, 2014; Devkota *et al.*, 2017), revealing a similar binding mode between ASDIs and ECFs, where ASDI's first three helices form a bundle that binds to  $\sigma_4$  domain, whereas their fourth helix, separated by a linker from the first three helices, binds to  $\sigma_2$  domain (Sineva, Savkina and Ades, 2017). Even though this general binding mode is preserved across the four resolved structures, each one seems to perform this binding using different residues (Fig. 4.1). The most extreme case is RsiW-SigW complex, where helix 3, which is sandwiched between  $\sigma_2$  and  $\sigma_4$

domains of the ECF in the other three solves structures, wraps around them instead (Fig. 4.1B) (Devkota *et al.*, 2017).

Recently, co-variation-based approaches have become one of the standards to reveal protein-protein interactions (de Juan, Pazos and Valencia, 2013). One of the *in silico* methods to determine the residues that are most likely to interact in a family of proteins is DCA, which requires large families of homologous proteins (M. Weigt *et al.*, 2009) (Section 1.6). Given the abundance of ECF  $\sigma$  factors and the fact that ASDIs are the most common regulator of these proteins, DCA is a realistic option for deciphering the details of the contact between both proteins. Aside from co-varying residues, proteins have residues that are specific of certain subfamilies and may specify ligand or protein interactions that only occur in a specific subfamily, the Specificity Determining Positions (SDPs) (de Juan, Pazos and Valencia, 2013).

Even though the co-evolution of ECFs and ASDIs has been suggested from their proximity in the genome and the conservation of ASDIs within ECF groups (Staroń *et al.*, 2009), a comprehensive classification of ASDIs and a measurement of the co-evolution with ECF  $\sigma$  factor has never been performed. This analysis is key for the application of DCA and for the extraction of SDPs, since co-evolution would support that the evolutionary pressure that modelled ASDI sequence is mostly dependent of its target ECF  $\sigma$  factor, hence, allowing the usage of co-evolution-based methods.

The similar secondary structure but the slightly different binding modes across ASDI structures suggest the presence of common residues that govern ASDI-ECF binding and other complex-specific contacts. In this section I describe the common principles that control ASDI binding to ECF  $\sigma$  factors.

#### 4.1. Class I anti- $\sigma$ domain (ASDI) retrieval and classification

In order to apply DCA to ASDIs, I collected ASDIs encoded in the genetic neighborhood of the coding sequences of ECF  $\sigma$  factors. Then, I classified them and compared their classification with the one of ECF  $\sigma$  factors. I extracted ASDI-containing proteins from the set of putative anti- $\sigma$  factors identified during the ECF classification (Section 3) using the HMMs for zinc-binding and non-zinc binding ASDIs constructed from the library of Staroń and colleagues (Staroń *et al.*, 2009). This step yielded 7,490 proteins. In order to further expand the size of the library, I built a new extended HMM from the ASDI of these sequences. I used this extended model to search for ASDIs in the genetic neighborhood of ECFs extracted during the library expansion, using only ECFs from representative and reference organisms as labelled by NCBI. This yielded 11,939 proteins, from which I removed the ones with ASDI shorter than 50 amino acids, since these could be divergent class II anti- $\sigma$  factors (Sineva, Savkina and Ades, 2017). The final number of ASDIs retrieved by this pipeline was 10,930, of which 10,806 have a non-redundant ASD. This shows that ~52% of the anti- $\sigma$  factors in the genetic neighborhood of ECFs are of class I, and that ~32% of the ECFs are regulated by ASDIs. The average size of the ASDI was  $100.85 \pm 33.20$  (standard deviation) amino acids. I observed that this library of putative anti- $\sigma$  factors is composed of a similar amount of zinc (~43%) and non-zinc binding (~57%)

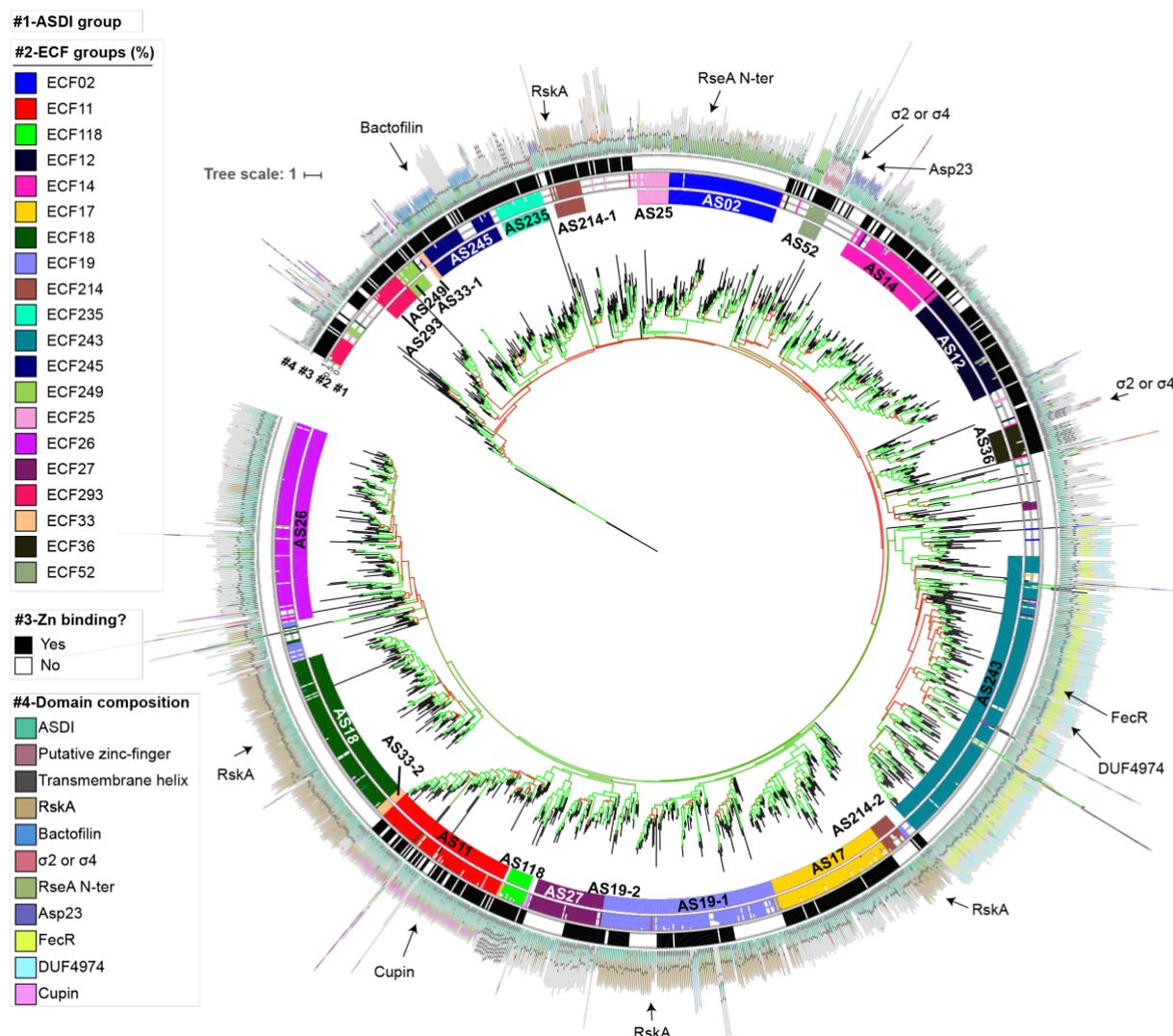
proteins. Most of the anti- $\sigma$  factors are bound to the membrane (~58%) and harbor one (~49% of the total of anti- $\sigma$  factors) transmembrane helix. I observed that ~59% of the soluble anti- $\sigma$  factors and ~48% of the membrane-bound anti- $\sigma$  factors are non-zinc binding. This contrasts with previous observations, where zinc-binding domains are preferred in soluble anti- $\sigma$  factors and membrane-bound anti- $\sigma$  factors usually lack zinc-binding domain (Campbell *et al.*, 2007).

I classified ASDIs extracted from these class I anti- $\sigma$  factors according to their sequence similarity into 1,475 subgroups of closely related sequences. For that, I used a divisive strategy, where the pool of sequences was subjected to a bisecting K-means clustering strategy until the maximum k-tuple distance among sequences in the cluster is smaller than 0.6 (Section 8.8). Then, the consensus sequences of subgroups were hierarchically clustered into a phylogenetic tree (Fig. 4.2). A simple inspection of the ASDI tree shows that the main ECF groups are preserved (Fig. 4.2, ring #2), supporting the co-evolution of ECFs and ASDIs. Given this similarity, I split the ASDI tree into monophyletic groups that regulate ECFs from the same group (Fig. 4.2, ring #1). This split usually agrees with high bootstrap values (Fig. 4.2 and Fig. 4.3), suggesting that this definition of ASDI groups is robust to changes in the dataset. As a result, ASDI groups were named with “AS”, followed by a number dependent on the ECF group they regulate. Even though ASDIs with the same sequence features and from the same area of the ASDI tree usually regulate members of the same ECF group, in some cases ASDI groups with different sequence features and located at some distance in the ASDI tree regulate ECFs of the same group. Two of these ASDI groups are AS19-1 and AS19-2, which regulate members of ECF19 and contain a zinc-binding domain (Fig. 2), but are divergent in the helix 1 (consensus HTLAGAYALDAL in AS19-1 and HLDPDQLALLA in AS19-2) and helix 2 (consensus of LDDERAAFERHL in AS19-1 and GEPLDADERAHL in AS19-2). ASDIs that regulate ECFs from the same subgroup are usually located together in the tree, but split into distinct ASDI subgroups (data not shown), probably due to the larger sequence diversity of anti- $\sigma$  factors respect to ECFs.

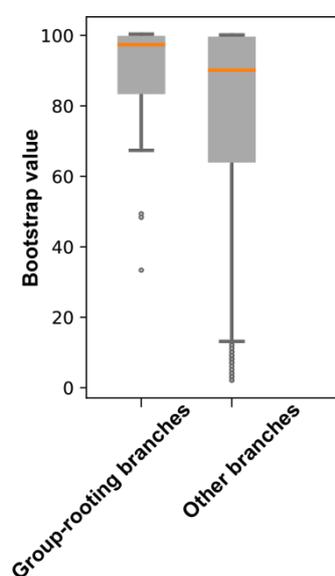
I observed that the presence of a mixture of zinc-binding and non-zinc binding ASDIs in the starting dataset does not affect their distribution across the tree, generating ASDI groups that are mixtures of zinc and non-zinc binding proteins, such as AS19-1 and AS27 (Fig. 4.2, ring #3). Exceptions are groups AS33-1 and AS33-2, whose difference is the presence or absence of the zinc-binding domain, respectively (Fig. 4.2, ring #3). I assessed the conservation of additional domains associated to ASDI-containing anti- $\sigma$  factors, since these domains generally specify the stimuli that trigger anti- $\sigma$  factor inhibition (Lewerke *et al.*, 2018; S. Li *et al.*, 2019). For that, I scanned full-length class I anti- $\sigma$  factors with Pfam models. I included the extended ASDI family model used for the expansion of the ASDI library to plot its position in the different class I anti- $\sigma$  factors. Additionally, I predicted the mode number of transmembrane helices in the different subgroups using the consensus prediction from online TopCons (Section 8.8). As a result, I observed that the protein domains associated to ASDIs are conserved for ASDIs from the same group, but differ between groups (Fig. 4.2, ring #4).

This suggests that ASDIs that regulate members of the same ECF group are inhibited by a similar mechanism, either by direct binding to the triggering molecule or by other adaptor proteins that, in turn, function as sensors.

Given the ample degree of correlation between ECF and ASDI classifications, I evaluated whether these families co-evolve. For this, I calculated the Pearson Correlation Coefficient (PCC) of the pairwise distance matrices of ASDIs and ECFs, as described by Goh and colleagues (Goh *et al.*, 2000). In order to determine the significance of the correlation coefficients considering the composition bias of proteins within the same organism, I included as negative controls RsbW-like anti- $\sigma$  factors and RpoD-like  $\sigma$  factors, emulating the strategy used by Dintner and colleagues (Dintner *et al.*, 2011). RsbW is the anti- $\sigma$  factor of the alternative  $\sigma$  factor  $\sigma^B$  and a protein kinase of the anti-anti- $\sigma$  factor RsbV in *B. subtilis* (Dufour and Haldenwang, 1994). RsbW-like anti- $\sigma$  factors have not been found to regulate ECF  $\sigma$  factors. RpoD is the housekeeping  $\sigma$  factor of *E. coli*. Housekeeping  $\sigma$  factors have not been found to be regulated by ASDIs. Therefore, RpoD-like and RsbW-like proteins should not interact with ASDIs and ECFs, respectively, and should display a low PCC that would serve as control for the lack of interaction. Indeed, low PCCs (around 0.5 to 0.6) were obtained for negative controls. These values are similar to the ones obtained by Dintner and colleagues (Dintner *et al.*, 2011). However, a PCC of 0.82 was obtained when correlating ECFs and ASDIs, showing that these families of proteins co-evolve (Table 4.1).



**Figure 4.2. ASDI phylogenetic tree.** Phylogenetic tree of the consensus sequences of subgroups of class I anti- $\sigma$  factor domains. Rings are explained as follows: #1) ASDI group, defined in this work, #2) ECF group of the ECFs encoded in the same genetic neighborhoods, #3) presence of Zn-binding motif, and #4) average domain composition of anti- $\sigma$  factors associated to each subgroup. Colors of the branches indicate bootstrap values from red (bootstrap=0) to green (bootstrap=100).



**Figure 4.3. Bootstrap value distribution in different branches of the ASDI tree** (Fig. 4.2). Rooting branches of ASDI trees have generally higher bootstrap values.

#### 4.2. DCA predicts two main contact interfaces between ASDIs and ECFs

I applied DCA (Martin Weigt *et al.*, 2009) to the full dataset of ASDI proteins and their cognate ECFs with the aim of studying the inhibitory mechanism of ASDI over ECFs. Results of this analysis revealed a large density of high DCA scores within  $\sigma_2$  and  $\sigma_4$  domains of the ECF  $\sigma$  factor, and also connecting both domains (Fig. 4.4A). This pattern agrees with previous DCA results in ECF  $\sigma$  factors (Wu *et al.*, 2019) and is indicative of the conserved secondary and tertiary structure on the protein. I also observed high scores interconnecting helices 1, 2 and 3 of the ASDI (Fig. 4.4A). In contrast, helix 4 appears isolated from the

rest of the ASD (Fig. 4.4A). This agrees with the crystal structures of ECF-ASDI complexes, where helices 1, 2 and 3 form a helix bundle connected to helix 4 by a flexible linker (Campbell *et al.*, 2003, 2007; Shukla *et al.*, 2014; Devkota *et al.*, 2017). I focused on the predictions that link ECFs and ASDIs, since these are likely the ones that inhibit ECF activity. A first glance of the contact map shows several high DCA scores linking the fourth helix of the ASDI with ECF's  $\sigma_2$  domain (Fig. 4.4A). Under closer inspection, I found that the top 14 inter-protein contact predictions (DCA score  $\geq 0.255$ ) are in close proximity in most of the crystal structures (Fig. 4.4 B and C). Of those, 12 are connecting ECF's  $\sigma_2$  domain and ASDI's helix 4, and two (#10 and #11) connect a single residue of ASDI's helix 1 to two residues located in  $\sigma_4$  domain of the ECF (Fig. 4.4E). In the first case, the area occluded by the anti- $\sigma$  factor includes ECF regions 2.1 and 2.2, whose main function is the binding to the clamp helices of the  $\beta'$  subunit of the RNAP (Wilson and Lamont, 2006; Lane and Darst, 2010b; L. Li *et al.*, 2019). It is likely that binding of ASDI's helix 4 to this area prevents ECF binding to the RNAP core, hampering ECF-dependent transcription when the anti- $\sigma$  factor is present. Instead, predictions #10 and #11 involve ECF helices 4.2 and 4.4, in two residues involved in the contact with the -35 element of the promoter (Lane and Darst, 2006; L. Li *et al.*, 2019).

**Table 4.1. Pearson Correlation Coefficient** of the distances between ECFs and ASDIs in organisms that contain RsbW-like and RpoD-like proteins, used as negative controls for the lack of correlation. Distances between pairs of proteins were measured using k-tuple distance implemented in Clustal Omega (Wilbur and Lipman, 1983; Sievers and Higgins, 2014) (Section 8.9).

	ECFs	ASDIs	RsbW	RpoD
ECFs	1.00			
ASDIs	0.82	1.00		
RsbW	0.56	0.63	1.00	
RpoD	0.48	0.50	0.67	1.00

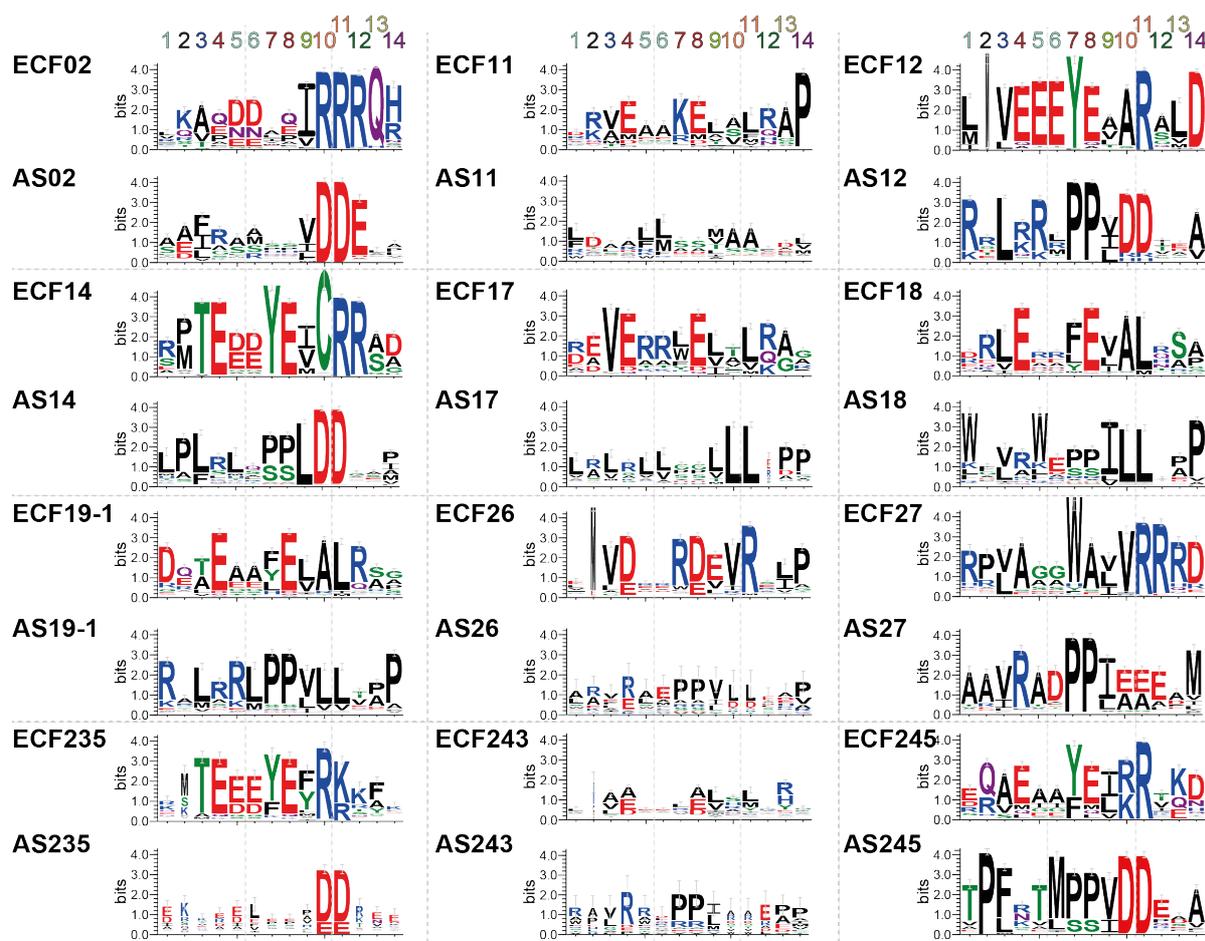
I plotted the specific residues that take part in the DCA predictions, both on the ECF and in the ASDI sides, for the different ASDI groups (Fig. 4.5). The resulting logos showed that many residue pairs are conserved within ECF/ASDI groups, while the identity of the conserved amino acids differs in the different groups (Fig. 4.5). Specifically, while contacts involving ASDI's helix 1 and  $\sigma_4$  domain (#10 and #11) are generally conserved, especially within groups, residues in ASDI's helix 4 exhibit a limited conservation that may or may not be reflected on the contact with  $\sigma_2$  domain of the ECF (Fig. 4.4D, Fig. 4.5). Predictions #10 and #11 feature two main types of predicted contacts, involving either a charged or a hydrophobic interaction (Fig. 4.5). This pattern is more evident for prediction #11, which tends to harbor a positive amino acid in the ECF (R178 in RpoE<sub>E.coli</sub>) and a negative residue in the ASDI (D11 in RseA<sub>E.coli</sub>), but in some cases this is replaced by a hydrophobic contact, typically with leucine on both ECF and ASDI (L177 in SigK and L18 in RskA, from *M. tuberculosis*). These results show that ASDI groups harbor different interaction motifs, suggesting a group-specific binding specificity, at least in helix 1 predictions given their group-specific conservation. However, the lack of major conservation in helix 4 predictions within ASDI groups and the fact that this helix is the one that holds most of the DCA predictions suggests that helix 4 is in charge of further differentiating the



2Q1Z), RsiW to SigW-RsiW from *B. subtilis* (PDB: 5WUQ) and RskA to SigK-RskA from *M. tuberculosis* (PDB: 4NQW). **D**: Multiple-sequence alignment of two selected ECF-ASDI pairs, RpoE-RseA from *E. coli* and SigK-RskA, from *M. tuberculosis*. Labels of the top 14 contacts indicate their position. The presence of  $\alpha$  helices and their names are depicted on top of the alignment in red boxes. Domain  $\sigma_4$  was split into four subregions for simplicity (4.1-4.4). The sequence logo depicts the composition of the full ECF and ASDI alignments. **E**: 3D depiction of the top 14 predictions in the structure of RpoE-RseA complex. ECF is colored in beige and anti- $\sigma$  factor in gray. Predicted contacts are labeled according to their rank. N and C-termini from ECF and anti- $\sigma$  factor are labelled.

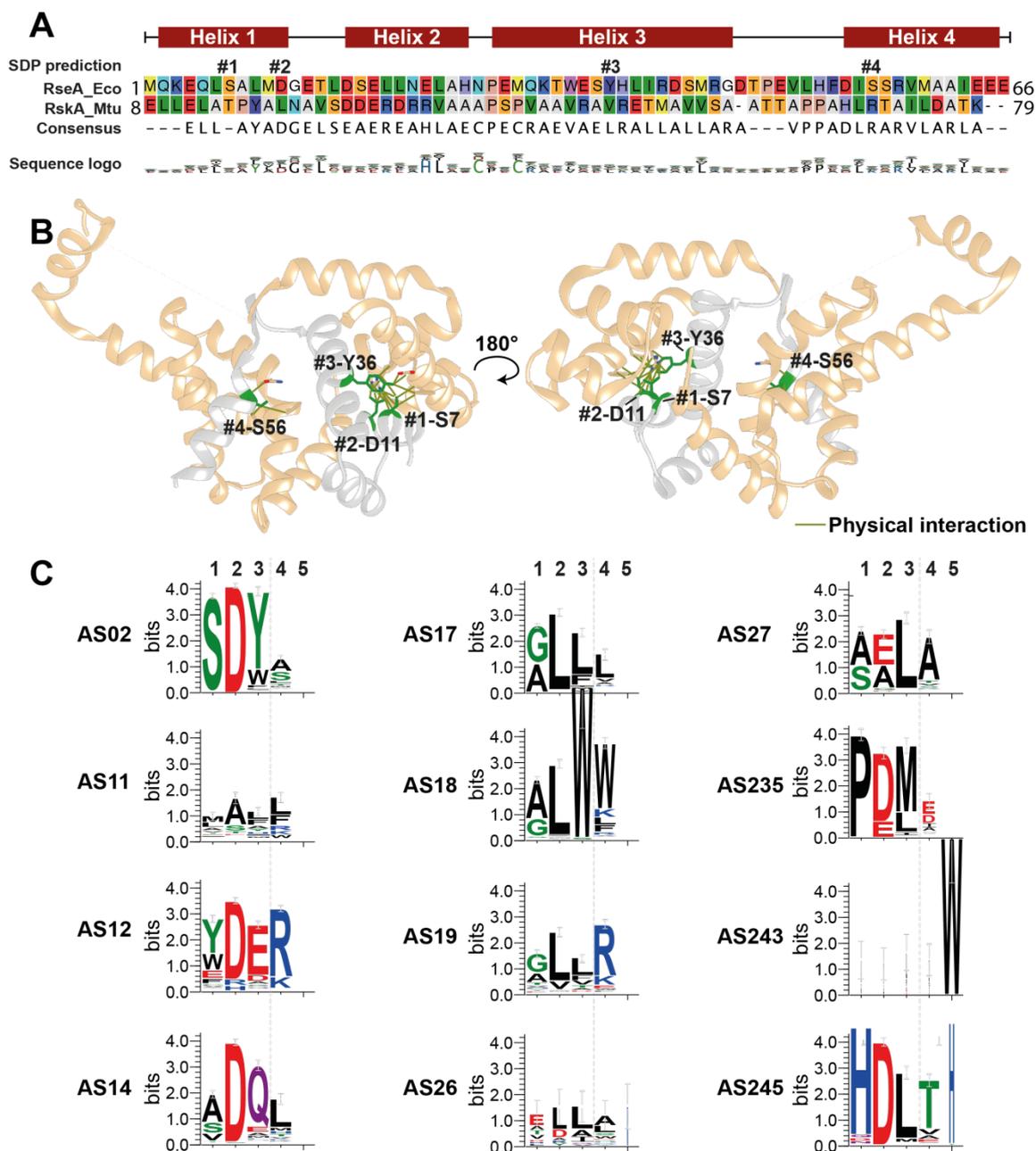
### 4.3. SDPs confirm that two main binding surfaces determine ECF/ASDI contact

DCA allowed for the identification of residues that are part of the contact interface in most of the ECF/ASDI complexes. However, most of the DCA predictions that lay in ASDI's helix 4 are not conserved within groups, showing that they do not determine specificity for ECFs of the same group. Therefore, I predicted specificity-determining positions (SDPs) for the 12 ASDI groups with more than 100 members using S3det (Rausell *et al.*, 2010). SDPs are residues that are specific from a subfamily of proteins and may specify ligand or protein interactions that only occur in a specific subfamily (Section 1.6). I predicted SDPs by comparing every pair of ASDI groups and took only the highest scoring SDP prediction of every ASDI group into further consideration (see Section 8.10 for more details). As a result, five SDPs were defined, two in helix 1, one in helix 3, one in helix 4 and the last one exclusively present in group AS243 (Fig. 4.6A). Proteins from group AS26 did not hold any prediction since they do not fit well into the multiple-sequence alignment of the full ASDI dataset, probably due to extensive differences at sequence level. Similarly, AS243's SDP corresponds almost exclusively to a gapped position in the alignment of the rest of the groups (Fig. 4.6C SDP#5). These differences at sequence level might reflect functional differences. In favor of this hypothesis, one member of AS243, FecR from *E. coli*, is distinguished from other non-AS243 ASDIs in that its 59 N-terminal amino acids are required for ECF activity (Ochs *et al.*, 1996) (Section 1.4.1.4). SDPs were named by running numbers (SDP#1 to SDP#5) from N- to C-terminus, or with their residue identifier in RseA<sub>*E. coli*</sub>, used as reference. Interestingly, all predicted SDPs are part of the contact interface with the ECF (Fig. 4.6B, Fig. 4.7). As expected, conserved position D11, predicted by DCA (Fig. 4.4B, #10 and #11), was part of the predicted SDPs (Fig. 4.6A, SDP#2). Yet, another residue predicted by DCA, V27 in helix 4 (Fig. 4.4B, #1 and #5), was also part of the SDPs (Fig. 4.6A, SDP#4). Predictions SDPs #1 and #3 connect S7 in helix 1 and Y36 in helix 3 (RseA<sub>*E. coli*</sub> coordinates) to  $\sigma_4$  domain, usually in its last helix (Fig. 4.6B, Fig. 4.7). Interestingly, SDPs #1, #2 and #3 form a cluster of interactions with the same area of the ECF, which usually corresponds to the last helix of the  $\sigma_4$  domain, except in SigE-ChrR structure, where the contact appears before this area (Fig. 4.6B, Fig. 4.7).

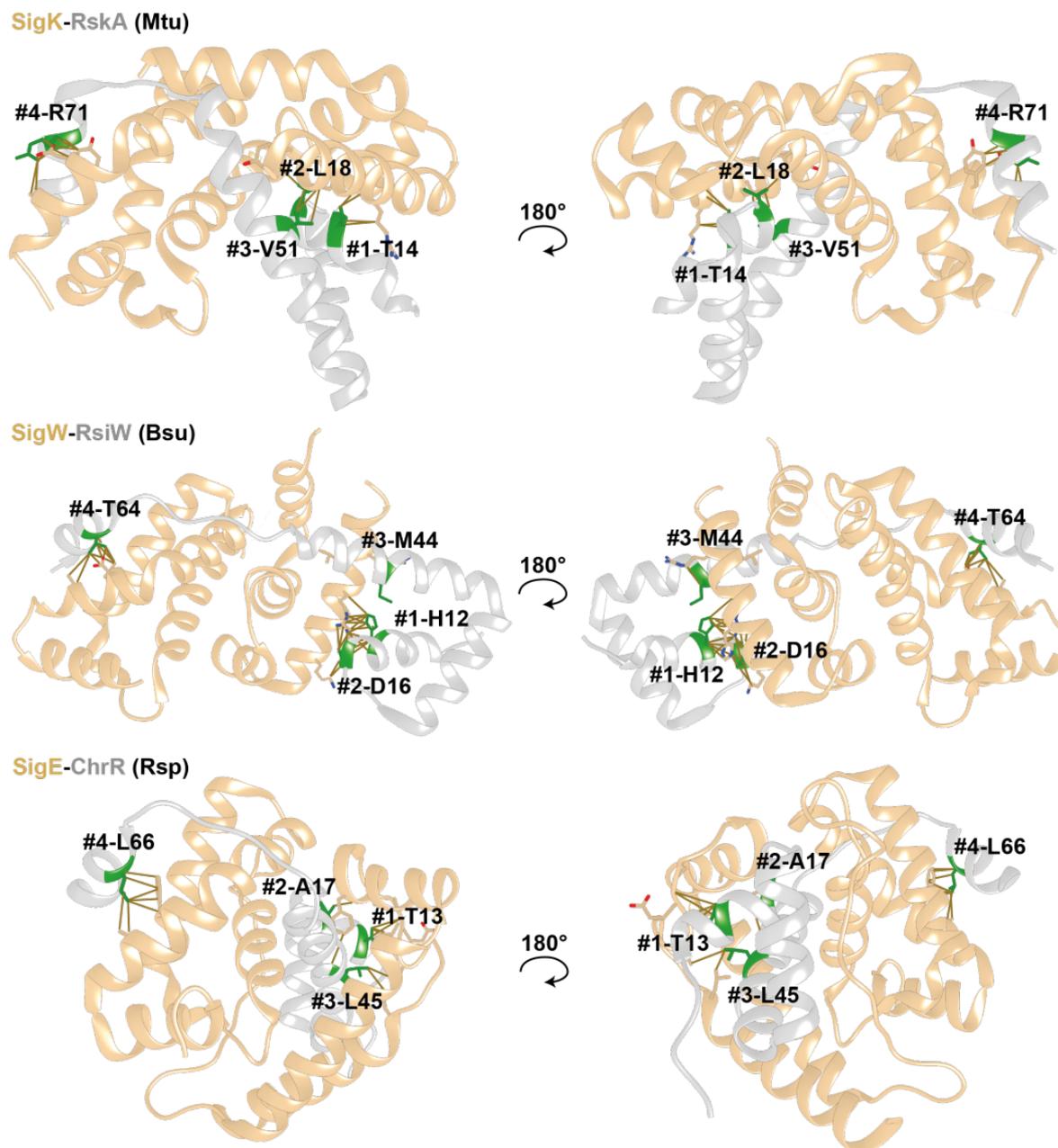


**Figure 4.5. Logos of the top 14 DCA predictions.** Logos of the ECF and anti- $\sigma$  factor residues involved in the top 14 DCA predictions, for the 12 ASDI groups with more than 100 members are shown.

Given that SDPs are characteristic features that distinguish subgroups of proteins, I found that SDP predictions are conserved within individual ASDI groups – except for AS26 and AS243 (Fig. 4.6C). Prediction SDP#1 is usually a hydrophobic amino acid or serine (Fig. 4.6C). Aromatic amino acids appear in this position in AS12 (Fig. 4.6C). Position SDP#2 is either negatively charged or hydrophobic, usually aspartate or leucine, as already discussed for DCA predictions #10 and #11 (Fig. 4). SDP#3, the only prediction in ASDI's helix 3, usually contains hydrophobic residues (9 out of 12 groups). Of those, aromatic residues are present in AS02 and AS18 (Fig. 4.6C). Charged residues are possible, but are only preferential in AS12, while polar residues appear in AS14 and AS245 (Fig. 4.6C). The last predicted SDP, SDP#4 in helix 4, is usually hydrophobic, but positively charged residues appear in AS12 and AS19, negative charges in ECF235 and a conserved threonine in ECF245 (Fig. 4.6C). Position SDP#5 is only present as tryptophan in AS243, and as histidine in some instances of AS245. Given that these residues are conserved within phylogenetic groups, face the ECF and feature different amino acids in different groups, it is likely that they take part in determining specificity towards the target ECF.



**Figure 4.6.** Description of the specificity-determining positions (SDPs) that distinguish different ASDI groups. **A:** Multiple-sequence alignment of the anti- $\sigma$  factors RseA from *E. coli* and RskA from *M. tuberculosis* showing the position of SDPs, labelled with numbers according to their position. Alpha helices and their names are indicated with red boxes on top of the ASDI sequences. The sequence logo shows the amino acidic composition of the full ASDI alignment. **B:** Structure of the RpoE-RseA complex from *E. coli* (PDB: 1OR7 (Campbell *et al.*, 2003)). SDPs are labeled as in A and their contacts with the ECF are shown by lines. ECF is represented in beige and anti- $\sigma$  factor in gray. **C:** Logo of SDPs in every ASDI group with more than 100 proteins. Positions are labelled as in A.



**Figure 4.7.** ASDI specificity-determining positions (SDPs) plotted in the structure of ECF-ASDI complexes. ECFs are colored in beige and anti- $\sigma$  factors in gray; SDPs are colored in green and labelled with their identifier as in Fig. 4.6. Complexes SigK-RskA from *M. tuberculosis* (Mtu, PDB: 4NQW (Shukla *et al.*, 2014)), SigW-RsiW from *B. subtilis* (Bsu, PDB: 5WUQ (Devkota *et al.*, 2017)) and SigE-ChrR from *R. sphaeroides* (Rsp, PDB: 2Q1Z (Campbell *et al.*, 2007)) are shown. Contacts with the ECF are represented by lines. A similar representation is available for RpoE-RseA pair in Fig. 4.6B.

#### 4.4. SDPs and DCA predictions show the general binding mode of ASDIs

A drawback of SDP predictions is that they do not provide information on their contacted residues. Looking at the four resolved structures of ECF-ASDI complexes (Campbell *et al.*, 2003, 2007; Shukla *et al.*, 2014; Devkota *et al.*, 2017) and resolving their contact interface using Voronoi tessellation (Olechnovič and Venclovas, 2014), ECF residues contacted by SDPs are sometimes shared by several structures and are also predicted by DCA (Table 4.2). Two physical contacts of SDP#4 (S56 in RseA from *E. coli*) with the ECF are also predicted by DCA, this is, V27 (DCA prediction #1) and N23

(DCA prediction #5) in the coordinates of RpoE<sub>E.coli</sub>. The contact between S56 (RseA) and N23 (RpoE) is present in the four crystal structures of ECF/ASDI complexes, strengthening the idea that DCA allows the prediction of contacts that appear in the overall protein family (Fig. 4.6B, Table 4.2). SDP#2 (D11 in RseA<sub>E.coli</sub> coordinates) harbors two DCA predictions, R149 (prediction #10) and R178 (prediction #11) in RpoE<sub>E.coli</sub> coordinates. Indeed, these contacts are observed in three out of four structures (Table 4.2). This shows that, even though DCA predictions are realized in most of the structures, the binding of ECF and ASDI still has some flexibility, suggesting that the contacts predicted by DCA might be missing in certain members of the family or in certain conformations and arguing in favor of an “induced fit” between ECFs and ASDIs, where the relative conformation of ECF-ASDI complex can vary to adjust for the chemical properties of the individual amino acids featured in both proteins. Predictions SDP#1 and SDP#3 are not part of the DCA predictions, but share physical interactions with RpoE<sub>E.coli</sub>'s R178, which is also contacted by SDP#2 in three of the structures (Table 4.2), indicating an important role of R178 in the inhibition by ASDI domains. However, even though R178 seems to be a pivotal residue for both DNA (Lane and Darst, 2006; L. Li *et al.*, 2019) and anti- $\sigma$  factor binding, it is not part of the observed contacts in SigE-ChrR complex (Table 4.2), showing again that while DCA predictions are often true contacts, some subfamilies of proteins show contact modes that deviate from this general scheme.

**Table 4.2. Description of the physical contacts between specificity-determining positions (SDPs) in ASDIs and ECF  $\sigma$  factors** in the four published crystal structures of these complexes. Structures are named after their ECF and are described by their ASDI group. RpoE represents RpoE-RseA complex from *E. coli* (PDB: 1OR7 (Campbell *et al.*, 2003)), SigE represents SigE-ChrR complex in *R. sphaeroides* (PDB: 2Q1Z (Campbell *et al.*, 2007)), SigK represents SigK-RskA complex from *M. tuberculosis* (PDB: 4NQW (Shukla *et al.*, 2014)) and SigW represents SigW-RsiW complex in *B. subtilis* (PDB: 5WUQ (Devkota *et al.*, 2017)). Crosses indicate contacts, as shown by Voronoi tessellation (Olechnovič and Venclovas, 2014). The ECF residue contacted is indicated between brackets. Contacts with ECF SDPs and contacts that are predicted by DCA are shown with crosses in their appropriate column.

SDP (RseA)	RpoE (Eco)	SigE (Rsp)	SigK (Mtu)	SigW (Bsu)	SDP on ECF?	DCA prediction?
	AS02	AS11	AS19	AS245		
#1 (S7)		X (A144)				
#1 (S7)		X (F145)			X	
#1 (S7)	X (Y156)					
#1 (S7)		X (R152)				
#1 (S7)		X (E153)				
#1 (S7)		X (L154)				
#1 (S7)		X (A155)				
#1 (S7)	X (R171)		X (K170)		X	
#1 (S7)	X (I174)		X (M173)			
#1 (S7)	X (F175)		X (R174)	X (H174)		
#1 (S7)	X (R178)		X (L177)	X (R177)	X	
#1 (S7)				X (E178)		
#1 (S7)			X (R181)			
#2 (D11)	X (S2)					
#2 (D11)	X (E3)					
#2 (D11)		X (F81)				
#2 (D11)		X (R85)	X (H85)		X	
#2 (D11)		X (R87)				
#2 (D11)		X (I89)	X (V89)			

#2 (D11)			X (I145)			
#2 (D11)	X (R149)	X (A144)	X (A148)	X (K148)		X
#2 (D11)	X (E150)	X (F145)	X (Y149)		X	
#2 (D11)	X (Y156)			X (L155)		
#2 (D11)				X (K170)	X	
#2 (D11)	X (I174)		X (M173)	X (I173)		
#2 (D11)				X (H174)		
#2 (D11)	X (R178)		X (L177)	X (R177)	X	X
#3 (Y36)		X (F81)				
#3 (Y36)	X (E150)	X (F145)			X	
#3 (Y36)		X (L154)				
#3 (Y36)		X (A155)				
#3 (Y36)		X (L162)				
#3 (Y36)	X (F175)					
#3 (Y36)	X (R178)		X (L177)	X (R177)	X	
#3 (Y36)	X (E179)			X (E178)		
#3 (Y36)	X (D182)		X (R181)	X (R181)		
#4 (S56)		X (F26)				
#4 (S56)	X (N23)	X (A27)	X (A27)	X (A21)		X
#4 (S56)	X (V26)	X (F30)	X (Y30)	X (V24)		
#4 (S56)	X (V27)	X (Q31)	X (D31)	X (D25)		X
#4 (S56)	X (Q30)		X (K34)	X (K28)		

Given that application of the SDP approach to ASDI's did not predict the contacted residues on the ECF side, the prediction of SDPs for the ECFs themselves could give further clues about the most relevant contact residues on the ECF  $\sigma$  factors. Following the same strategy as for the prediction of SDPs on ASDIs, I predicted nine SDPs on the ECF side, four of which are contacted by ASDI's SDPs (Table 4.2). Interestingly, one of the SDPs found on ECFs is RpoE<sub>E.coli</sub>'s R178 in helix 4.4 (Fig. 4.4D), involved in DNA binding (Lane and Darst, 2006; L. Li *et al.*, 2019). This ECF residue is 1) predicted by DCA, 2) an SDP that distinguishes ECF groups, and 3) contacted by an SDP in the ASDI side (D11 from RseA<sub>E.coli</sub>). This highlights the pivotal role of R178 for anti- $\sigma$  factor inhibition. The second ASDI residue identified by both by S3det and DCA is equivalent to S56 in RseA<sub>E.coli</sub> (SDP#4, DCA #1), which interacts with the residue equivalent to V27 in RpoE<sub>E.coli</sub>. Nevertheless, V27, on the ECF side, is not an SDP and is generally less conserved than its ASDI interaction partner (S56 RseA<sub>E.coli</sub>). This lack of conservation of V27 across ECF  $\sigma$  factors (Fig. 4.4D) contrasts with its important role in the interaction with  $\beta'$  subunit of the RNAP, since the equivalent residue in SigH from *M. tuberculosis* (I34 SigH<sub>M.tuberculosis</sub>) is in contact with the  $\beta'$  subunit in residue I365 (PDB: 5ZX2 and 5ZX3 (L. Li *et al.*, 2019)).

Taken together, SDP and DCA predictions suggest two different docking points used by ASDI to inhibit ECF activity, an extensive one that connects ASDI's helix 4 and ECF's  $\sigma_2$  domain and blocks the contact of the ECF with the  $\beta'$  subunit of the RNA polymerase, and a second one between ASDI's helices 1 and 3, and  $\sigma_4$  domain, which occludes residues essential for -35 element binding. Contact predictions between ASDI's helix 4 and  $\sigma_2$  domain are mostly composed by diverse residues that are

independent of the ECF/ASDI group. An exception is the interaction pair S56 (RseA<sub>*E.coli*</sub>) and V27 (RpoE<sub>*E.coli*</sub>), where the residues at position S56 are characteristic for individual ASDI groups and V27 is involved in binding to the  $\beta'$  subunit of the RNAP. In contrast, contact predictions between ASDI's helices 1 and 3, and the  $\sigma_4$  domain are centered on RpoE<sub>*E.coli*</sub>'s R178, which is conserved within many ECF groups, is involved in binding to the -35 element of the DNA and interacts with some ASDI residues, including DCA prediction D11 (RseA<sub>*E.coli*</sub>), also conserved within ASDI groups.

#### 4.5. The structure of ECF26/AS26 could differ from other ECF/ASDI complexes

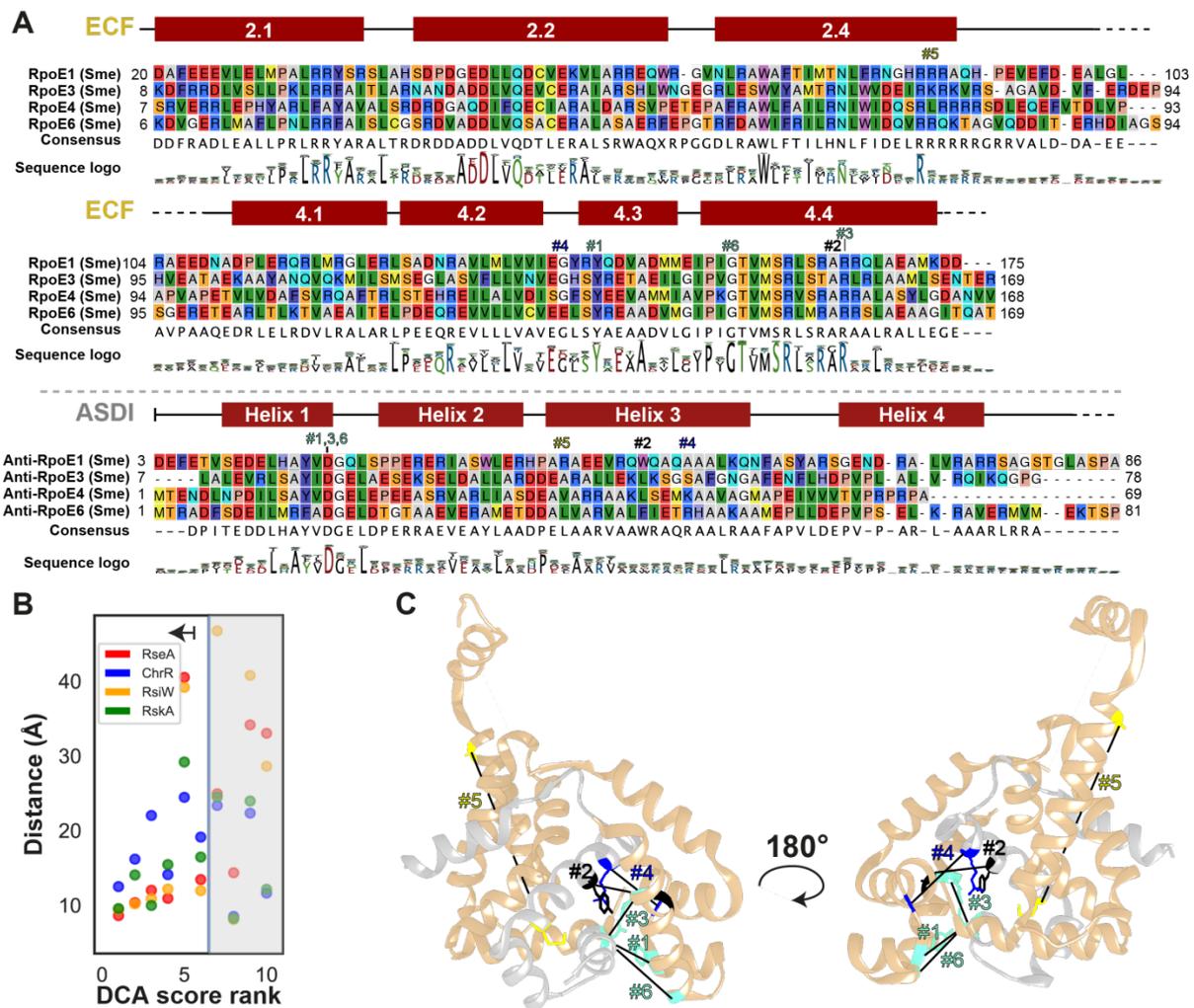
I further focused on AS26 since the ASD of this group seems to be divergent from other ASDIs, hampering their alignment with the rest of ASDIs (data not shown). Furthermore, four members of ECF26 (RpoE1, RpoE3, RpoE4 and RpoE6) have been experimentally addressed with their respective anti- $\sigma$  factors in the nitrogen-fixing Alphaproteobacterium *Sinorhizobium meliloti*. These ECFs share a large sequence similarity, with an identity ranging from 46% to 61%, and some common targets (Lang *et al.*, 2018). RpoE1 and RpoE4 are involved in the detoxification of sulfite and sulfite respiration (Bastiat *et al.*, 2012). These two proteins cross-activate *sort-sorU-azu2* operon (Bastiat *et al.*, 2012; Lang *et al.*, 2018). RpoE3 is involved in the expression of four genes un unknown function, including a putative RpoE1 target, and is not auto-regulated (Lang *et al.*, 2018). RpoE6 is involved in the transcription of ~40 genes and has a substantial functional overlap with RpoE2 (Lang *et al.*, 2018), from group ECF15, the general stress ECF  $\sigma$  factor in Alphaproteobacteria (Section 1.4.1.3). However, their anti- $\sigma$  factors, with less than 20% sequence identity, are functionally orthogonal, this is, they do not regulate other ECF26 in *S. meliloti* (personal communication Dr. Doreen Meier, Desiree Körner and Prof. Dr. Anke Becker, unpublished).

Given the similarity at sequence level of members of ECF26 in *S. meliloti* but their orthogonality respect to their anti- $\sigma$  factor, I focused on this group to reveal how could this specificity arise. An alignment of anti-ECF26 ASDIs showed that the amino acid conservation is reduced from mid helix 3 towards the transmembrane helix, located after helix 4 (Fig. 4.8A). This differs from the full ASDI family alignment, where helix 4 has generally similar conservation levels as helix 1 (Fig. 4.6A). Indeed, the Clustal Omega is not able to align well anti- $\sigma$  factors from group AS26, indicating differences at a sequence level also within this group. However, the ECF sequence is conserved across members of ECF26, except for both termini, which are extended in some cases, and the linker between  $\sigma_2$  and  $\sigma_4$  domains (Fig. 4.8A).

Encouraged by the large number of ECF-ASDI pairs that belong to group 26 (n=588) and their apparent sequence similarity, I performed DCA to unravel new contacts that could shed light into the differences between members of group 26 and other ECF-ASDI pairs, or contacts that could help to understand the orthogonality observed in members of group 26 in *S. meliloti*. The distance between  $\alpha$ -carbons in the top 10 predictions is generally higher than predictions for the overall ECF-ASDI contact (Fig. 4.4C and Fig. 4.8B). This could be due to 1) DCA results have lower quality in group 26

given that the starting number of protein pairs is smaller (588 vs 10,934), or 2) the available crystal structures (none of them from members of group 26) do not reflect well the structure of the complex ECF26-AS26. The latter is a plausible idea since the four structures resolved to date have a slightly different structure (Fig. 4.1). The top 6 DCA predictions for group 26 generally had a distance between  $\alpha$ -carbons shorter than 20Å in the four crystal structures of ECF-ASDI (Fig. 4.8B), suggesting a possible contact in pairs from group 26 if they had an alternative conformation.

DCA predictions for group 26 DCA will be referred by their coordinates in RpoE or RseA in *E. coli* to be able to compare with the DCA predictions in the full ASDI family. Three predictions involved residue D11 in RseA<sub>E.coli</sub> (#1, #3 and #6), in helix 1. This residue is predicted to contact the last part of ECF's domain  $\sigma_4$  in 3 different regions, namely helix 4.3 (#1-Y156 RpoE<sub>E.coli</sub>) and 4.4 (#3-R178 and #6-G168 in RpoE<sub>E.coli</sub>) (Fig. 4.8 A and C). One of these predictions, D11-R178, was already identified by the DCA of the full dataset (Fig. 4.4B, #11), whereas both predictions #1 and #3 are in physical contact in RpoE-RseA structure (Table 4.2). Instead, prediction #6 appears for the first time for members of group 26 and, given that it is near D11(RseA<sub>E.coli</sub>) in the four structures of ECF-ASDI complexes (Fig. 4.8B), a slight modification of the conformation in members of ECF26-AS26 could allow this contact. Aside from predictions with ASDI's helix 1, DCA in group 26 showed three predictions in helix 3, which did not hold any prediction in the full dataset DCA (Fig. 4.8 A and C, #2, #4 and #5). Predictions #2 (RpoE<sub>E.coli</sub>'s A177 with RseA<sub>E.coli</sub>'s Y36) and #4 (RpoE<sub>E.coli</sub>'s G154 with RseA<sub>E.coli</sub>'s R40) link helix 3 to ECF's  $\sigma_4$  domain. The proximity of these amino acids in the structures ECF-ASDI complexes (Fig. 4.8B) indicates that they are feasible in members of group 26. Instead, prediction #5, which links ASDI's helix 3 (RseA<sub>E.coli</sub>'s M29) to ECF's  $\sigma_2$  domain (RpoE<sub>E.coli</sub>'s A88), has a distance over 20Å in the four structures, indicating that it could be a false positive. Future resolutions of the structure of ECF-ASDI complexes from group 4 would answer to whether this prediction is, indeed, a contact in members of this group.



**Figure 4.8. Analysis of ECFs and ASDIs from group 26.** **A:** Multiple-sequence alignment of members of ECF26 and AS26 ( $n=588$ ). The sequences of the four pairs found in *S. meliloti* are shown. Consensus sequence and logo correspond to the full group. Alpha helices in the mapped structure of RpoE-RseA pair, from *E. coli*, are shown and labelled. Region  $\sigma_4$  is split into four subregions according to its  $\alpha$  helices for simplicity. Labels on the sequence indicate the top 6 DCA results for group 26 alone. **B:** distance between the  $\alpha$ -carbons of the top 10 DCA predictions for group 26 mapped into the four structures of ECF-ASDI complexes. RseA corresponds to RpoE-RseA complex from *E. coli* (PDB: 1OR7 (Campbell *et al.*, 2003)), ChrR to SigE-ChrR from *R. sphaeroides* (PDB: 2Q1Z), RsiW to SigW-RsiW from *B. subtilis* (PDB: 5WUQ) and RskA to SigK-RskA from *M. tuberculosis* (PDB: 4NQW). The top 6 predictions were chosen as positives. **C:** DCA results for group 26. Colors indicate clusters of contacts that share one amino acid. Lines link DCA predictions. Labels indicate the rank of the prediction.

Predictions #1,3,6 sit in a single position on helix 1 of the ASDI, equivalent to D11 in RseA<sub>E.coli</sub>. This position harbors an aspartate in most of the members of AS26 (Fig. 4.8A). Moreover, these predictions also display a large conservation in the ECF side (Fig. 4.8A). This agrees with observations on D11 in the full dataset ECF (Fig. 4.4D) and argues in favor of D11 (RseA<sub>E.coli</sub>) as a pivotal point in the contact between ECFs and ASDIs, also in the group 26. In contrast, predictions in ASDI's helix 3 in members of group 26 are not conserved (Fig. 4.8A). In the case of AS26 in *S. meliloti*, these positions harbor different amino acids in all the sequences (Fig. 4.8A). This variability on the anti- $\sigma$  factor is reflected on the ECF side. Even though the ECF sequence is generally more conserved (Fig. 4.4D, Fig. 4.8A), the residues that harbor DCA predictions that contact ASDI's helix

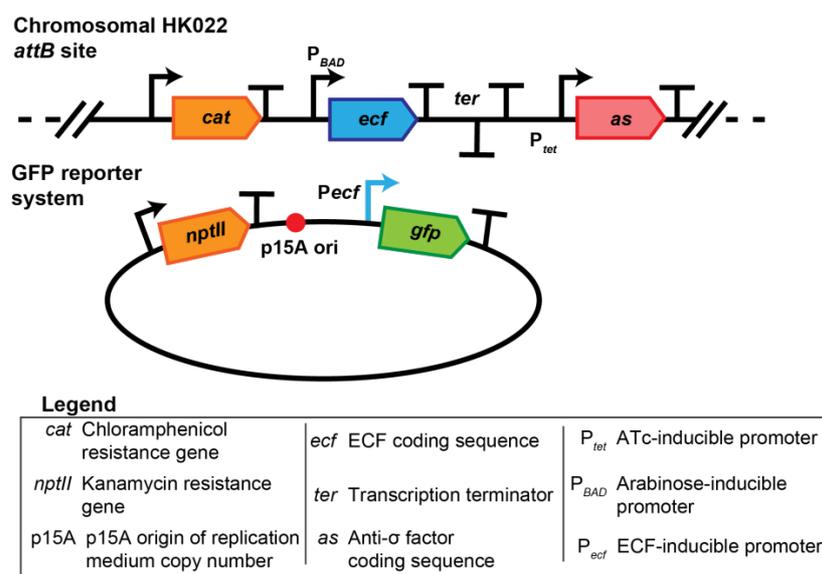
3 are less conserved than predictions that contact helix 1 when looking at the complete group 26 (Fig. 4.8A, logos). This argues in favor of this variable residues determining specificity in group 26. However, when focusing on the members of ECF26 from *S. meliloti*, these positions on the ECF side seem rather conserved (Fig. 4.8A). Therefore, even though these residues could determine specificity to a certain degree, they do not seem to be the ones that give orthogonality to the four pairs of ECF-anti- $\sigma$  factor from group 26 in *S. meliloti*. In contrast to observation in the full ECF-ASDI pairs, ASDI's helix 4 does not hold any prediction. Possible reasons are that 1) ASDI's helix 4 is not important for the inhibition of members of ECF16, or 2) helix 4 binds to the ECF in different configurations in distinct members of AS26, which hampers DCA prediction. In favor of the first, deletion mutants without helix 4 are able to repress the activity of members of ECF26 in *S. meliloti* (Meier *et al.*, unpublished). This hypothesis would also explain why helix 4 of anti-RpoE4 is degenerated, while this protein is still functional (Fig. 4.8A). However, experimental conformation solving the structure ECF26-AS26 complexes is required to confirm these ideas.

#### 4.6. Testing the inactivation of anti- $\sigma$ factors by changes in membrane potential

Anti- $\sigma$  factors are usually transmembrane proteins that sequester the ECF in an inactive conformation. Different environmental signal can inactivate anti- $\sigma$  factors (Mascher, 2013) (Section 1.4.1). These environmental signals are diverse and could involve, among others, unfolded outer membrane proteins, LPSs, siderophores bound to metals (Chevalier *et al.*, 2019), heavy metals (Grosse, Friedrich and Nies, 2007) and reactive oxidative species (Sineva, Savkina and Ades, 2017). Since the signals transduced by anti- $\sigma$  factors are so diverse, it could be possible that changes in voltage across the cell membrane trigger their inactivation. Membrane potential, or voltage, is defined by the difference in ionic concentration across the membrane. Due to the proton motive force, *E. coli* cells have more positive charges outside the cell than inside. However  $K^+$  cation is in a higher concentration in the cytoplasm (Felle *et al.*, 1980). Intracellular  $K^+$  concentration has been determined to be  $\sim 210$ mM during the early logarithmic phase in *E. coli* (Schultz and Solomon, 1961). Cells are depolarized when the difference of potential across the membrane gets reduced, or hyperpolarized when it increases.

In order to test whether anti- $\sigma$  factors are sensitive to changes in membrane potential, I built six genetic circuits using different ECFs and their cognate anti- $\sigma$  factors. In these circuits, ECF expression was controlled by an arabinose inducible  $P_{BAD}$  promoter while anti- $\sigma$  factor expression was controlled by an anhydrotetracycline (ATc) inducible  $P_{tet}$  promoter. The ECF/anti- $\sigma$  factors (AS) used were ECF/AS15\_436, ECF/AS16\_3622, ECF/AS22\_4450, ECF/AS28\_1088, ECF/AS38\_1322 and ECF/AS02\_2817, the latter native to *E. coli* (RpoE-RseA system). These genetic parts were originally obtained from (Rhodius *et al.*, 2013). I selected these ECF/AS pairs due to their lack of toxicity, the activity of both ECF and anti- $\sigma$  factor in *E. coli* and the presence of transmembrane helices in their anti- $\sigma$  factor (Rhodius *et al.*, 2013; Pinto *et al.*, 2018). These ECF/anti- $\sigma$  circuits were assembled and subsequently integrated into the genome of *E. coli* (as described in Section 8.19), generating six

different *E. coli* strains. For each ECF/AS, I designed a reporter construct consisting of the target ECF promoter ( $P_{ecf}$ ) fused with a *gfp* reporter gene (Bisicchia, Botella and Devine, 2010) and assembled them on medium copy vectors (Section 8.19). These medium copy plasmids were used to evaluate ECF activity. The newly generated plasmids were then transformed in the *E. coli* strains carrying the cognate ECF/AS circuit, generating the strains GFC0414-GFC0419 (Section 8.19; Fig. 4.9; Table 8.3).

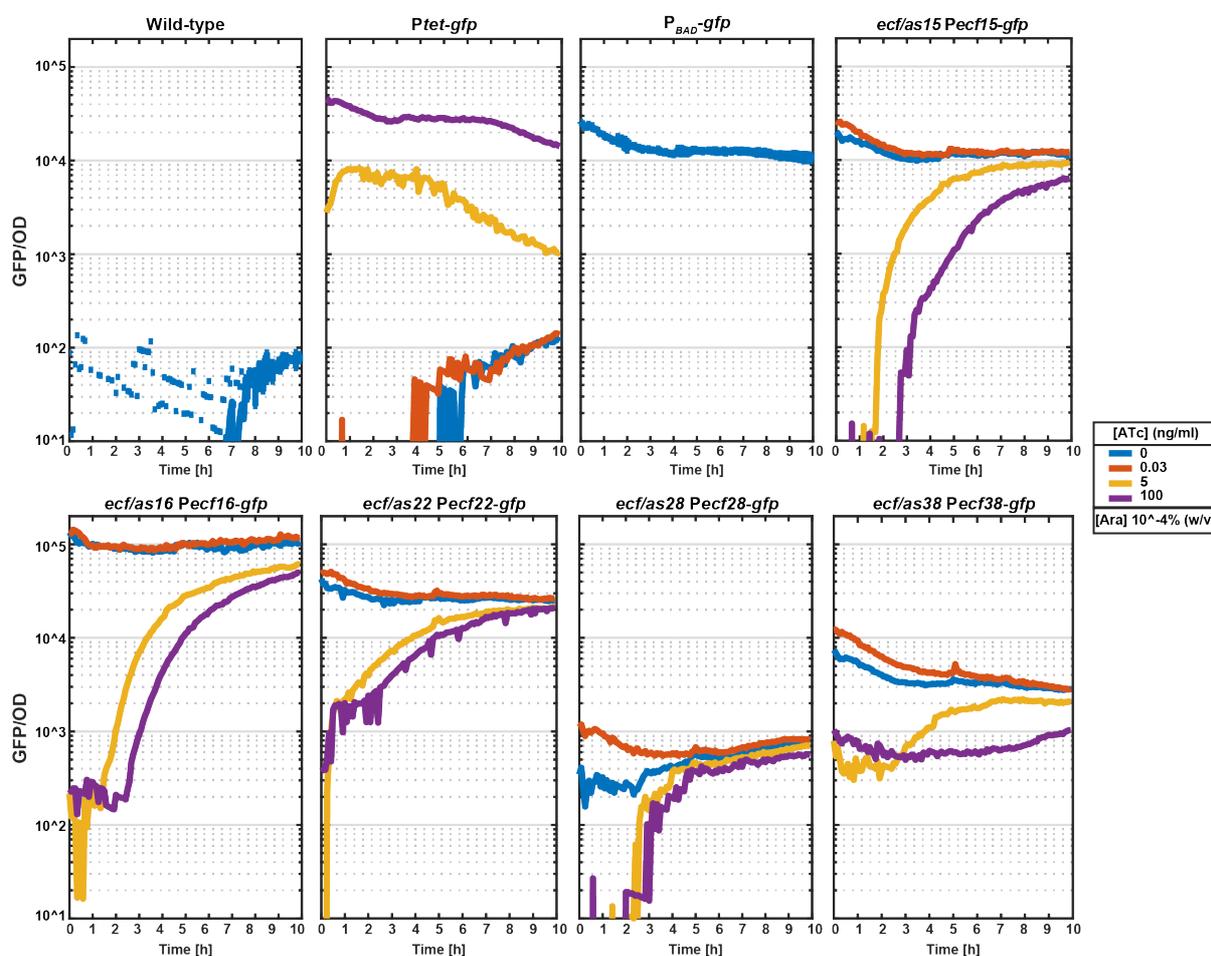


**Figure 4.9. Blueprint of genetic constructs used for testing** anti- $\sigma$  factors response to changes in the membrane voltage. The strains GFC0414-GFC0418 possess the ECF/AS construct integrated into the HK022 *attB* site, and the GFP reporter system. The strain GFC019 relies on the native RpoE-RseA system and contains only the reporter plasmid, where *gfp* is driven by  $P_{ecf02\_2817}$  promoter (Rhodius *et al.*, 2013).

Using these constructs, I initially tested different concentrations of arabinose and ATc inducers so as to achieve a state where the removal of ATc and, hence, of the anti- $\sigma$  factor, is able to trigger ECF activation and consequently *gfp* expression in less than 10 hours. In this manner, slight changes that affect the activity of the anti- $\sigma$  factors are quickly realized into an increase on *gfp* expression. For finding this critical concentration of ATc and arabinose, I cultured strains GFC0414-GFC0418 overnight in MOPS minimal medium supplemented with 0.5% (v/v) glycerol and concentrations ranging from 0.2%-10<sup>-5</sup> % (w/v) arabinose and 0-100ng/mL ATc (Section 8.19). Before the measurement, I washed out ATc to stop anti- $\sigma$  factor expression, and I tracked the fluorescence emitted by GFP and the OD<sub>600</sub> every 5min for 10 hours. Results of this experiment showed that cultures with 10<sup>-4</sup> % (w/v) arabinose and 5ng/mL ATc performed the desired dynamics (Fig. 4.10).

I treated cells with an ionophore in order to test whether anti- $\sigma$  factors and/or ECFs respond to changes in voltage. Ionophores are drugs able to transport ions across cell membranes, changing the voltage of the cells. During this study I used valinomycin. Valinomycin is a neutral cyclic peptide that makes the membrane permeable to K<sup>+</sup> (Ahmed and Booth, 1983). Therefore, its effect in cell potential depends on the concentration of K<sup>+</sup> in the medium. For high concentrations of K<sup>+</sup> in the medium,

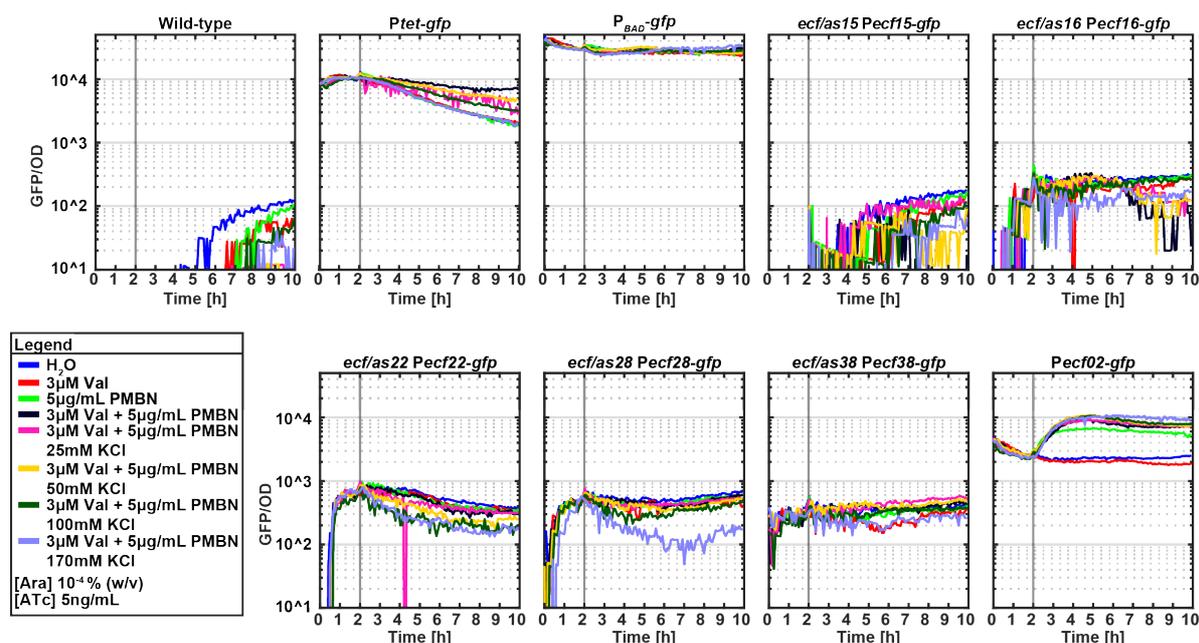
valinomycin transports these positive charges inside the cell, depolarizing the membrane, whereas in media with small amounts of  $K^+$ , valinomycin expels intracellular  $K^+$  and hyperpolarizes the cell (Ahmed and Booth, 1983). *E. coli* outer membrane is not permeable to valinomycin (Ahmed and Booth, 1983), driving it naturally resistant to this ionophore. However, polymyxin B nonapeptide (PMBN) sensitizes *E. coli* to valinomycin (Alatossava, Vaara and Baschong, 1984). PMBN is a cationic cyclic peptide obtained from the removal of the terminal amino acid of polymyxin B (Chihara *et al.*, 1973). This modification makes PMBN lose most of its antimicrobial activity (Chihara *et al.*, 1973), but it still allows it to bind to LPS and perturb the outer membrane of Gram negative bacteria (Alatossava, Vaara and Baschong, 1984; Tsubery *et al.*, 2000).



**Figure 4.10. Ability of ECFs to overcome anti- $\sigma$  factor inhibition after removing their inducer (ATc), as shown by GFP-emitted fluorescence.** Fluorescence emitted by GFP (GFP) and OD at 600nm ( $OD_{600}$ ) were measured every 5min in a 10-hour time course experiment, by using a TECAN<sup>®</sup> Infinite 200 PRO plate reader. Strains were grown in MOPS minimal medium supplemented with 0.5% (v/v) glycerol,  $10^{-4}$ % arabinose (w/v) and different concentrations of ATc. ATc was removed prior to measurement. ATc was not removed from the medium in  $P_{tet}$ -*gfp* control strain.  $P_{tet}$ -*gfp* and  $P_{BAD}$ -*gfp* strains (GFC0013 and GFC0014) were used as positive controls of the activity of  $P_{tet}$  and  $P_{BAD}$  promoters. The rest of the strains (GFC0414-GFC0418) contained  $P_{BAD}$ -*ecf*,  $P_{tet}$ -*as* integrated into HK022 *attB*, and  $P_{ecf}$ -*gfp* in a reporter plasmid, as in Fig. 4.9. The average fluorescence intensity in a wild-type strain without *gfp* (SV01) was used for blanking fluorescence, whereas the average  $OD_{600}$  of MOPS minimal medium was used for  $OD_{600}$  blanking. The average of two technical replicates is shown for wild-type and  $P_{BAD}$ -*gfp*. Error bars indicate standard deviation and are shown only for the positive direction. Results are based on one replicate for the rest of the strains. The fluorescence detection limit is set to  $10^{-3}$  (AU). The desired behavior is achieved with  $10^{-4}$ % (w/v) arabinose and 5ng/mL ATc.

Using the appropriate ATc and arabinose concentrations, if anti- $\sigma$  factors were inhibited by changes in voltage, the addition of ionophores would result in an increased *gfp* expression. For testing this idea, I

grew the different strains (GFC0414-GFC0419) in M9 media with arabinose ( $10^{-4}$  % (w/v)) and ATc (5ng/mL). After two hours, I treated these cultures with PMBN (5 $\mu$ g/mL), valinomycin (3 $\mu$ M), as suggested by (Alatossava, Vaara and Baschong, 1984), and different concentrations of KCl (up to 170 $\mu$ M). M9 media was chosen since it contains a controlled  $K^+$  concentration ( $\sim$ 22mM); however, MOPS and MSgg minimal media were also tested without PMBN (data not shown). DMSO, used for dissolving valinomycin, reached a maximum concentration of 0.075% (v/v) during the experiments and did not affect cell viability (data not shown).



**Figure 4.11. Testing the potential induction of ECF/anti- $\sigma$  factor systems with valinomycin.** Cultures were grown in M9 minimal medium supplemented with 0.5% glycerol, 5ng/mL ATc and  $10^{-4}$  % (w/v) arabinose. Fluorescence intensity emitted by GFP (GFP) and OD at 600nm (OD<sub>600</sub>) were measured every 5min in a 10-hour time course experiment, by using a TECAN<sup>®</sup> Infinite 200 PRO plate reader. Induction with different reagents (legend) was done after two hours (gray line). The average fluorescence intensity in a wild-type strain without *gfp* (SV01) was used for blanking fluorescence, whereas the average OD<sub>600</sub> of M9 minimal medium was used for OD<sub>600</sub> blanking. Strains that contained reporter plasmid with either *P<sub>ter</sub>-gfp* or *P<sub>BAD</sub>-gfp* (GFC0013 and GFP0014) were used as controls of the activity of arabinose and ATc. A strain containing a reporter plasmid with *Pecf02\_2817-gfp* (GFC0419) was used as control for PMBN activity. The rest of the strains (GFC0414-GFC0418) contained *P<sub>BAD</sub>-ecf*, *P<sub>ter</sub>-as* integrated into HK022 *attB*, and *P<sub>ecf</sub>-gfp* in a reporter plasmid, as in Fig. 4.9. These results are based on only one replicate. The fluorescence detection limit was set to  $10^{-3}$  (AU).

Results of this experiment (Fig. 4.11) showed that strains containing *P<sub>ter</sub>-gfp* and *P<sub>BAD</sub>-gfp* were induced in similar order of magnitude as in MOPS minimal medium (Fig. 4.10), indicating that ATc and arabinose concentrations of 5ng/mL and  $10^{-4}$  % (w/v), obtained in MOPS minimal medium, could still be valid for M9 minimal medium. However, this should be verified in future experiments. A decrease in expression of *gfp* over time was observed for *P<sub>ter</sub>-gfp* strain, similarly as in MOPS minimal medium experiments (Fig. 4.10 and Fig. 4.11). This could be a consequence of the degradation of ATc over time in aqueous medium at 37°C (Politi *et al.*, 2014), although the loss of the plasmid containing *P<sub>ter</sub>-gfp* cannot be ruled out. The wild-type strain (SV01) displayed some GFP/OD values over background after 5 hours, likely due to blanking with its average fluorescence intensity instead of with the fluorescence intensity observed at each time point. GFP/OD values increased faster in the uninduced culture, probably because of the growth defects observed in cultures treated with inducers

(data not shown). As a positive control of PMBN, the strain containing  $P_{ecf02}$ -*gfp* displayed and increased GFP/OD intensity when PMBN was added to the medium (Fig. 4.11). Indeed, RpoE is induced by changes in LPS structure (Tam and Missiakas, 2005). The rest of the strains (GFC0414-GFC0418) did not show any increase of the GFP signal after addition of the inducers (Fig. 4.11). In general, these strains showed a reduced GFP/OD signal in the first part of the experiment, probably due to a fluorescence intensity below the detection limit at low cell density. The strain containing  $ecf28/as28$   $P_{ecf28}$ -*gfp* and, to a lesser extent, the strain containing  $ecf22/as22$   $P_{ecf22}$ -*gfp*, showed a reduction in GFP/OD values after induction (Fig. 4.11). This was more prominent for higher doses of KCl. One reason could be the reduced viability of  $ecf28/as28$   $P_{ecf28}$ -*gfp*-containing strain under 170mM KCl, although the viability of the strain that contains  $ecf22/as22$   $P_{ecf22}$ -*gfp* was only slightly affected under this treatment (data not shown). As previously observed (Alatossava, Vaara and Baschong, 1984), individual treatment with PMBN or valinomycin did not diminish viability in a large extent respect to the uninduced culture (data not shown). However, the combination of valinomycin and PMBN without external addition of KCl greatly reduced cell viability. This effect was generally reverted for higher  $K^+$  concentrations, with the exception of the strain that contained  $ecf28/as28$   $P_{ecf28}$ -*gfp* (data not shown). A reason could be that maximal  $K^+$  concentration after induction was  $\sim 190$ mM (170mM from the inducer and 22mM from the M9 minimal medium), whereas the intracellular concentration of  $K^+$  is  $\sim 210$ mM (Schultz and Solomon, 1961). Then, higher  $K^+$  concentrations in the inducer would be closer to intracellular concentration and would have a smaller impact on bacteria homeostasis after treatment with valinomycin and PMBN. However, this does not explain why the strain that contained  $ecf28/as28$   $P_{ecf28}$ -*gfp* showed a reduced viability after treatment with the highest  $K^+$  concentration.

All in all, these data indicate that the ECF/anti- $\sigma$  factor systems tested do not respond to voltage changes in *E. coli*. Further experiments carried out using MOPS and MSgg minimal media (without PMBN) led to the same conclusion (data not shown). Improvements of these experiments would verify whether  $10^{-4}$  % (w/v) arabinose and 5ng/mL ATc are the proper inducer concentrations to make ECF/anti- $\sigma$  factor systems inducible in M9 minimal medium. Moreover, time-point measurements need to be used to blank fluorescence, and several biological replicates of each experiment would need to be performed in order to confirm these results.

#### 4.7. Discussion and summary

In this section I address the binding principles that govern the contact between class I anti- $\sigma$  factors and ECF  $\sigma$  factors using a combination of phylogenetic and covariation-based methods, including DCA and SDP prediction. I first classified ASDIs into subgroups of closely related sequences, which were further hierarchically clustered. I observed a good correlation between ECF and ASDI classifications, supported by a Pearson correlation coefficient significantly larger than negative controls. This prompted me to define ASDI groups according to the ECF group of their cognate

partner. Subsequently, I focused on defining the positions that are important for the contact between both proteins. While DCA predicts co-varying pairs of residues that are likely to interact, SDPs are residues that characterize phylogenetic groups within families of proteins (de Juan, Pazos and Valencia, 2013). The combination of these two methods results in the identification of residues involved in protein-protein contacts that are characteristic of phylogenetic groups and could be involved in defining interaction specificity. I confirmed the physical proximity of the top 14 DCA predictions in the four crystal structures of ASDI/ECF complexes. Furthermore, two out of the five SDPs were also DCA predictions, confirming that some of the contacts between ECFs and anti- $\sigma$  factors are group-specific and partially explaining the functional cross-talk between ECFs and ASDIs of the same group.

One of the outcomes of this work is the first classification of ASDIs. The expansion of the ASDI dataset provided a full overview of the diversity of these proteins. These results show that ~32% of the ECFs are regulated by ASDIs, in agreement with previous reports (Campbell *et al.*, 2007). Moreover, the amount of zinc-binding motif-containing ASDIs remains ~40%. This expanded ASDI library has important differences in respect to the original ASDI extraction in that 1) it contains more cytoplasmic anti- $\sigma$  factors (~42% in this work respect to ~28% in the first ASDI extraction (Campbell *et al.*, 2007)), 2) cytoplasmic anti- $\sigma$  factors are not overrepresented in zinc-binding motifs, this is, 41% of the soluble anti- $\sigma$  factors are zinc-binding in this work, whereas this value is 92% in the first extraction (Campbell *et al.*, 2007), and 3) membrane-bound anti- $\sigma$  factors are not overrepresented in non-zinc binding proteins, this is, 48% of the transmembrane anti- $\sigma$  factors are non-zinc binding in this work, whereas this value is reduced to ~25% in the original ASDI library (Campbell *et al.*, 2007). These data suggest that ASDIs are more diverse than previously thought, and argues against a functional role of zinc-binding domain exclusively in soluble anti- $\sigma$  factors. This is supported by the ASDI tree, where zinc and non-zinc binding ASDI groups are mixed across the tree and sometimes even within the same group, as in the case of AS27, and AS19-1. It is tempting to speculate a structural role of the zinc-binding motif in these mixed zinc and non-zinc binding groups, as shown for RsiW from *B. subtilis* (group AS245) (Devkota *et al.*, 2017).

Analysis of DCA predictions and SDPs revealed a large interaction area between ASDI's helix 4 and ECF's helices 2.1 and 2.2. A secondary, albeit conserved within groups, binding interface between ASDI's helix 4 and ECF's helices 4.2 and 4.4 was also found. These data indicate a modular binding between ECFs and ASDIs, where ASDI's helix 1 binds to ECF's  $\sigma_4$  domain and ASDI's helix 4 binds to  $\sigma_2$  domain. This modularity of the ASDI interaction is reflected in the function of the ECF residues involved in the predictions. On one side, contacted residues in regions 2.1 and 2.2 are mostly involved in the contact with the clamp helices of the  $\beta'$  subunit of the RNAP (Lane and Darst, 2010b; L. Li *et al.*, 2019); on the other hand, predicted contacts in  $\sigma_4$  are part of the contact interface with the -35 element of the promoter (Lane and Darst, 2006; L. Li *et al.*, 2019). Even though this modular binding mode might be true in the overall ECF-ASDI complexes as an average, DCA predictions on the

contact between members of ECF26 and AS26 differ. On one hand, group 26 preserves predictions between ASDI's helix 1 and  $\sigma_4$  domain; however, predictions between ASDI's helix 4 and  $\sigma_2$  domain, the main contact interface when looking at the overall ECF-ASDI dataset, are missing. This shows that individual phylogenetic groups might have different binding conformations, as already seen in the four available crystal structures of the ECF-ASDI complex (Fig. 4.1).

ASDI helix 2 does not harbor any predicted contact with the ECF, as supported by the four crystal structures of ECF-ASDI complexes (Campbell *et al.*, 2003, 2007; Shukla *et al.*, 2014; Devkota *et al.*, 2017). In three out of four structures, ASDI's helix 3 is sandwiched between ECF's  $\sigma_2$  and  $\sigma_4$  domains (Campbell *et al.*, 2003, 2007; Shukla *et al.*, 2014). However, the SDPs predicted in this area contact the same area of the ECF's  $\sigma_4$  domain as the DCA prediction in helix 1 (Fig. 4.6B), arguing in favor of a more primordial binding to  $\sigma_4$  domain than to  $\sigma_2$  domain. Indeed, in one of the four ECF-ASDI structures (SigW-RsiW from *B. subtilis*, PDB: 5WUQ) only  $\sigma_4$  domain is contacted by ASDI helix 3 (Devkota *et al.*, 2017). DCA predictions on the interaction between members of ECF26 and AS26 revealed that 3 out the top 6 predictions are in helix 3 (Fig. 4.8A). Two of these predictions are plausible ( $<20\text{\AA}$ ) in most of the ECF-ASDI structures (Fig. 4.8B), and link ASDI's helix 3 to  $\sigma_4$  domain, supporting again that binding of this helix to  $\sigma_4$  domain is more important than binding to  $\sigma_2$  domain.

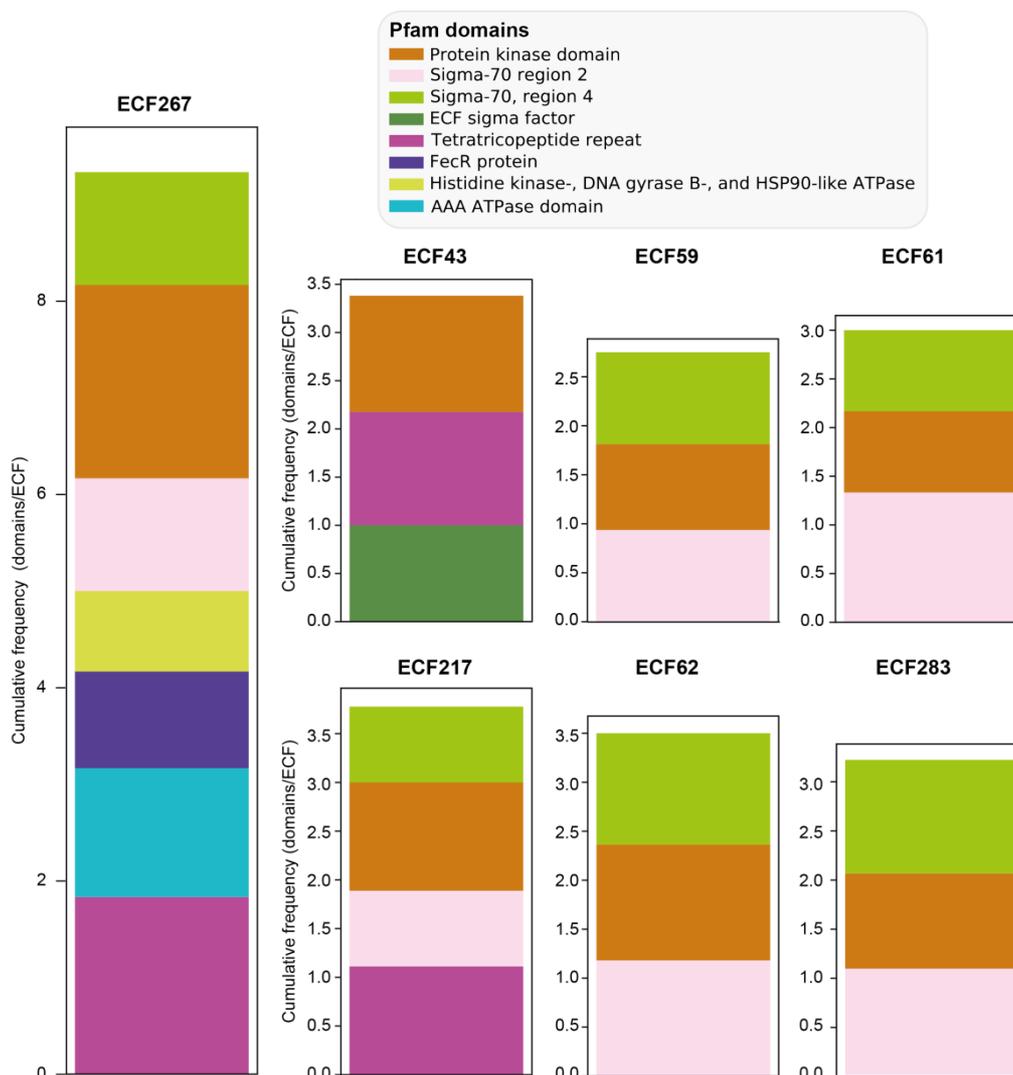
Amino acids involved in DCA predictions have different conservation levels across ASDI groups. Residues that take part in contacts between ASDI's helix 1 and ECF's  $\sigma_4$  (DCA predictions #10 and #11) are conserved for most of the groups. Interestingly, this area, which connects D11 on the ASDI (RseA<sub>*E.coli*</sub>) to R149 and R178 on the ECF (RpoE<sub>*E.coli*</sub>) bears two main types of interactions, hydrophobic or charged. Random mutagenesis in RseA<sub>*E.coli*</sub> (AS02) showed that mutation of D11 to histidine inhibits RseA activity (Missiakas *et al.*, 1997), confirming the importance of these contacts for the ASDI mechanism. Moreover, three out of the top 6 DCA predictions on the interaction between ECF26 and AS26 involved D11 (Fig. 4.8A). Given their group-specific conservation and the striking polarity differences between the two binding types, it is feasible to speculate that D11 defines coarse-grained specificity of ASDIs for ECFs of the same binding type, usually found in the same phylogenetic group. However, ASDIs are usually specific to their own target ECF and do not usually crosstalk with members of the same group (Rhodius *et al.*, 2013), indicating that there are more sources of specificity in residues that are not conserved within groups. One potential source of this specificity are the residues predicted by DCA in helix 4. These residues are generally not conserved within groups (Fig. 4.5) and binding the ECF's  $\sigma_2$  domain in all the solved crystal structures of ASDI-ECF complexes (Campbell *et al.*, 2003, 2007; Shukla *et al.*, 2014; Devkota *et al.*, 2017). This lack of major conservation is extended to the predicted contacts on the ECF side, which are generally in charge of binding to the  $\beta'$  subunit of the RNAP (Wilson and Lamont, 2006; Lane and Darst, 2010b; L. Li *et al.*, 2019).

This section focuses on class I anti- $\sigma$  factors, their co-evolution and interaction with ECFs. These results reveal that the binding between ASDIs and ECFs is modular in that 1) two distinct helices on the ASDI, helix 4 and helix 1, bind to each of the ECF's  $\sigma_2$  and  $\sigma_4$  domains, respectively, 2) these two regions have two separate functions, probably blocked by anti- $\sigma$  factor, namely, RNAP binding and DNA binding, respectively, and 3) the different level of conservation of the predicted interacting residues in both positions suggest the helix 1 determines group-dependent ASDI specificity, whereas helix 4 residues determine case-dependent specificity. Future experiments will test the importance of each of the predicted contacts. Moreover, our results support that crystal structures of the complex between ECF and ASDI of non-crystalized groups, such as group 26, have an alternative conformation. Even though ASDI's helix 3 seems to be more important in the inhibition of members of ECF26, the details of this interaction will be revealed by the resolution of their crystal structure.

## 5. ECF $\sigma$ factor phosphorylation

Among the conserved elements found in the genetic neighborhood of ECF  $\sigma$  factors in comparative genomic studies are Hanks-type serine/threonine kinases (STKs) (Section 1.4.2). This microsynteny between ECFs of certain groups and STKs was identified during the founding classification (ECF43) (Staroń *et al.*, 2009) and in following updates in Planctomycetes (ECF59, ECF60, ECF61 and ECF62) (Jogler *et al.*, 2012), and correlates with the lack of co-encoded anti- $\sigma$  factors (Fig. 3.8D). It has been hypothesized that the microsynteny between ECF  $\sigma$  factors and STKs could be a consequence of the regulation of STKs over ECF activity via phosphorylation, compensating the lack of anti- $\sigma$  factor (Mascher, 2013). This idea is strengthened with the ECF classification presented in Section 3, where a total of seven groups were found to be encoded in microsynteny with STKs (Fig. 5.1). Recent work by Bayer-Santos and colleagues revealed that EcfK, a member of ECF43, is required for the expression of a type 6 secretion system (T6SS) in *Xanthomonas citri*, and that this expression is dependent upon the STK PknS, encoded in the same genetic neighborhood (Bayer-Santos *et al.*, 2018). However, this work did not identify direct phosphorylation of PknS over EcfK.

This section aims at studying a possible phosphorylation in ECF  $\sigma$  factors, its functional role, its conservation and its evolutionary origins. In the first part, I focus on group ECF43, since this is the most abundant and taxonomically widespread group associated to STKs. One member of ECF43, EcfP from *Vibrio parahaemolyticus* (locus *vp0055*) was experimentally addressed as part of this study in experiments conducted by Dr. Shankar Chandrashekar Iyer and Dr. Simon Ringgaard, from the Max-Planck Institute for Terrestrial Microbiology, Marburg. *V. parahaemolyticus* is the major cause of seafood-borne gastroenteritis in humans worldwide (Letchumanan, Chan and Lee, 2014). Its reference strain contains five ECF  $\sigma$  factors from subgroups ECF02s1, ECF11s5, ECF28s2, ECF28s3, and ECF43 (EcfP). EcfP is encoded in the same operon as the STK PknT (locus *vp0057*) and other two genes of unknown function (*vp0056* and *vp0054*). Given that members of the ECF groups susceptible of phosphorylation, including EcfP, were not retrieved by the ECF extraction pipeline (Section 3), I expanded these groups and made predictions on their phosphorylation site. In the last part I speculate about the possible evolutionary origins of ECF  $\sigma$  factor phosphorylation.

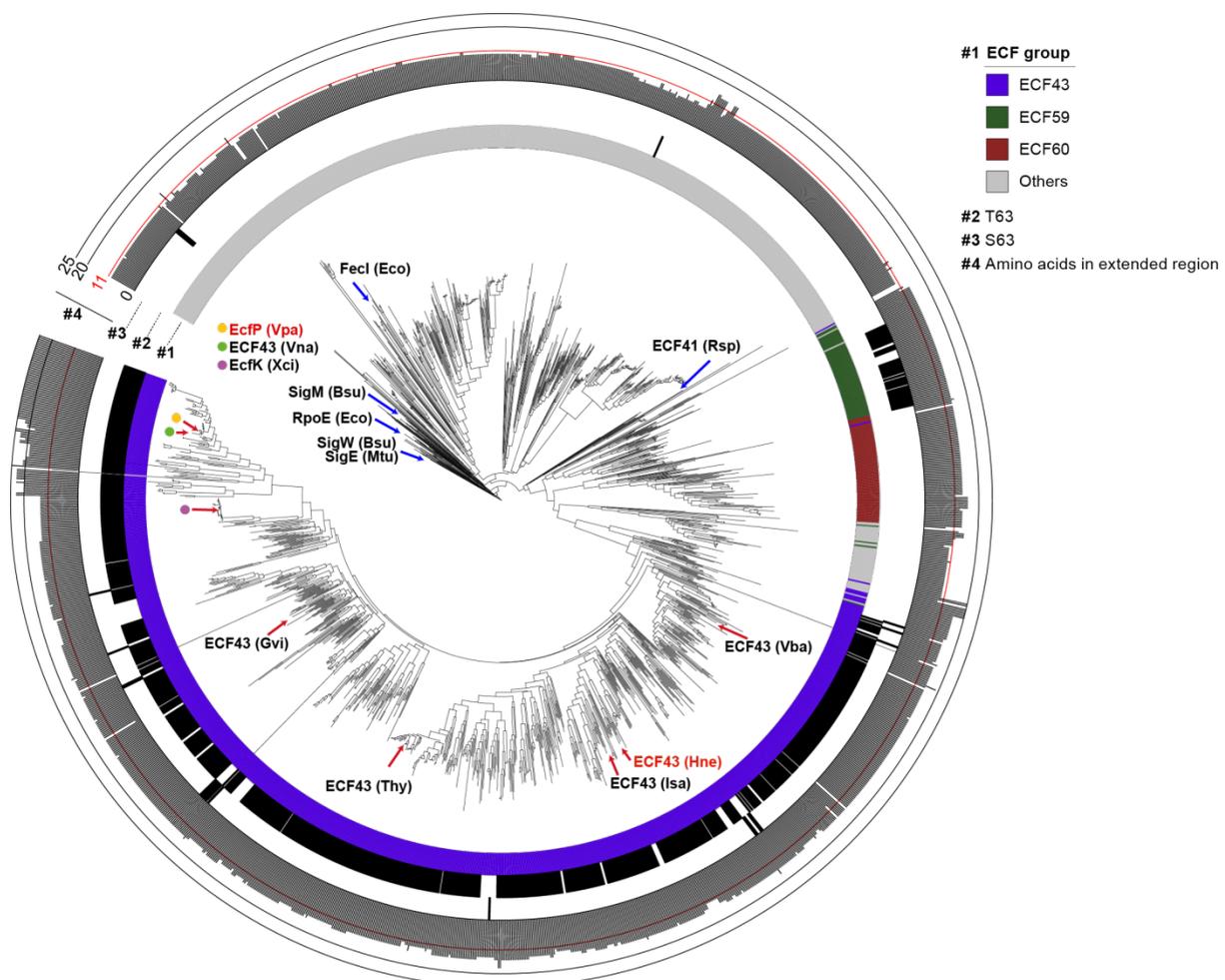


**Figure 5.1. Cumulative frequency of conserved domains in the genetic neighborhood of ECF  $\sigma$  factors from selected groups.** Conserved domains are defined as Pfam domains that are present in the genetic neighborhood ( $\pm 10$  coding sequences) of more than 75% of the ECFs for each group. Only groups with conserved STKs (Pfam accession: PF00069) are shown. Proteins from original group ECF60 were not classified by the pipeline since only eight members of ECF60 were extracted.

### 5.1. Members of ECF43 contain a deviant non-charged motif

I focused on group ECF43 since it is the largest STK-associated group. In the ECF classification presented in Section 3, members of ECF43 are exclusively present in Proteobacteria. However, notable members of this group, such as EcfP from *V. parahaemolyticus*, were not retrieved as part of the ECF expansion (Section 3.5). With the purpose of recovering the missing ECFs and analyzing the reason why they failed to be retrieved, I searched for proteins similar to EcfP encoded in the proximity of proteins similar to EcfP's kinase, PknT. First, I searched for proteins with sequence similarity to EcfP and PknT using online PSI-BLAST (E-value $<10$ ), and then I used the model built from the alignment of these sequences to search for matches in proteins encoded with less than 5Kbp in NCBI (version February 2017). This search yielded 1,603 ECF-STK pairs with less than 98% combined identity, this means that, taking together ECF and STK, their amino acid identity to any

other pair of the library is always smaller than 98% (Chandrashekar Iyer *et al.*, accepted) (Section 8.11).



**Figure 5.2. Phylogenetic tree of EcfP-like proteins.** Ring #1 represents the association of ECF  $\sigma$  factor to original groups groups. Ring #2 and #3 depict the presence of threonine or serine residues, respectively, in the position equivalent to T63 of EcfP. Ring #4 indicates the length of the region that connects  $\sigma_{2.1}$  and  $\sigma_{2.2}$ , enclosed between dashed lines in the multiple-sequence alignment of Fig 5.3A. Extended variants, present members of ECF43, ECF59 and ECF60, are those with length greater than 11. The closest homologs to EcfP contain longer extensions, of approx. 25 amino acids. Red arrows and dots indicate specific ECFs from ECF43. ECFs that are not regulated by STK and served as controls are indicated by blue arrows. The naming code for the species is as follows: Bsu = *Bacillus subtilis*, Eco = *Escherichia coli*, Gvi = *Gloeobacter violaceus*, Hne = *Hyphomonas neptunium*, Isa = *Ideonella sakaiensis*, Mtu = *Mycobacterium tuberculosis*, Rsp = *Rhodobacter sphaeroides*, Thy = *Thermomonas hydrothermalis*, Vba = *Verrucomicrobiaceae bacterium*, Vna = *Vibrio natriegens*, Vpa = *Vibrio parahaemolyticus* and Xci = *Xanthomonas citri*. Figure adapted from (Chandrashekar Iyer *et al.*, accepted).

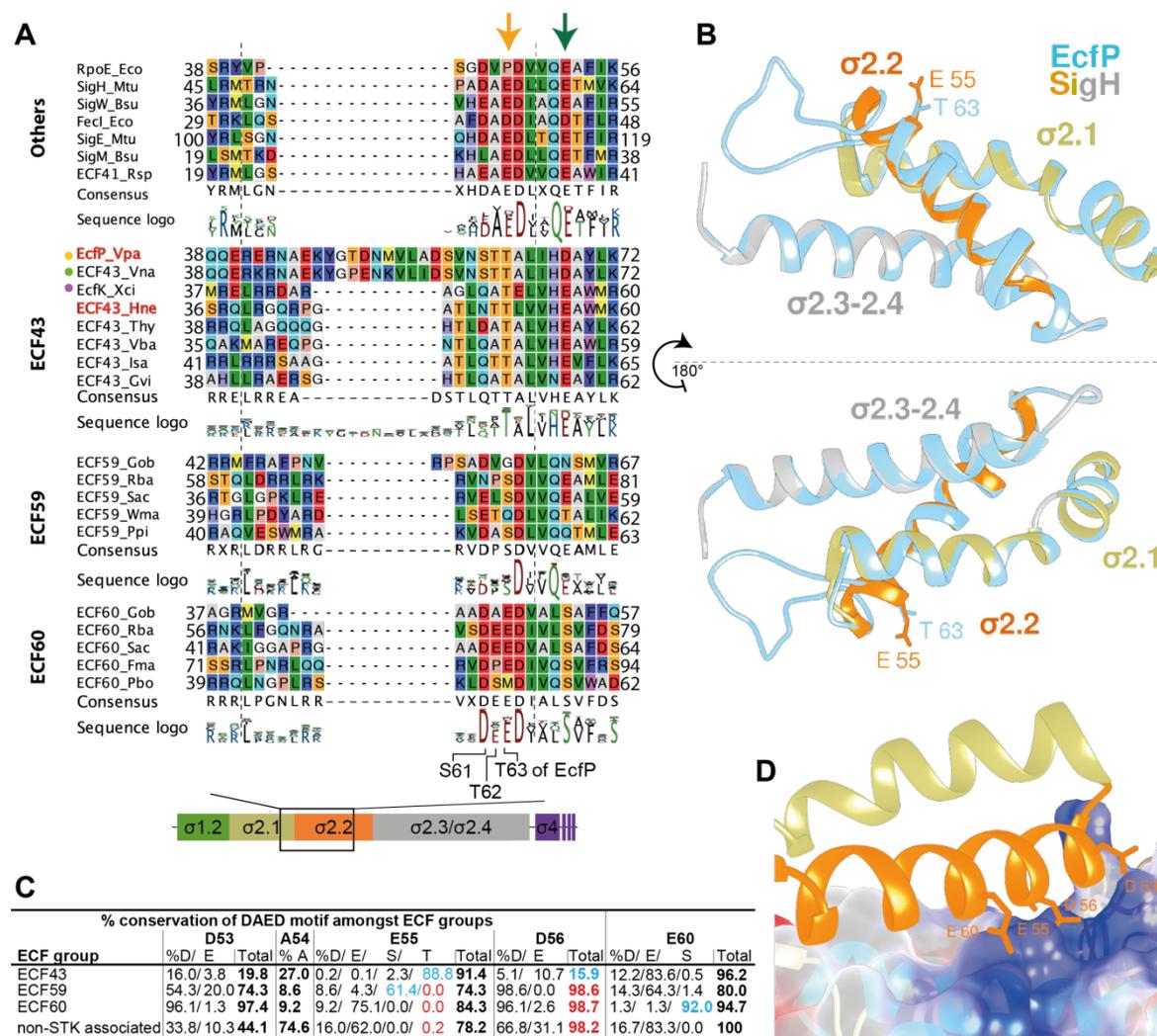
After the expansion, I built a maximum likelihood phylogenetic tree from the resulting ECF  $\sigma$  factors (Fig. 5.2). I included ECF sequences from other ECF groups aside from ECF43, as negative controls (Fig. 5.2, blue arrows). These ECFs are scattered across a region of the ECF tree. Assigning ECF sequences to original ECF groups (Section 3.5), I found that, indeed, the area of the tree where negative controls map is not composed of members of ECF43, but members of other ECF groups, preferentially ECF01 (12.23% of the ECFs) or ECF39 (8.30% of the ECFs) (Chandrashekar Iyer *et al.*, accepted). The STKs encoded in the proximity of these false positives is likely not controlling their activity, given that members of these groups are typically associated to anti- $\sigma$  factors (Staroń *et al.*, 2009) (Chandrashekar Iyer *et al.*, accepted). These ECFs will be called “non-STK associated”

ECFs. However, since original group ECF01 is diverse (Staroń *et al.*, 2009) (Section 3.3), a functional role of the STK in the activity of members of original ECF01 cannot be discarded. The main ECF group that composes the phylogenetic tree is ECF43 (53.15% of the ECFs) followed by ECF59 and ECF60 (4.37% and 4.68% of the ECFs, respectively), which form three monophyletic clades (Fig 5.2) (Chandrashekar Iyer *et al.*, accepted). EcfP from *V. parahaemolyticus* is indeed part of ECF43 clade (Fig. 5.2).

A multiple-sequence alignment (MSA) including members of ECF43 and the negative controls included in the phylogenetic tree revealed divergent areas in members of ECF43. Members of ECF43 contain an extended region between  $\sigma_{2.1}$  and  $\sigma_{2.2}$  that does not appear in negative controls (Fig. 5.3A, Fig. 5.2 ring #4) (Chandrashekar Iyer *et al.*, accepted). Negative controls contain an average of  $9.83 \pm 0.37$  (standard deviation) amino acids, while this is  $15.45 \pm 3.08$  amino acids for members of ECF43 (Fig. 5.2 ring #4) (Chandrashekar Iyer *et al.*, accepted). This region is particularly long ( $23.07 \pm 3.29$  amino acids) in the closest ECFs to EcfP, which are mainly present in Vibrionales and Alteromonadales (Fig. 5.2 ring #4). Interestingly, members of ECF59 and ECF60, with an average of  $13.91 \pm 0.53$  and  $13.96 \pm 1.76$  amino acids, also contain this extended region (Fig. 5.2 ring #4, Fig. 5.3A). Another observation derived from this MSA was that the first part of  $\sigma_{2.2}$  helix, located after the extended loop present in ECF groups linked to STKs, is divergent in members of ECF43. In canonical ECF  $\sigma$  factors, region  $\sigma_{2.2}$  is an  $\alpha$  helix that contains conserved negative amino acids with a DAED motif in its N-terminus (Fig. 5.3A) (Chandrashekar Iyer *et al.*, accepted). The negatively charged amino acids face the same side of the  $\alpha$  helix and contribute to the binding to the positive charges of the clamp helices in  $\beta'$  subunit of the RNAP (Fig. 5.3D). The DAED motif is replaced by a consensus QTT in ECF43 and, specifically, STT in EcfP (Fig. 5.3A). The last threonine of this motif, which corresponds to T63 in EcfP, is conserved in most of the members of ECF43 (88.8% of the members of ECF43) but harbors aspartate or glutamate in canonical ECFs (78% of the non-STK associated ECFs) (Fig. 5.3 A and C) (Chandrashekar Iyer *et al.*, accepted). Indeed, T63 is conserved across ECF43 clade of the phylogenetic tree (Fig. 5.2 ring #2) (Chandrashekar Iyer *et al.*, accepted).

In the structural alignment of a homology model of the structure of the  $\sigma_2$  domain of EcfP with SigH from *M. tuberculosis*, T63 overlays with E55 in SigH, which is in charge of binding to the  $\beta'$  subunit of the RNAP (PDB: 5ZX2, (L. Li *et al.*, 2019)) (Fig. 5.3B). Even though T63 is conserved only in ECF43, members of ECF59 contain a serine in the same position in 61.4% of the sequences (Fig. 5.3 A and C, Fig. 5.2 ring #3). Instead, 92% of the members of ECF60 replace with serine another negative charge from  $\sigma_{2.2}$ , which corresponds to E60 in SigH from *M. tuberculosis* and is conserved in 95.8% of the canonical ECFs (Fig. 5.3 A and C) (Chandrashekar Iyer *et al.*, accepted). E60 faces the same side of the ECF as E55 and is also involved in binding to the  $\beta'$  subunit of the RNAP (PDB: 5ZX2, (L. Li *et al.*, 2019)) (Fig. 5.3D). Given the absence of important negative charges required for the contact with the clamp helices  $\beta'$  subunit of the RNAP, the conservation of the STK and the absence of any other putative regulator of ECF  $\sigma$  factor activity, it is possible that 1) phosphorylation

compensates for the lack of negative charges and enhances the binding of members of ECF43 to the RNA polymerase, and 2) this phosphorylation is triggered by extracellular cues (Chandrashekar Iyer *et al.*, accepted).



**Figure 5.3. Sequence analysis of members of ECF43.** A: Multiple-sequence alignment of the region between  $\sigma_{2.1}$  and  $\sigma_{2.2}$  with the associated sequence logo of the full set of proteins associated to each group. The conserved DAED motif present in controls is replaced by non-charged residues in members of ECF43, and partially replaced in ECF59. T63 in *V. parahaemolyticus* (orange arrow) is conserved in members of ECF43, phosphorylation of this residue was studied in EcfP and ECF43\_Hne (also named HNE1495) (red labels). The conserved residue with a negative charge in the position of E60 in SigH from *M. tuberculosis* (green arrow) is substituted by a serine in ECF60. Stripped lines indicate the region used for the calculation of the extended  $\sigma_{2.1}$ - $\sigma_{2.2}$  region in Fig. 5.2. The species abbreviation goes as follows: Bsu = *Bacillus subtilis*, Eco = *Escherichia coli*, Fma = *Fuerstia marisgermanicae*, Gob = *Gemmata obscuriglobus*, Gvi = *Gloeobacter violaceus*, Hne = *Hyphomonas neptunium*, Isa = *Ideonella sakaiensis*, Mtu = *Mycobacterium tuberculosis*, Pbo = *Paludisphaera borealis*, Ppi = *Planctomicrobium piriforme*, Rba = *Rhodopirellula baltica*, Rsp = *Rhodobacter sphaeroides*, Sac = *Singulisphaera acidiphila*, Thy = *Thermomonas hydrothermalis*, Vba = *Verrucomicrobiaceae bacterium*, Vna = *Vibrio natriegens*, Vpa = *Vibrio parahaemolyticus*, Wma = *Wenzhouxiangella marina* and Xci = *Xanthomonas citri*. B: homology model of the  $\sigma_2$  domain from EcfP (blue) overlaid to the  $\sigma_2$  domain from SigH in the RNA polymerase open complex of *M. tuberculosis* (PDB: 5ZX2 (L. Li *et al.*, 2019)). EcfP structure was modelled with Swiss-model (Waterhouse *et al.*, 2018) using SigH as template (PDB: 5ZX2 (L. Li *et al.*, 2019)). E55 from SigH (orange) has the same orientation as T63 (blue) in EcfP, suggesting that T63 is lays near the positively charged surface of the clamp helices of the  $\beta'$  subunit. C: conservation of the DAED motif and other negatively charged residues of  $\sigma_{2.2}$  responsible for mediating the interaction with the  $\beta'$  subunit of the core RNAP enzyme. Amino acid coordinates refer to SigH from *M. tuberculosis*. D: binding of SigH ( $\sigma_{2.1}$ -beige,  $\sigma_{2.2}$ -orange) from *M. tuberculosis* to the clamp helices (light blue) of the  $\beta'$  subunit of the RNA polymerase (PDB: 5ZX2 (L. Li *et al.*, 2019)). Negative residues from  $\sigma_{2.2}$  are shown and labeled. The surface of the  $\beta'$  subunit is colored according to electrostatic charge, where red is negative and blue is positive charge. Figure adapted from (Chandrashekar Iyer *et al.*, accepted).

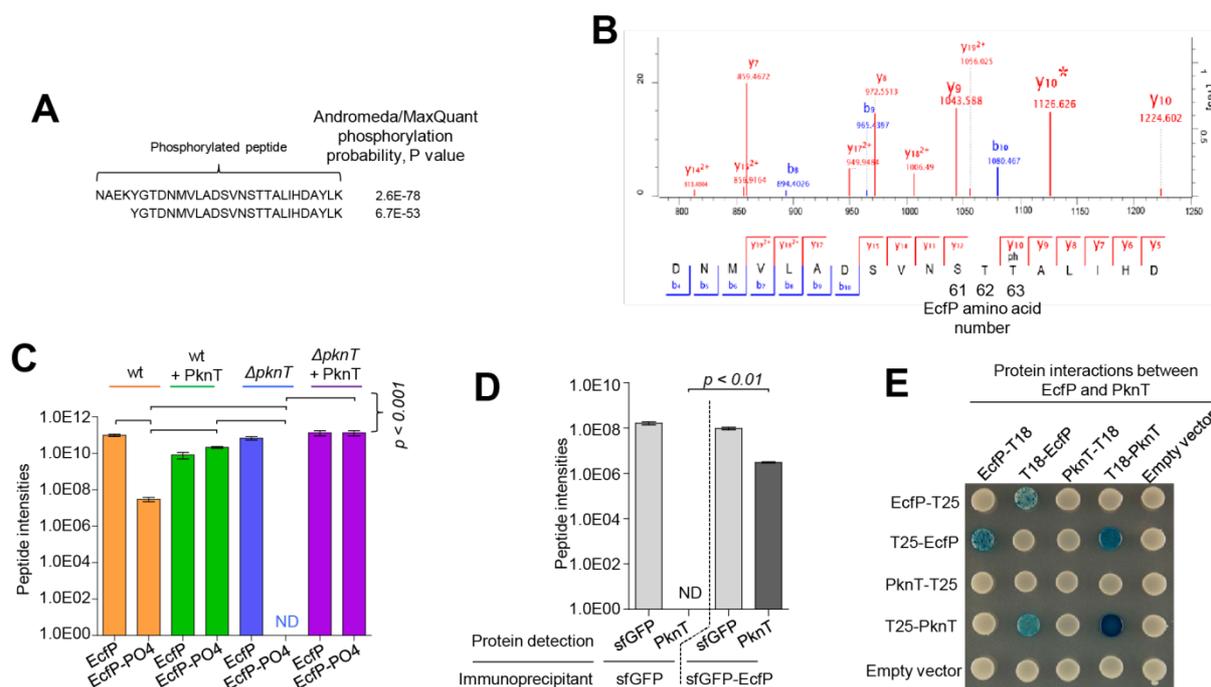
## 5.2. EcfP phosphorylation in T63 is required for RNA polymerase binding

EcfP is the only member of ECF43 in *V. parahaemolyticus*, which allows for the isolated study of its function, without a possible crosstalk by any other member of this group. Experimental work to assess the phosphorylation of EcfP was carried out by Dr. Shankar Chandrashekar Iyer and Dr. Simon Ringgaard (Max-Planck Institute for Terrestrial Microbiology, Marburg). EcfP phosphorylation was assessed by affinity purification of sfGFP (superfolder Green Fluorescent Protein)-tagged EcfP, ectopically expressed in wild-type cells, followed by MS/MS (tandem mass spectrometry) analysis. Two overlapping phosphorylated peptides of EcfP were detected (Fig. 5.4A), showing that, indeed, EcfP is phosphorylated *in vivo*. Importantly, both these peptides contain T63, suggested to be target of phosphorylation in the comparative genomic study of Section 5.1 (Fig. 5.4A) (Chandrashekar Iyer *et al.*, accepted). Further MS/MS analysis showed that the detected phospho-peptides were phosphorylated in T63 (Fig. 5.4B) (Chandrashekar Iyer *et al.*, accepted).

Overexpression of PknT increased the fraction of phospho-EcfP 100-fold, reaching 50% of the total EcfP (Fig. 5.4C). However, deletion of PknT abolished EcfP phosphorylation, which was recovered after ectopic expression of PknT (Fig. 5.4C) (Chandrashekar Iyer *et al.*, accepted). This showed that phosphorylation of EcfP is dependent on PknT. Direct interaction of EcfP by PknT was further supported by two lines of evidence (Chandrashekar Iyer *et al.*, accepted). First, PknT specifically co-immunoprecipitated with sfGFP-EcfP, but not with sfGFP alone, using wild-type cells that expressed either sfGFP or sfGFP-EcfP and were analyzed by LC-MS (liquid chromatography coupled with mass spectrometry) (Fig. 5.4D). Second, EcfP and PknT showed positive interaction in bacteria-two-hybrid (BTH) assays (Fig. 5.4E). This assay also revealed that EcfP may self-interact (Fig. 5.4E). It is important to point out that these experiments were done in standard laboratory conditions, without input stimulus.

In order to find the response mechanism mediated by EcfP and PknT, deletion mutants were cultured under the present of different stressors, and their viability was measured (Chandrashekar Iyer *et al.*, accepted). Deletion of *ecfP* and *pknT* separately yielded *V. parahaemolyticus* strains highly sensitive to polymyxin antibiotics (polymyxin B and E) (Chandrashekar Iyer *et al.*, accepted). These strains reduced their colony-forming units (CFU) by  $10^7$ -fold compared to wild-type (Fig. 5.5A). However, no phenotype was found for *vp0054* nor *vp0056* deletion mutants (Chandrashekar Iyer *et al.*, accepted). Polymyxins are cationic antimicrobial peptides that primarily target lipopolysaccharides (LPSs), breaking up the Gram negative outer cell membrane (Velkov *et al.*, 2013). Interestingly, the phosphoablative mutant, *ecfPT63A*, was highly sensitive to polymyxin B, in a similar manner as  $\Delta$ *ecfP* (Fig. 5.5B), indicating that T63 and its ability to be phosphorylated are essential for the function of EcfP (Chandrashekar Iyer *et al.*, accepted). The viability of a phosphomimetic version of EcfP, *ecfPT63E* exhibits a ~100-fold increase in the survival compared to  $\Delta$ *ecfP* strain in the presence of polymyxin B (Fig. 5.5C). Therefore, the mutation of T63 by a negatively charged amino acid is

able to partially complement the lack of phosphorylation, essentially working as a phosphomimetic variant (Chandrashekar Iyer *et al.*, accepted).

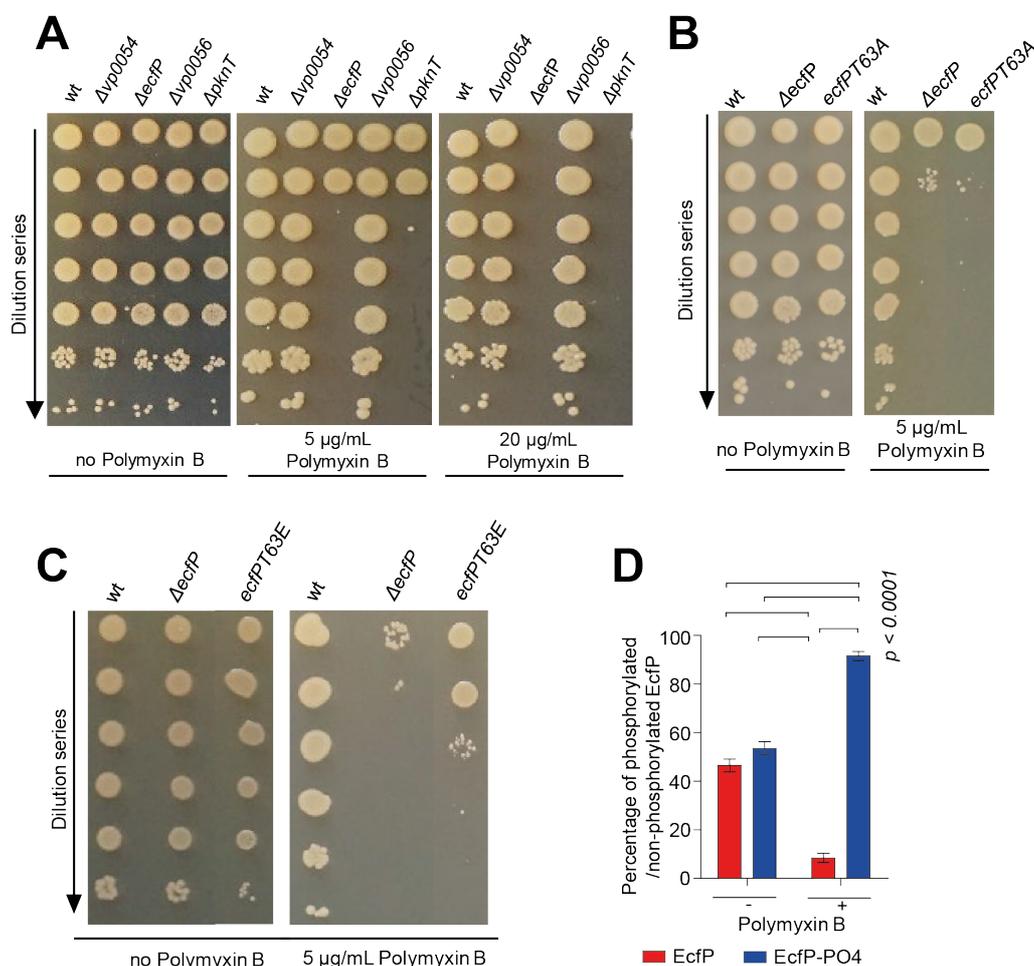


**Figure 5.4. EcfP  $\sigma$  factor phosphorylation at T63 by threonine kinase PknT.** **A:** sequences of the detected phosphorylated EcfP peptides with their corresponding phosphorylation probability as calculated by Andromeda embedded in MaxQuant. **B:** MS/MS spectrum of the mapped phosphorylation site in the phosphopeptides of ECF  $\sigma$  factor EcfP. The identified b- and y-fragment ions are shown in blue and red, respectively. The y-ion series allowed for the identification of the phosphorylation modification. For the y10 ion, a neutral loss of the phosphate (H3PO4-98Da) was detected, showing that T63 of EcfP is phosphorylated. The neutral loss peak is marked with an asterisk, y10\*. **C:** label-free, LC-MS quantification of co-immunoprecipitated sfGFP-EcfP in different *V. parahaemolyticus* strains. Bars show summed peptide intensities of unphosphorylated EcfP and phosphorylated EcfP peptides in wild-type cells (orange), wild-type cells with ectopic overexpression of PknT (green), in a  $\Delta pknT$  background (blue), and in a  $\Delta pknT$  background with ectopic expression of PknT to test for complementation of the phenotype (purple). **D:** bar plot showing summed protein peptide intensities of a LC-MS after co-immunoprecipitation using beads with attached anti-GFP antibodies on wild-type *V. parahaemolyticus* cells expressing sfGFP (negative control) or sfGFP-EcfP, respectively. The assay shows that PknT is significantly co-immunoprecipitated with sfGFP-EcfP but not sfGFP alone. **E:** bacterial-two-hybrid assay testing for protein-protein interaction between EcfP and PknT. Blue colony formation suggests that a direct interaction occurs. All the above are based on three biological replicates (per condition or sample group). Error bars indicate SEM and P-values were calculated by Student's t-test. "ND": non-detected. Figure adapted from (Chandrashekar Iyer *et al.*, accepted).

With the aim of studying whether PknT phosphorylates EcfP in the presence of polymyxin B, Phos-tag gel analysis was performed (Chandrashekar Iyer *et al.*, accepted). For that, wild-type cells ectopically expressing PknT in the presence and amount of phospho-EcfP respect to the absence of polymyxin B (Fig. 5.5D), supporting that polymyxin B triggers the phosphorylation of EcfP (Chandrashekar Iyer *et al.*, accepted).

The functional role of T63 phosphorylation could be related to binding to RNAP binding, as described in Section 5.1 and (Chandrashekar Iyer *et al.*, accepted). In this model, the lack of negative charges due to the lack of a DAED motif in the  $\sigma_{2.2}$  of members of ECF43 is compensated by the phosphorylation of T63, which faces the RNAP in a homology model of EcfP (Fig. 5.3D) (Chandrashekar Iyer *et al.*, accepted). In order to test this hypothesis, sfGFP-EcfP was expressed in *E. coli* in the presence or absence of PknT. Then, sfGFP-EcfP was co-immunoprecipitated and tested for the presence of  $\beta/\beta'$  subunits of the RNAP by Western blot (Chandrashekar Iyer *et al.*, accepted). *E.*

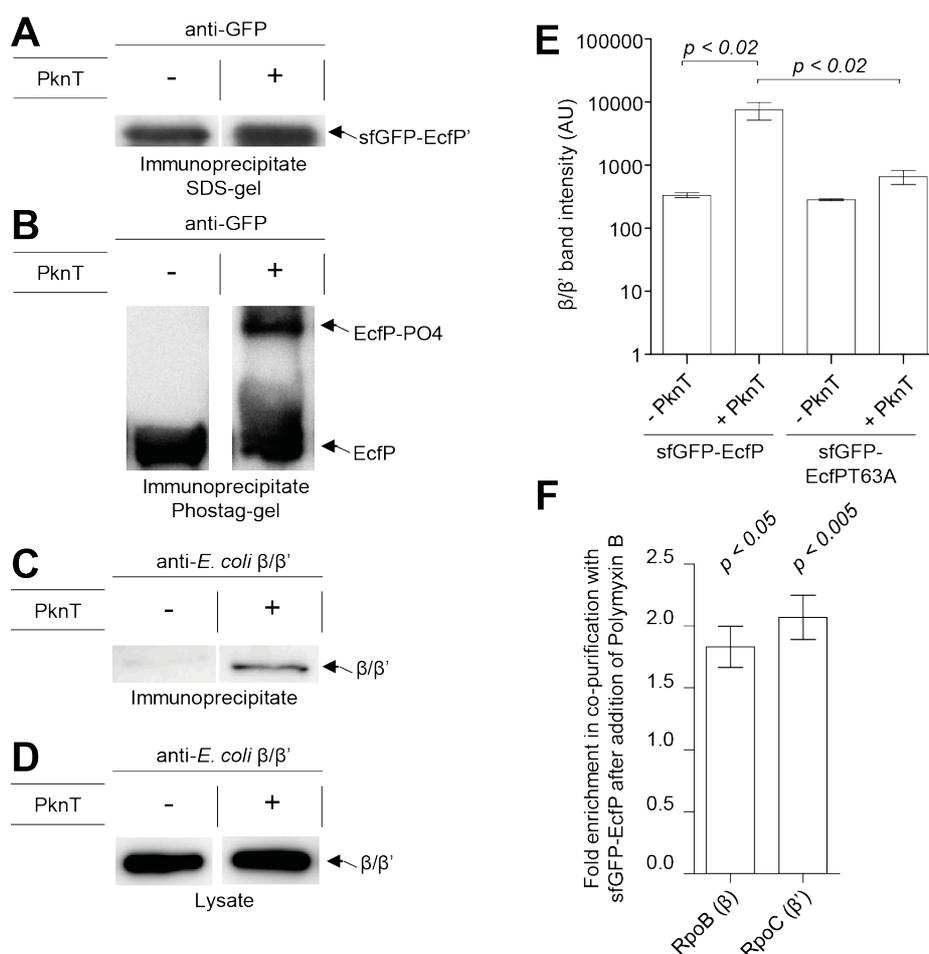
*coli* is used in this experiment since antibodies against its  $\beta/\beta'$  subunits were commercially available. Furthermore, *E. coli* allows for the activity of heterologous ECF  $\sigma$  factors (Rhodius *et al.*, 2013), indicating that they successfully binding to *E. coli*'s RNAP. Controls show that sfGFP-EcfP is present in the same amount regardless of PknT (Fig. 5.6A). However, phospho-EcfP is only detected in the presence of PknT (Fig. 5.6B) (Chandrashekar Iyer *et al.*, accepted). Moreover, whereas the amount of  $\beta/\beta'$  does not change in the cell lysate upon PknT deletion (Fig. 5.6D), the sfGFP-EcfP immunoprecipitate shows  $\sim 20$ -fold more  $\beta/\beta'$  co-immunoprecipitation when PknT is present (Fig. 5.6C). This shows that 1) EcfP is only phosphorylated in the presence of PknT in a heterologous system, 2) the amount of  $\beta/\beta'$  co-immunoprecipitated with EcfP is larger when PknT is present (Chandrashekar Iyer *et al.*, accepted).



**Figure 5.5. EcfP and PknT regulate polymyxin B resistance of *Vibrio parahaemolyticus*.** A-C: spot dilution assay of *V. parahaemolyticus* wild-type and mutant variants on LB growth medium in the presence and absence of polymyxin B. A: both *ecfP* and *pknT* deletion mutants, but neither *vp0054* nor *vp0056* are sensitive to polymyxin B. B: *ecfPT63A* mutant shows the same polymyxin B sensitivity as *ecfP* deletion mutant. C: *ecfPT63E* phosphomimic mutant is partially able to recover polymyxin B resistance. D: bar plot showing the percentage of phosphorylated and non-phosphorylated sfGFP-EcfP in the presence or absence of polymyxin B in *V. parahaemolyticus* after a Western blot on Phos-tag gel using anti-GFP antibodies. Three biological replicates were performed and analyzed for all the above panels. Error bars indicate SEM and P-values were calculated by Student's t-test. Figure adapted from (Chandrashekar Iyer *et al.*, accepted).

To further check that the phosphorylation of EcfP is triggering its binding to RNAP, the same experiment was done with the *ecfPT63A* mutant, which cannot be phosphorylated by PknT

(Chandrashekar Iyer *et al.*, accepted). In this case, the levels of co-immunoprecipitated  $\beta/\beta'$  subunit are not significantly different in the presence or absence of PknT, and  $\beta/\beta'$  band intensity is equivalent to the one of sfGFP-EcfP in the absence of PknT (Fig. 5.6E) (Chandrashekar Iyer *et al.*, accepted). The ~2-fold enrichment in the co-purification of  $\beta$  and  $\beta'$  subunits of RNAP with sfGFP-EcfP after polymyxin B addition (Fig. 4.6F) supports that EcfP responds to this antibiotic by an increased binding to the RNAP (Chandrashekar Iyer *et al.*, accepted).



**Figure 5.6. Phosphorylation of EcfP at T63 enhances its interaction with  $\beta/\beta'$  subunit of RNAP.** A-D: Analysis for co-immunoprecipitation of *E. coli*  $\beta/\beta'$  with sfGFP-EcfP in the absence or presence of PknT. A-B: Western blot of the immunoprecipitate analyzed on (A) an SDS gel and (B) a Phos-tag gel using anti-GFP antibodies in the presence or absence of PknT. C: Western blot of the immunoprecipitate using anti-*E. coli*  $\beta/\beta'$  antibodies in the presence or absence of PknT. D: Western blot of cleared lysates using anti-*E. coli*  $\beta/\beta'$  antibodies in the presence or absence of PknT. E: Bar plot showing the quantification of the amount of *E. coli*  $\beta/\beta'$  in immunoprecipitated samples of sfGFP-EcfP and sfGFP-EcfPT63A, respectively, in the absence or presence of PknT. F: Bar plot showing the fold enrichment in co-purification of  $\beta$  and  $\beta'$  subunits of RNAP with sfGFP-EcfP after addition of polymyxin B. Three biological replicates were performed for each experiment. Error bars indicate SEM and P-values were calculated by Student's t-test. Figure adapted from (Chandrashekar Iyer *et al.*, accepted).

These results, kindly provided by Dr. Shankar Chandrashekar Iyer and Dr. Simon Ringgaard (Max-Planck Institute for Terrestrial Microbiology, Marburg), show that, indeed, the lack of DAED motif at  $\sigma_{2.2}$ , is compensated by phosphorylation of T63 in EcfP from *V. parahaemolyticus*. The phosphorylation of EcfP, triggered by polymyxin antibiotics and mediated by the STK PknT,

promotes EcfP binding to the RNAP and the response to polymyxin stress. This is the first time that ECF phosphorylation is found to be part of a bacterial signal transduction mechanism.

### 5.3. Members of ECF43 are widespread across bacteria

In the ECF expansion presented in Section 3.1, the number of members of ECF43 expanded from 36 to 69 unique proteins. These 69 proteins are only present in Alpha, Beta and Gammaproteobacteria. Thanks to the retrieval and analysis of EcfP homologs, I expanded the number of members of ECF43 to 931 unique protein sequences (see Section 8.11 for details in the procedure). These proteins can be found in 13 bacteria phyla, including Proteobacteria (83.73% of the members of ECF43), Acidobacteria (5.86%) and Planctomycetes (5.13%). However, the number of phyla that contain an EcfP homolog with T63 or S63 in reference/representative organisms (as defined by NCBI, where RefSeq assemblies are preferred over GenBank if both exist), is reduced to only 8 phyla (Fig. 5.7). Since these sequences contain an extended region between  $\sigma_{2.1}$  and  $\sigma_{2.2}$ , are usually members of ECF43, contain the conserved T63 in most the cases, and are encoded near a STK, it is likely that they are regulated via phosphorylation in a similar manner as EcfP in *V. parahaemolyticus*. Indeed, the residue equivalent to T63 of EcfP was found to be phosphorylated in another member of ECF43 from *Hyphomonas neptunium* (locus HNE1495) (data not shown), named Ecf43\_Hne in Figures 5.2 and 5.3A (data from Dr. Shankar Chandrashekar Iyer and Dr. Simon Ringgaard, from the Max-Planck Institute for Terrestrial Microbiology, Marburg). These results show that ECF phosphorylation is a mechanism present in distantly related Gamma and Alphaproteobacteria, arguing for the phosphorylation of other members of group ECF43. Phosphorylation occurs on equivalent residues – T63 in *V. parahaemolyticus* and T51 in *H. neptunium* – suggesting a common response mechanism, where the negative charge added by the phosphorylation allows the ECF to bind to the RNAP and activate downstream responses.

*V. parahaemolyticus* uses EcfP to cope with polymyxins, which target LPS in the outer membrane. Therefore, their presence needs to be sensed by an extracytoplasmic or transmembrane protein. About 90% of the STKs associated to members of ECF43 are transmembrane (data not shown). Therefore, it is likely that these transmembrane STKs sense the extracytoplasmic cue and, in turn, activate their associated ECF  $\sigma$  factor to respond to the stimulus. STKs associated to members of ECF43 presented in Section 3 contain extracytoplasmic areas with tetratricopeptide repeats (TPRs) (data not shown). TPRs mediate protein-protein interactions and occur ubiquitously across Bacteria and Eukaryotes (Cervený *et al.*, 2013). I extended this study to the STKs associated to members of the extended ECF43 introduced in this section. This would help to reveal whether response to polymyxin antibiotics is conserved across ECF43 or members of ECF43 rather respond to different input signals. Not surprisingly, most of the STKs associated to members of ECF43 contain TPRs in their C-terminal region, corresponding to their extracytoplasmic area (Fig. 5.8). However, this is not the case for most of the proteins from Vibrionales, where the kinase exhibits a shorter extracytoplasmic region that does

not hold any Pfam domain (Fig. 5.8). Indeed, the extracytoplasmic area of PknT is only ~100aa long. Therefore, response to polymyxin antibiotics could be conserved in close homologs of EcfP from Vibrionales.

Aside from EcfP in *V. parahaemolyticus*, the only other member of ECF43 with an assigned function is EcfK from *X. citri* (encoded in locus *xac4128*). This ECF is responsible of resistance to *Dictyostelium* predation through the activation of a T6SS (Bayer-Santos *et al.*, 2018). The STK associated to this system, PknS (encoded in locus *xac4127*), contains a long extracytoplasmic region with several TPRs. Given that 1) the extracytoplasmic domains vary across the STKs associated to members of ECF43, and that 2) the function of EcfP (*V. parahaemolyticus*) and EcfK (*X. citri*) differs, it seems that different members of ECF43 specialized in different types of extracytoplasmic stress. However, a similar input signal across members of ECF43 cannot be discarded since their STKs do not necessarily bind to the input cue directly, and instead, they might bind to intermediate proteins that, in turn, sense a similar input signal. In favor of this idea, resistance mechanism against polymyxins and predation both involve LPS modifications (Velkov *et al.*, 2013; Duncan *et al.*, 2018). In the case of anti- $\sigma$  factors, there are several examples where triggering molecules do not bind directly to the anti- $\sigma$  factor, but to other upstream proteins. For instance, RseA, the anti- $\sigma$  factor of RpoE in *E. coli*, is not the primary sensor of C-terminal regions of unfolded OMP, but its site-1 protease, DegS (Section 1.4.1.1). In the case of FecR, the anti- $\sigma$  factor from FecI in *E. coli*, the sensing of ferric-citrate is done by FecA, its outer membrane transporter, which then transfers the signal to FecR (Section 1.4.1.4).

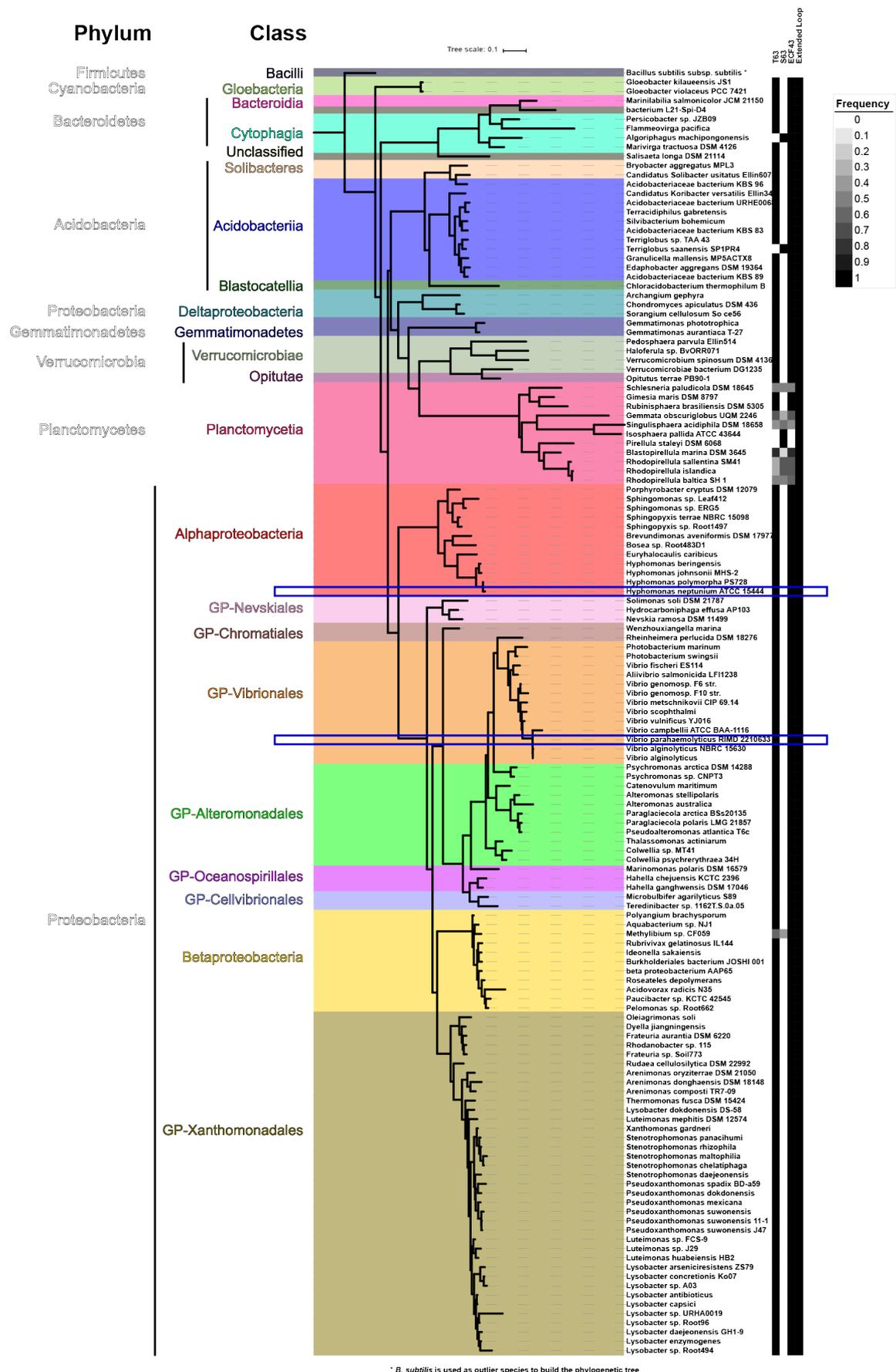
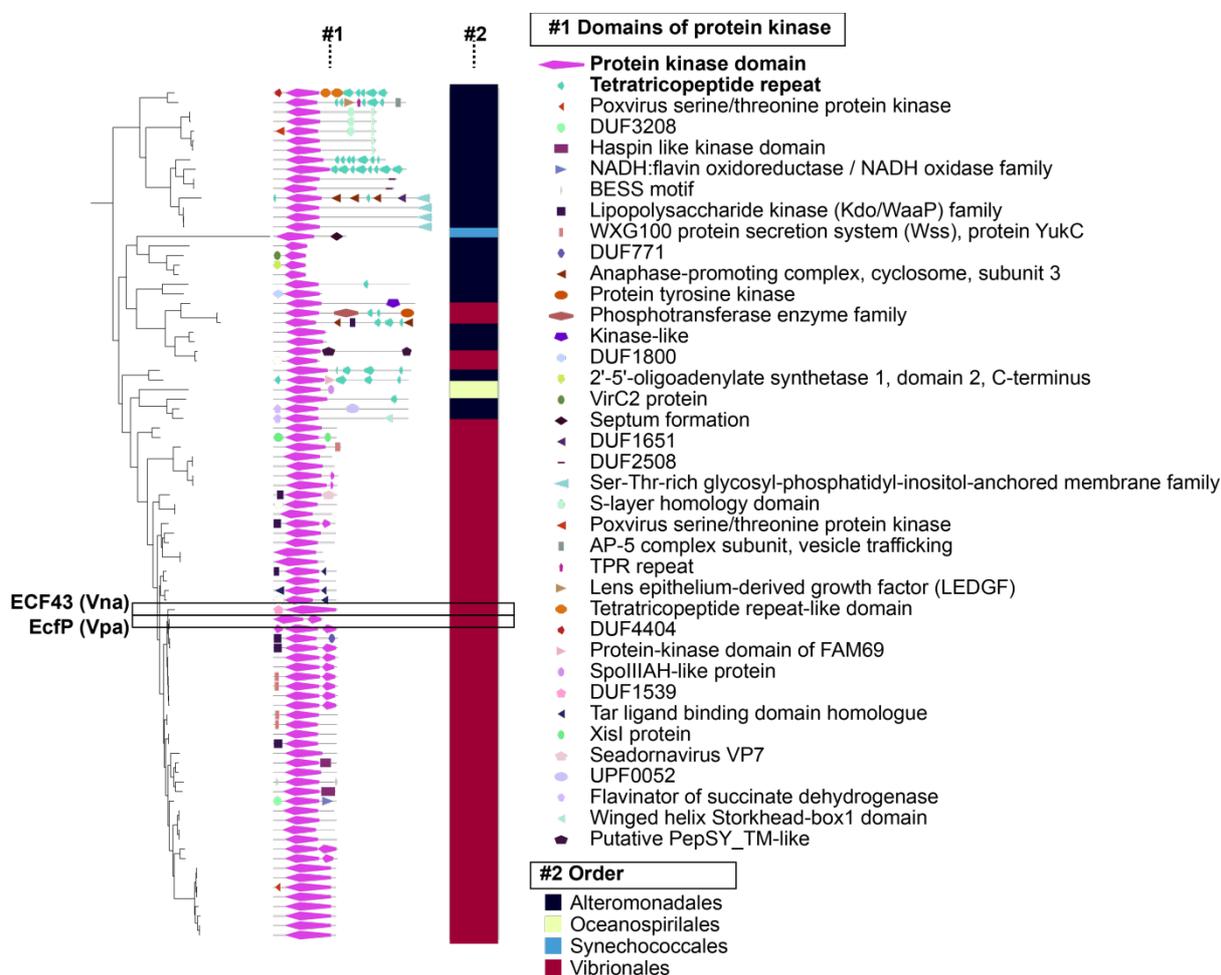


Figure 5.7. Phylogenetic tree of species encoding close EcfP homologs in their genomes. The branches of the tree represent the phylogenetic distance of the 16S rDNA sequences of representative and reference organisms that containing an ECF  $\sigma$  factor with Ser or Thr

in a position equivalent to T63 in EcfP of *V. parahaemolyticus* in the library of EcfP homologs. Representative and reference organism are as defined by NCBI, where RefSeq genomes are preferred over GenBank assemblies. Taxonomic class and phylum are indicated by the shades of the tree and the accompanying labels. The tree was rooted in the 16S rDNA of *Bacillus subtilis*, used as outlier since Firmicutes do not contain any ECF with Ser/Thr63. The heatmap on the right shows the frequency of ECF variants with S63, T63, with an extended region between  $\sigma_{2.1}$  and  $\sigma_{2.2}$  and members of group ECF43. Frequencies identical to 1 (=100%) indicate that all EcfP homologs in this species exhibit the respective feature, while frequencies below 1 indicate that there exist other EcfP homologs in the species, which do not exhibit the given feature. Blue frames indicate PknT from *V. parahaemolyticus* and STK\_Hne from *H. neptunium*. Figure adapted from (Chandrashekar Iyer *et al.*, accepted).



**Figure 5.8. Different members of ECF43 are associated to distinct protein kinase domain architectures.** The phylogenetic tree on the left is a fragment of the ECF tree of the homologs of EcfP show in Fig. 5.2. For simplicity, only the closest variants to EcfP from *Vibrio parahaemolyticus* (Vpa) are shown. Column #1 depicts the Pfam domains found in the STK associated to each ECF and their position. Column #2 shows the taxonomic order of the organisms of origin of the proteins. The extracytoplasmic domains, located in C-termini from the protein kinase domain, vary across protein kinases, potentially suggesting that STKs respond to different stimuli or have different sensing mechanisms. Figure adapted from (Chandrashekar Iyer *et al.*, accepted).

#### 5.4. ECF phosphorylation could be possible in other ECF groups

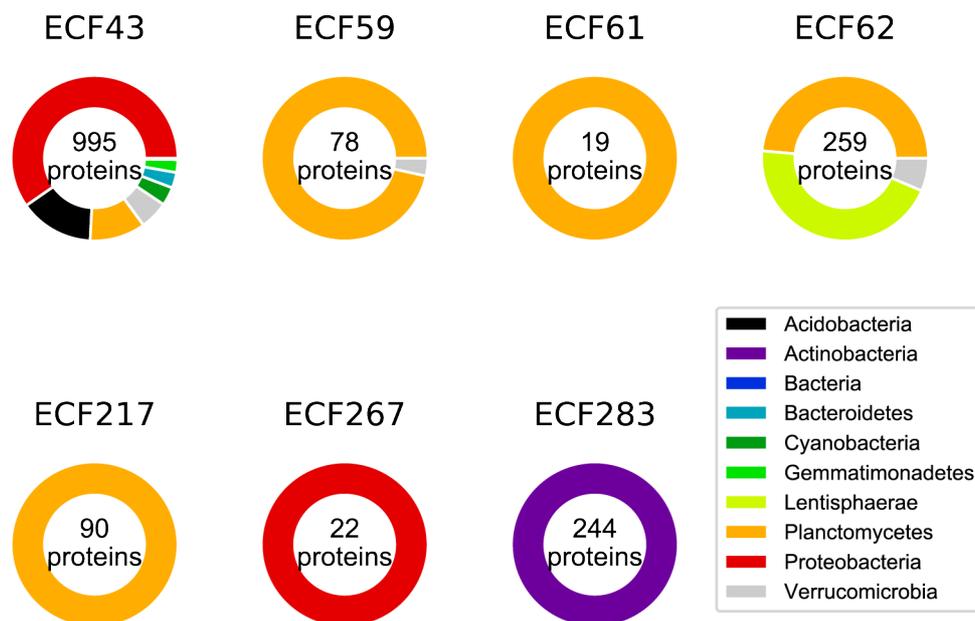
Given the success of the *in silico* approach to predict the phosphosite in group ECF43, I decided to extract more proteins from each of the STK-associated ECF groups of the new ECF classification (Fig. 3.8, Fig. 5.1) and study divergent areas or residues that could be target of phosphorylation. These seven groups associated to STKs (ECF43, ECF59, ECF61, ECF62, ECF217, ECF267 and ECF283) (Fig. 5.1), are not associated to putative anti- $\sigma$  factors, except ECF267, which is encoded near a single-pass transmembrane FecR-like anti- $\sigma$  factor fused to TPRs (Table S3.1). Except for ECF43, there is no functional characterization of ECFs from any STK-associated groups. This makes

them a perfect target for the discovery of new phosphosites, but at the same time makes necessary to test the hypothesis derived from *in silico* approaches.

For the expansion of the groups associated to STK, there is no target protein, such EcfP from *V. parahaemolyticus*, that could guide the search. Hence, I focused on the proteins from all the genomes in NCBI that hit HMMs from STK-associated groups. Proteins were assigned to an STK-associated group when 1) they meet the criteria to be part of the STK-associated group, as derived from Section 3.5, and 2) they are encoded near ( $\leq 5$ Kbp) the coding sequence of a protein with a protein kinase domain (Pfam: Pkinase, PF00069). The criteria to define a protein as part of an STK-associated ECF group is slightly different from the strategy in Section 3.5 due to the computational burden imposed by running hmmscan function with the HMMs of every ECF group against every potential candidate member. With the aim of saving computational time and knowing that the probability score, based on the logistic regression of Eq. 3.1, alone has a good discriminative power to distinguish members from non-members of a group (data not shown), I only tested the probability scores achieved by the seven STK-associated groups for the ECF assignment. The probability of a protein belonging to a group, as determined by Equation 3.1, was calculated only for groups with HMMER bit scores  $\geq$  noise or trusted cut-offs. Proteins that do not feature a score  $\geq$  noise or trusted cut-offs against any group are left unclassified. Then, proteins are assigned to the group with the largest probability, as long as this value is above the threshold of 0.34%, as derived for full-length ECFs (Section 3.5). More details of this pipeline are available in Section 8.12. One limitation of this strategy is that I do not consider the possible classification of the protein against non-STK-associated ECF groups. Another difference respect to the original pipeline is that full-length sequences were used for the group assignment, instead of the stripped  $\sigma_2$ - $\sigma_4$  domains. The reason is that  $\sigma_2$  and  $\sigma_4$  domains could be divergent in STK-associated proteins, and hence, not hit by Pfam models for these regions.

As a result, I retrieved 4,719 ECF-STK pairs. Of these, 1,707 are unique ECF protein sequences. Most of the proteins are part of ECF43, but groups ECF262 and ECF283 also contain over 100 unique ECFs (Fig. 5.9). This alternative expansion strategy managed to retrieve more unique protein sequences from ECF43 than when looking for homologs of EcfP (995 pairs versus 931). This is likely due to the less stringent HMM built from the full ECF43, respect to performing the search with an HMM derived from a single protein sequence. Moreover, I was able to increase the size of all the STK-associated groups, but ECF267 and ECF283, which lost 6 and 30 non-redundant ECFs, respectively. This loss is due to the lack of a predicted protein kinase in the genetic neighborhood of some ECFs. The taxonomic origin of the retrieved proteins is diverse (Fig. 5.9). Most of the STK-associated groups are restricted to a single phylum (Fig. 5.9). Proteins from ECF61 and ECF217 are appear only in Planctomycetes, whereas proteins from ECF267 appear exclusively in Proteobacteria and ECFs from ECF283 are limited to Actinobacteria (Fig. 5.9). Exceptions are ECF43, ECF59 and ECF62 (Fig. 5.9). Interestingly, I found one protein from ECF43 present in the plant *Ricinus*

*communis* (Streptophyta). Even though it seems to contain only  $\sigma_4$  domain, this ECF could play a role in plastids.



**Figure 5.9. Taxonomic composition of the proteins from STK-associated groups after their expansion.** In order to lessen bias to more commonly sequence organisms, pie charts refer to ECFs from representative and reference organism, as defined by NCBI (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>), where only RefSeq entries are included when both RefSeq and GenBank assemblies are available for the same organism. The number of unique ECF sequences contained in each group is indicated as label in the pie chart.

As learned from ECF43, amino acid positions that are usually negatively charged, but substituted by conserved serine or threonine residues in STK-associated groups might be target of phosphorylation. This study could also be performed using one of the multiple phosphosite prediction tools available online (KinasePhos (Huang *et al.*, 2005), GPS (Xue *et al.*, 2008), and many others). However, these tools are typically optimized for eukaryotic models and for specific studied Hanks-type kinases (Blom, Gammeltoft and Brunak, 1999; Huang *et al.*, 2005; Xue *et al.*, 2008). As an example of the difference between bacterial and eukaryotic Hanks-type protein kinases, the latter usually also have tyrosine kinase activity, while this function is performed by a different protein family in bacteria, the bacterial tyrosine kinases (BY-kinases) (Pereira, Goss and Dworkin, 2011; Mijakovic, Grangeasse and Turgay, 2016). I created an MSA with the ECFs assigned to the seven extended STK-associated ECF groups, plus two random ECFs from each of the 150 ECF-groups that are not associated to STKs. The latter were used as background ECFs. Then, I focus on the top 15 columns of the alignment that contained the highest combined frequency of aspartate or glutamate (D+E) in background proteins. The frequency of D+E in these columns ranged from 98.29% to 36.3% in background ECFs. Then, I calculated the combined frequency of serine or threonine residues (S+T) in the same positions of the alignment. As a result, I was able to observe positions that typically contain negative charges in standard ECF  $\sigma$  factors, but are substituted by conserved serine or threonine in ECFs associated to STKs (Table 5.1 and 5.2). These positions could be target of phosphorylation. As a positive control, the position equivalent to the phosphorylated T63 of EcfP in *V. parahaemolyticus*

(P47 in RpoE from *E. coli*), contains a serine or a threonine residue (preferentially threonine) in 91.36% of the members of ECF43 (Table 5.1 and 5.2). This same residue could also be target of phosphorylation in ECF59 and ECF217, where 53.83% and 85.56% of the proteins contain serine or threonine, respectively (Table 5.1). However, in the case of ECF59 and ECF217 the residue that sits preferentially in the position of P47 is a serine (Table 5.2). These two groups contain another putative phosphorylation site in the linker (RpoE residue E119), which contains serine or threonine more often than position P47, with 65.38% and 91.11% of the members of ECF59 and ECF217 containing serine or threonine, respectively (Table 5.1). Interestingly, both residues in region 2.2 and linker harbor preferentially serine (Table 5.2), which indicates that both could be phosphorylated as part of an *in vivo* regulatory response. Another residue in the linker (N122 in RpoE) could be phosphorylated in ECF283, since 83.61% of the members of this group contain serine or threonine in this position (preferentially threonine) (Table 5.1 and 5.2). Residues in the area that links  $\sigma_2$  and  $\sigma_4$  domain are in intimate contact with the active site of the RNAP complex (Fang *et al.*, 2019). This area, which fulfills the same function as  $\sigma_{3.2}$  in non-type IV  $\sigma$  factors, occupies the RNA exit channel, where it mimics RNA and organizes ssDNA during transcription initiation (Fang *et al.*, 2019).  $\sigma_{3.2}$  needs to be expelled from the core RNAP during promoter escape (>5bp transcript) in order to allow transcription elongation (Fang *et al.*, 2019) (Section 1.2.1). It is plausible that phosphorylation of the linker modifies ECF activity, either changing RNAP or promoter affinity. Interestingly, 3/7 groups have a potential phosphorylation site in the linker area.

**Table 5.1. Combined frequency of serine and threonine residues (S+T) in the top 15 columns of the ECF multiple-sequence alignment with the highest combined frequency of aspartate and glutamate (D+E) in background ECFs.** Background ECFs are two randomly selected ECFs from each of the 150 ECF groups that are not associated to STKs. The ECF region where a certain residue is located and the corresponding amino acid in RpoE from *E. coli* are indicated in columns one and two, respectively. Columns three and four refer to the D+E and T+S frequencies in background ECFs. The rest of the columns refer to S+T frequency in STK-associated ECF groups. The number of non-redundant ECF protein sequences from which the frequencies are drawn is indicated in the last row. Bold numbers draw attention to positions that could be phosphorylated in each ECF group.

Region	Position RpoE ( <i>E. coli</i> )	Background (D+E)	Background (S+T)	ECF43 (S+T)	ECF59 (S+T)	ECF61 (S+T)	ECF62 (S+T)	ECF217 (S+T)	ECF267 (S+T)	ECF283 (S+T)
N-terminus	D7	36.3	2.4	0	43.59	0	24.71	15.56	4.55	28.28
2.1	D18	47.95	6.51	2.81	5.13	0	8.49	5.56	0	0
2.2	D45	56.16	10.27	24.72	10.26	42.11	2.32	17.78	13.64	0
	P47	74.32	1.37	<b>91.36</b>	<b>53.85</b>	0	2.32	<b>85.56</b>	4.55	0
	D48	98.29	0.34	9.85	0	0	0	0	0	0
	E52	88.7	2.4	0.2	0	0	3.09	1.11	0	0
2.4	N84	48.63	12.33	1.01	3.85	0	10.04	13.33	36.36	40.98
Linker	D99	47.95	4.45	0	11.54	10.53	13.13	3.33	0	0
	E119	59.25	4.79	9.35	<b>65.38</b>	0	2.7	<b>91.11</b>	0	0
	N120	39.04	8.9	10.85	10.26	15.79	6.56	11.11	9.09	<b>83.61</b>
4.1	E126	42.81	6.16	1.41	0	36.84	1.16	0	4.55	0
	E140	45.55	2.74	2.71	3.85	0	6.18	8.89	4.55	0.41
4.2-4.3	D152	40.75	2.4	4.62	0	5.26	1.16	0	0	8.61
	G153	36.3	1.03	0.3	0	0	2.32	0	0	5.33
4.3	E158	79.45	2.4	0.5	1.28	0	11.58	3.33	18.18	0
4.3	A161	42.12	7.88	6.13	6.41	5.26	7.34	10	4.55	11.89
<b>Number of unique ECFs</b>			292	995	78	19	259	90	22	244

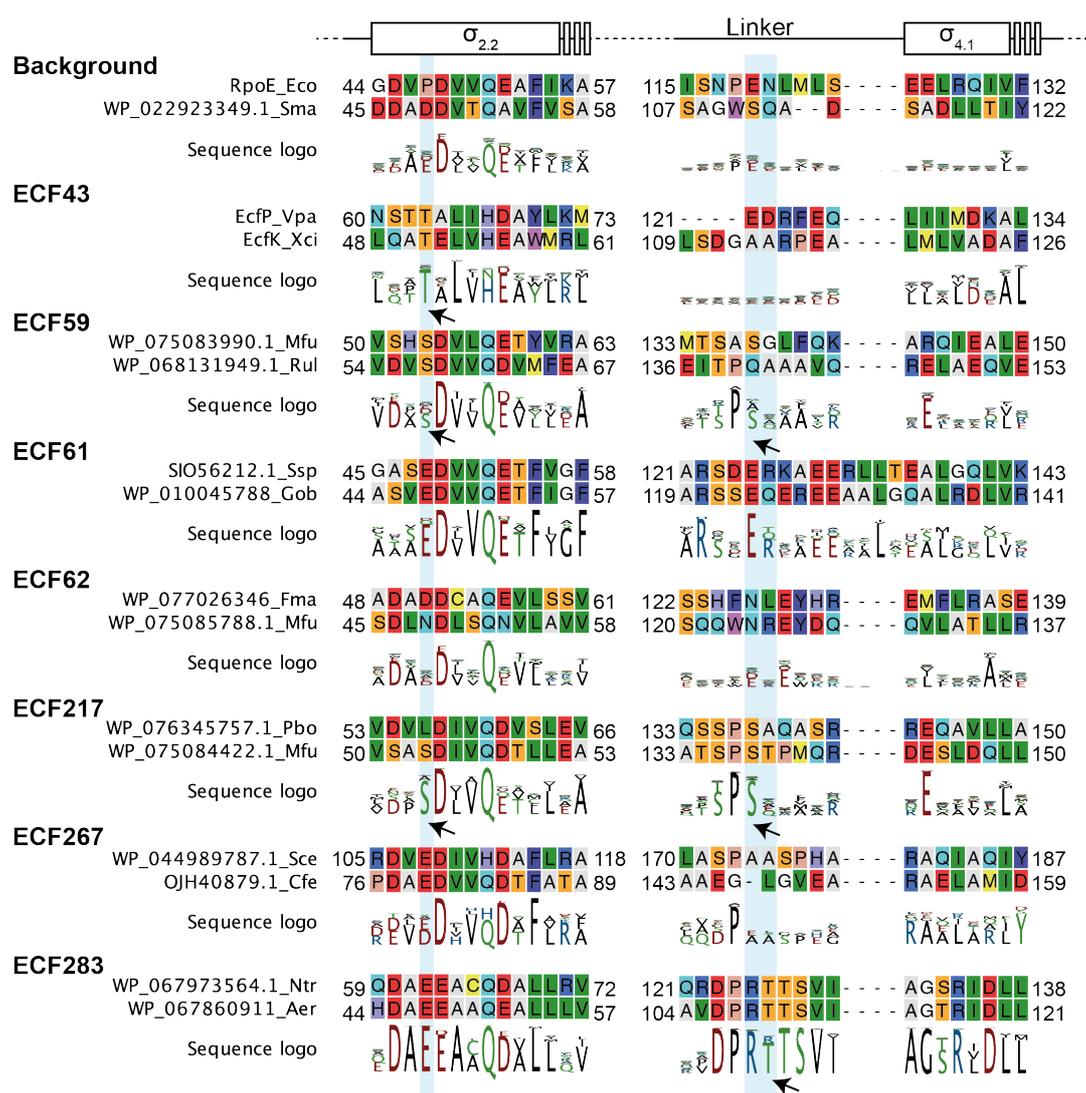
**Table 5.2. Most common amino acid in the top 15 columns** of the ECF multiple-sequence alignment with the highest combined frequency of aspartate and glutamate in background ECFs. Background ECFs are two randomly selected ECFs from each of the 150 ECF groups that are not associated to STKs. The ECF region of the column and the corresponding amino acid in RpoE from *E. coli* are indicated in columns one and two, respectively. Green cells indicate negatively-charged residues (aspartate and glutamate), whereas orange cells indicated phosphorylatable residues by STKs (threonine and serine). Bold residues draw attention to positions that could be phosphorylated in each ECF group.

Region	Position RpoE ( <i>E. coli</i> )	Background	ECF43	ECF59	ECF61	ECF62	ECF217	ECF267	ECF283
N-terminus	D7	D	M	<b>T</b>	D	R	I	D	L
2.1	D18	D	D	D	D	D	D	D	D
2.2	D45	D	Q	D	A	D	D	E	D
	P47	E	<b>T</b>	<b>S</b>	E	D	<b>S</b>	D	E
	D48	D	A	D	D	D	D	D	E
	E52	E	E	E	E	E	E	D	D
2.4	N84	D	D	D	D	D	<b>S</b>	Q	
Linker	D99	D	A	A	D	D	D	A	A
	E119	E	D	<b>S</b>	E	E	<b>S</b>	A	R
	N120	E	D	Q	R	R	Q	A	<b>T</b>
4.1	E126	E	L	E	A	L	E	A	G
	E140	E	P	E	F	P	E	A	P
4.2-4.3	D152	E	G	E	K	E	E	E	C
	G153	G	G	E	G	G	G	G	Q
4.3	E158	E	E	E	D	E	E	E	E
4.3	A161	E	E	E	E	E	E	A	E

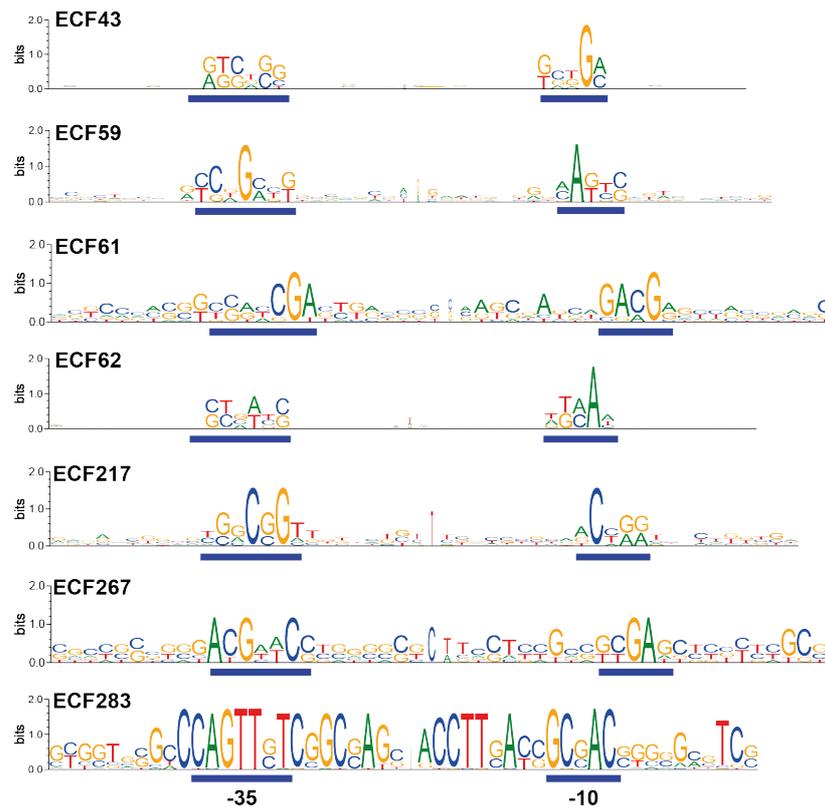
I predicted bipartite sequences upstream of the coding sequences of members of each STK-associated group using BioProspector (Liu, Brutlag and Liu, 2001) (see [Section 8.6](#) for details). The resulting predictions for promoter target motifs did not show any clear ECF target element (Fig. 5.11). This means that, either the database of proteins is too small or ECFs from these groups are not self-regulated. Some predictions, for instance ECF217 and ECF283, could be real promoters. Future experiments will attempt to use these promoter motifs as reporters of ECF  $\sigma$  factor activity in members of their group.

In summary, I found putative phosphosites in four out of seven STK-associated ECF groups, including the experimentally confirmed phosphosite of ECF43. The predicted phosphosite is either in  $\sigma_{2.2}$  helix (P47 in RpoE from *E. coli*) or in the linker (E119 or N120 in RpoE from *E. coli*), or in both regions. In the case of members of ECF43, the phosphorylated threonine in helix  $\sigma_{2.2}$  provides a negative charge that enhances binding to RNAP (Section 5.2). The other two groups with a putative phosphosite in P47 are ECF59 and ECF217. However, these groups feature a serine residue in the same position and contain a putative phosphosite in the linker. It is unclear whether the usage of serine instead of threonine would change the functional role of the phosphorylation in this residue. Future experiments will target these groups to find out whether the serine located in P47 is indeed phosphorylated, whether this phosphorylation changes the binding affinity of members of ECF59 and ECF217 to the RNAP and what is the role played by the putative phosphorylation site in the linker. Moreover, ECF283 also contains a putative phosphosite in the linker in positions equivalent to E119 or N120 in RpoE from *E. coli*. The linker region is not as conserved as  $\sigma_2$  and  $\sigma_4$  domains, and the percentage of background ECFs with aspartate or glutamate in E119 and, especially N120, is smaller

than in  $\sigma_{2.2}$ 's residue (Table 5.1). Nevertheless, the linker region is in contact with the active cleft of the RNAP complex (Fang *et al.*, 2019), and its phosphorylation could change the affinity to the RNAP, promoter escape or promoter specificity. Members of ECF283 are perfect candidates for the study of this potentially new mechanism of ECF  $\sigma$  factor regulation since 1) they lack of a putative phosphosite in  $\sigma_{2.2}$  helix, 2) ECF283 is the second most abundant group with a predicted phosphosite, and 3) their predicted putative target promoter is feasible and, if confirmed, could easily be used in reporter experiments, where ECF283 promoter would control the expression of a reporter gene that would indicate the activity of ECF283 under different conditions.



**Figure 5.10. Multiple-sequence alignment of STK-associated ECF groups and background proteins.** Background ECFs are composed by two randomly-selected ECFs from each of the 150 ECF groups that are not associated to ECF  $\sigma$  factors. Two proteins from each set were selected for representation. The logo refers to all the members of a given set. Names of sequences include their common name, or NCBI protein identifier, followed by the organism where the protein is present. Positions where phosphorylated amino acids were predicted are shaded, and arrows point to the specific predicted phosphosites. Organism codes go as follows: Aer=*Aeromicrobium erythreum*, Cfe=*Cystobacter ferrugineus*, Eco=*Escherichia coli*, Fma=*Fuerstia marisgermanicae*, Gob=*Gemmata obscuriglobus*, Mfu=*Mariniblastus fucicola*, Ntr=*Nocardopsis trehalosi*, Pbo=*Paludisphaera borealis*, Rul=*Roseimaritima ulvae*, Sce=*Sorangium cellulosum*, Sma=*Serinicoccus marinus*, Ssp=*Singulisphaera* sp. GP187, Vpa=*Vibrio parahaemolyticus*, Xci=*Xanthomonas citri*.



**Figure 5.11. Predicted target promoter motifs for STK-associated ECF groups.** These motifs are the result of running BioProspector (Liu, Brutlag and Liu, 2001) on the 200bp upstream of the ECF operon, where operon is defined as the set of coding sequences transcribed in the same direction and with a distance shorter than 50bp. In order to lessen taxonomic bias, I only include ECFs from representative and reference organism, as defined by NCBI (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>), where only RefSeq entries are included when both RefSeq and GenBank assemblies are available for the same organism. BioProspector predictions for -35 and -10 elements are indicated. Logos are the result of pilling DNA sequences, centered at the two regions predicted by BioProspector.

### 5.5. PknT closest relatives are part of T6SS clusters in *Vibrio* spp.

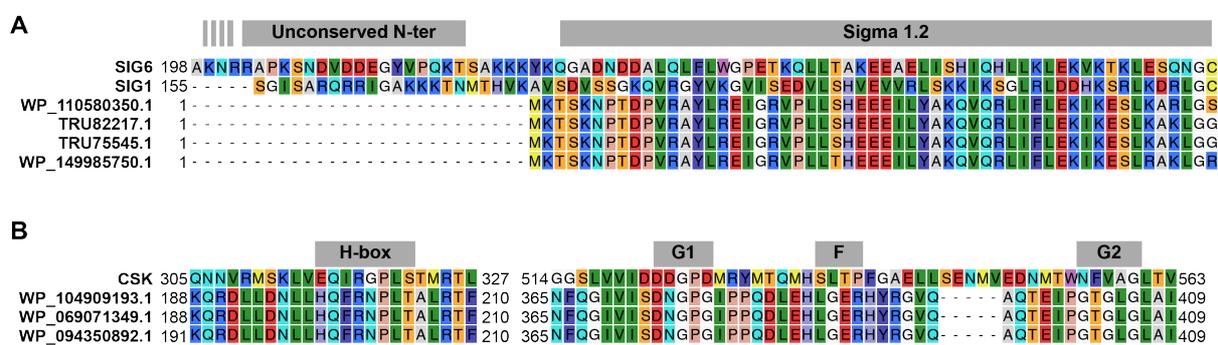
Even though this is the first systematic characterization of ECF  $\sigma$  factor phosphorylation,  $\sigma$  phosphorylation by STKs has been described in plant and green algae chloroplasts. However, this system has fundamental differences from ECF43 phosphorylation. In this paragraph I will compare the most studied  $\sigma$  factor in plants, SIG1 from *Arabidopsis thaliana*, with EcfP regulation in *V. parahaemolyticus*. First, the six  $\sigma$  factors in *A. thaliana* are derivatives of group 1  $\sigma$  factors with an extended N-terminal unconserved region that functions as region 1.1, or gatekeeper, when phosphorylated (Puthiyaveetil, Ibrahim and Allen, 2013). In contrast, region 1.1 is absent in ECF  $\sigma$  factors. Second, non-phosphorylated SIG1 is active in plastids (Puthiyaveetil, Ibrahim and Allen, 2013). However, EcfP seems to be only active upon phosphorylation (Section 5.2). Phosphorylation of SIG1 at T170 (Shimizu *et al.*, 2010) changes its promiscuity towards target promoters (Puthiyaveetil, Ibrahim and Allen, 2013). Non-phosphorylated region 1.1 allows for the promiscuous recognition of promoters divergent from consensus (Puthiyaveetil, Ibrahim and Allen, 2013). Upon phosphorylation of region 1.1, SIG1 becomes more stringent towards consensus promoters, hereby displacing the transcriptional machinery of the chloroplasts to respond to changes in the quality of light (Puthiyaveetil, Ibrahim and Allen, 2013). Third, SIG1 regulation by phosphorylation evolved by

evolutionary tinkering of CSK, a histidine kinase from a 2CS in Cyanobacteria and non-green algae that lost its response regulator and changed specificity towards threonine residues (Puthiyaveetil, Ibrahim and Allen, 2013). The evolution of the EcfP-PknT signal transduction module in *V. parahaemolyticus* is unknown, but it seems unlikely that it evolved from 2CSs given that, in contrast to CSK, PknT kinase does not harbor any histidine kinases domain. However, regulation of  $\sigma$  factors from plastids is more complex and involves a Hanks-type kinase associated to the bacterial multisubunit RNAP named plastid transcription kinase (PTK) (Schweer, Türkeri, Kolpack, *et al.*, 2010). PTK is closely related to the  $\alpha$ -subunit of the nucleocytoplasmic casein kinase 2 (CK2), hence called in some instances cpCK2 (Schweer, Türkeri, Link, *et al.*, 2010). PTK/cpCK2 targets multiple phosphosites of different plastid  $\sigma$  factors in *A. thaliana*, but it has been more studied in the context of SIG6 (Schweer, Türkeri, Link, *et al.*, 2010). In SIG6, PTK/cpCK2 phosphorylation is dependent on another unknown “pathfinder” STK, at least for the phosphorylation of the residue that shows the strongest mutant phenotype, Ser174, laying in the unconserved N-terminus (Schweer, Türkeri, Link, *et al.*, 2010). Interestingly, PKT/cpCK2 is also target of CSK (Puthiyaveetil, Ibrahim and Allen, 2013), suggesting that CSK acts upstream of PTK/cpCK2, which in turn phosphorylates several  $\sigma$  factors and subunits of the bacteria-like RNAP (Puthiyaveetil, Ibrahim and Allen, 2013). The differences between  $\sigma$  factor regulation by phosphorylation in chloroplasts and in members of ECF43 suggest that both systems arose independently.

If phosphorylation of ECFs and plant  $\sigma$  factors would have emerged from a common origin and then diversify to modify  $\sigma$  factor activity in different manners, then the same type of phosphorylation that occurs in plastids should happen in bacterial  $\sigma$  factors. In order to test a potential divergent evolution mechanism from a common ancestor, I tested the presence of proteins similar to SIG1 (NCBI: NP\_176666.1) in Cyanobacteria, the closest bacterial relatives to plant plastids (Mereschowsky, 1905), using BLAST (online version, default options). This yielded proteins with a maximal identity of 38.28% that did not include the divergent unconserved N-terminus of SIG1, nor any residue equivalent to SIG1’s phosphorylatable T170. However, the N-terminal part of these proteins is rich in serine and threonine (Fig. 5.12A). To further test divergent evolution, I searched for proteins similar at a sequence level to CSK kinase (NCBI: NP\_564908.1) in Cyanobacteria using BLAST (online version and default parameters). This yielded histidine kinases with a maximal identity of 24.36% and canonical histidine kinase H, G1 and G2 boxes (Fig. 5.12B). The H-box harbors glutamate in CSK instead of the autophosphorylated histidine in cyanobacterial proteins (Fig. 5.12B). This mutation is associated to the change of target of CSK, which phosphorylates threonine instead of the aspartate of response regulators (Puthiyaveetil *et al.*, 2008). Additionally, regions G1 and G2, important for ATP stabilization and gamma-phosphate hydrolysis, differ between CSK and cyanobacterial CSK-like proteins (Fig. 5.12B), suggesting a different ATP hydrolysis mechanism in cyanobacterial and CSK-like plant homologs (Ibrahim *et al.*, 2016). Moreover, cyanobacterial CSK-like proteins lack the N-terminus of CSK (~110 aa). Plant  $\sigma$  factors are also regulated by Hanks-type kinases such as

PKT/cpCK2. Looking for proteins similar to PKT/cpCK2 (NCBI: O64816.1) in Cyanobacteria yielded proteins with a maximal identity of 28.16% (alignment not shown). However, it is not possible to assess whether these proteins target  $\sigma$  factors in Cyanobacteria.

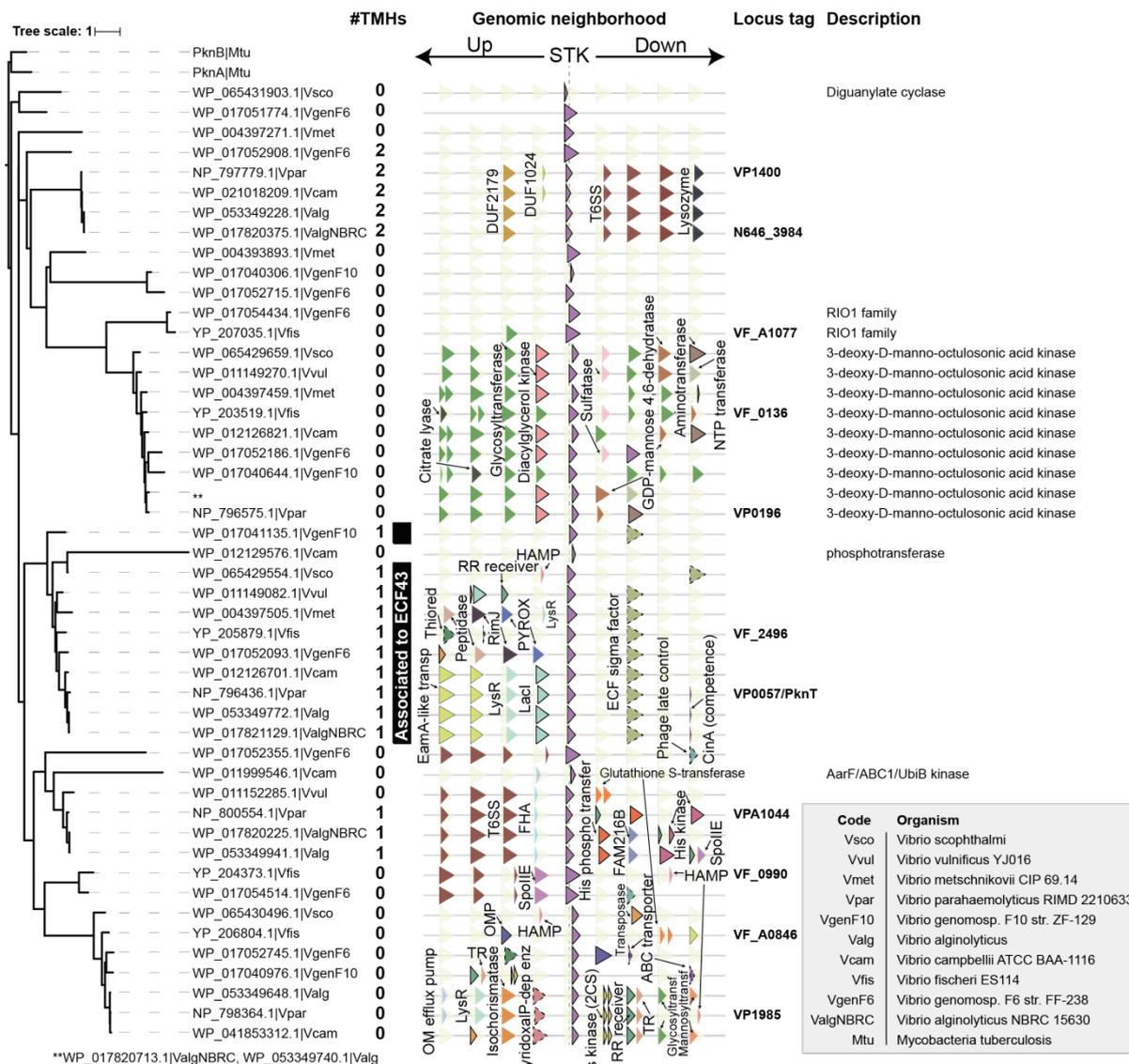
In conclusion, even though  $\sigma$  factors from Cyanobacteria could be phosphorylated, this phosphorylation is not likely mediated by CSK-like proteins and it does not seem likely to occur in region 1.1. This suggests that phosphorylation does not regulate promoter stringency in Cyanobacteria in a similar manner as in SIG1. Furthermore, given the lack of CSK and SIG1 homologs in Cyanobacteria, it does not seem likely that ECF  $\sigma$  factor phosphorylation and plastid  $\sigma$  factor phosphorylation share a common origin. However, this possibility cannot be completely ruled out since I only searched for CSK and SIG1 homologs in cyanobacterial genomes, and there might be other bacteria that possess a similar system, which could have been lost in cyanobacteria after the endosymbiotic event. An even more unlikely scenario is that CSK and/or SIG1 homologs were only present in the specific Cyanobacteria that took part in the endosymbiotic event that generated plants and green algae, and that this bacterial lineage was later lost. This is even more unlikely given that in red algae and diatoms CSK still seems to be coupled to a response regulator, indicating that the repurposing of CSK as STK happened mostly in the green lineage (Puthiyaveetil and Allen, 2009) and supporting the idea that ECF phosphorylation in bacteria and in plants have a different origin.



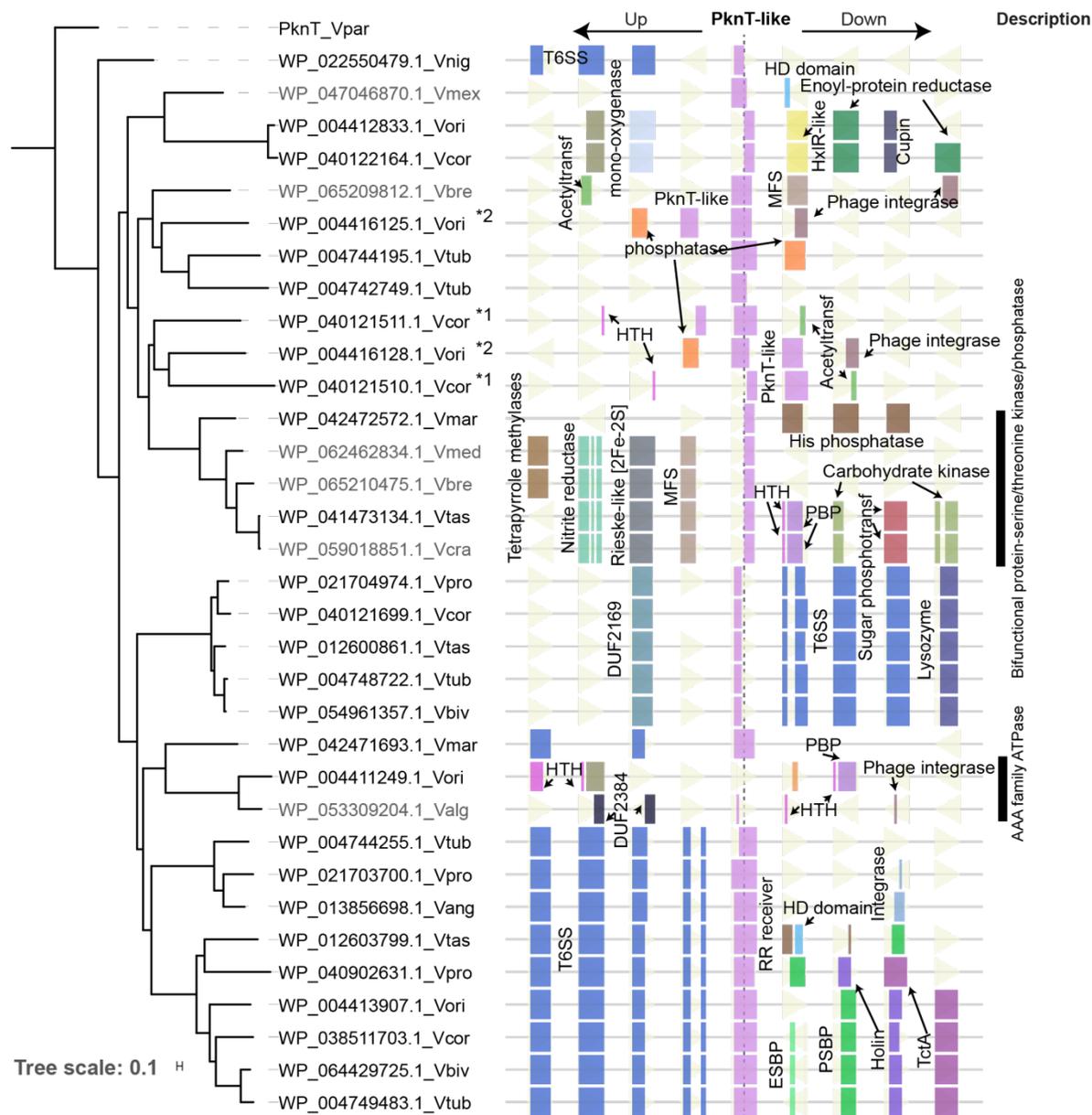
**Figure 5.12. Multiple-sequence alignment of the most similar cyanobacterial proteins to plastid  $\sigma$  factors and CSK kinase. A:** *Arabidopsis thaliana* SIG1 (NCBI: NP\_176666.1) and SIG6 (NCBI: Q9LD95.1) with the top four BLAST results when searching for SIG1 in Cyanobacteria using online BLAST with default options. The four proteins shown in the alignment are from *Microcystis* genus. The unconserved region with the position where SIG1 is phosphorylated (T170) is shown (Shimizu *et al.*, 2010). S174, the most important residue in SIG6's phosphorylation (Schwee, Türkeri, Link, *et al.*, 2010) is not shown. Cyanobacterial proteins do not contain this region. **B:** Top three most similar cyanobacterial proteins to *A. thaliana* CSK (NCBI: NP\_564908.1) from *Nostoc* genus. CSK contains divergent H-box, G1 and G2 regions that are not found in any cyanobacterial proteins.

Since plants do not seem to be the origin of members of ECF43, I focused on members of *Vibrio* spp. to find the evolutionary origin of these proteins. There are 10 reference/representative genomes from *Vibrio* spp. that encode members of ECF43. Focusing on these genomes, I found 49 proteins with sequence similarity to STKs. A phylogenetic tree of the protein kinase domain of these proteins revealed that STKs associated to members of ECF43 generally cluster together (Fig. 5.13, black edge). This search also identified 3-deoxy-D-manno-octulosonic acid (Kdo) kinases, which participate in the biosynthesis of lipid A from the LPS (Harper *et al.*, 2010). Kdo kinases clustered together in the

same clade (Fig. 5.13). Aside from EcfP, *V. parahaemolyticus* contains other three STKs, of which two are genomically linked to T6SSs (Fig. 5.13). There are two T6SSs in *V. parahaemolyticus*, T6SS1 and T6SS2 (Broberg, Calder and Orth, 2011). T6SS1 is most active under high salt conditions, such as in sea water (Salomon *et al.*, 2013; Li *et al.*, 2017). In contrast, T6SS2 is found in all tested strains of *V. parahaemolyticus* and is only active under low salt conditions (Yu *et al.*, 2012; Salomon *et al.*, 2013). The STK VP1400 is part of one of the seven putative operons that encode T6SS1, while another STK, VPA1044, is part of one of the three operons that compose T6SS2 (Makino *et al.*, 2003). The last STK from *V. parahaemolyticus*, VP1985, is encoded near a histidine kinase and a response regulator, which are potentially part of a 2CS (Fig. 5.13). The closest relatives to PknT are VPA1044-related (Fig. 5.13). The presence of a transmembrane helix in VPA1044-like kinases indicates that they could be in charge of the transmembrane transduction of a putative sensing system, likewise STKs associated to members of ECF43. However, instead of an ECF  $\sigma$  factor, the targets of regulation of VPA1044-like kinases are more likely to include the forkhead-associated (FHA) domain-containing protein encoded upstream (Fig. 5.13), likewise the STK PpkA from *P. aeruginosa* (Mougous *et al.*, 2007). PpkA is required for the assembly of the T6SS encoded in the Hcp secretion island I (Mougous *et al.*, 2007). The relationship between VPA1044, the histidine phosphotransferase and the histidine kinase contained in its genetic neighborhood is unknown. One possibility is that the histidine phosphotransferase or the histidine kinase are targets of VPA1044-like phosphorylation (Fig. 5.13). In principle they could be part of the same sensing pathway that would regulate the expression of T6SS2. However, this has never been tested to the best of my knowledge. Further linking ECF43-associated STKs and T6SSs, the STK PknS from *X. citri* is needed for the activity of EcfK, a member of ECF43 (Bayer-Santos *et al.*, 2018). EcfK in turn regulates the expression of a T6SS (Bayer-Santos *et al.*, 2018), suggesting a functional connection between STKs that take part on T6SS regulation and STKs associated to members of ECF43.



**Figure 5.13. Phylogenetic tree of protein kinase domains of proteins from reference and representative organisms that contain members of ECF43 in *Vibrio* spp.** PknA and PknB from *M. tuberculosis* are included as outliers, used to root the tree. Sequences are labeled with their NCBI identifier plus a code for the organism of origin (see legend). STKs associated to members of ECF43 are labeled. The genetic neighborhood of kinases is shown. For clarity, only proteins that appear in several genetic neighborhoods are depicted. Symbols do not indicate transcription direction. Locus tags and other descriptors linked to the kinase are shown in the last two columns.



**Figure 5.14. Phylogenetic tree of PknT-like proteins in organisms that do not contain any member of ECF43 in their genome.** Only reference and representative organisms, as defined by NCBI, are considered. Only RefSeq genomes are considered if both RefSeq and GenBank records are available for the same organism. PknT-like kinases from organisms where none of these proteins is associated to

T6SSs are indicated with gray labels. Proteins encoded  $\pm 4$  positions from the PknT-like coding sequence are indicated with beige triangles. The direction of the triangle indicates direction of transcription. Positions with no protein annotation (no beige triangle) correspond to pseudogenes. Domains of the genetic neighborhood that are present in several proteins are showed and labeled. Two genetic neighborhoods, labeled with an asterisk, are repeated since they contain two PknT-like proteins. Description of the PknT-like proteins, extracted from NCBI, is shown on the right when it differs from protein kinase or serine/threonine protein kinase.

I further looked for PknT-like proteins in the 20 representative or reference *Vibrio* spp. genomes that do not contain any member of ECF43. I used BLAST with PknT as query and E-value<0.01 (Section 8.13). I did not include in this search GenBank genomes when an equivalent RefSeq genome exists. The aim was to find close relatives of PknT in genomes where no ECF43-associated PknT ortholog is present. Out of the 20 genomes, 14 have a protein with sequence similarity to PknT. Looking at the genomic neighborhood of these PknT-like kinases, ~48% (16 out of 33) are encoded near elements from T6SSs. These T6SS-associated, PknT-like kinases appear in nine out of the 14 *Vibrio* spp. In the remaining five *Vibrio* genomes, three PknT-like kinases appear to be “bifunctional protein-serine/threonine kinases/phosphatases” (Fig. 5.14). Therefore, the association of PknT-like STKs to T6SSs is not only restricted to *Vibrio* spp. that harbor members of ECF43.

From the 30 *Vibrio* spp. analyzed during this section, 14 contain a PknT-like protein associated to a T6SS, but no member of ECF43 in their genomes. In contrast, 7 *Vibrio* spp. contain both T6SS-associated STK and ECF43, 6 *Vibrio* spp. contain neither and 3 contain a member of ECF43, but lack any detected T6SS-associated STK. The fact that the lack of ECF43, but the presence of T6SS-associated STK is the most common situation in *Vibrio* spp. genomes suggests that T6SS appeared before in *Vibrio* spp. than ECF43-bases systems.

## 5.6. Discussion and summary

Even though anti- $\sigma$  factors are present in 114 out of 157 ECF groups (Section 3.4), seven ECF groups lack these proteins and contain instead microsyntenic STKs. Previous reports have suggested a potential regulatory role of these STKs in the function of their associated ECF  $\sigma$  factors (Mascher, 2013). The regulatory role of a STK, PknS, over a member of ECF43, EcfK from *X. citri*, has been proven *in vivo* (Bayer-Santos *et al.*, 2018), but the direct phosphorylation of EcfK by PknS and the biological effect of this phosphorylation have never been shown. In this section I studied ECF  $\sigma$  factor phosphorylation computationally. First, I focused on members of ECF43, which contain a conserved threonine residue in  $\sigma_{2.2}$ , equivalent to T63 in EcfP from *V. parahaemolyticus*. Phosphomimetic mutants of the residue equivalent to T63 in EcfK are constitutively active, supporting their phosphorylation as a mean of activating members of ECF43 (Bayer-Santos *et al.*, 2018). Dr. Shankar Chandrashekar Iyer and Dr. Simon Ringgaard proved that T63 is phosphorylated *in vivo* in two members of ECF43, EcfP from *V. parahaemolyticus* and ECF43\_Hne from *H. neptunium*, and that this phosphorylation depends on the microsyntenic STK in the case of EcfP (Section 5.2). T63 is located in region  $\sigma_{2.2}$ , in an area that normally contacts the clamp helices of the  $\beta'$  subunit of the RNAP through charged interactions (L. Li *et al.*, 2019). We speculated that phosphorylation would

increase the affinity of members of ECF43 to the RNAP. In support of this idea, results of Dr. Shankar Chandrashekar Iyer and Dr. Simon Ringgaard showed that the co-immunoprecipitated  $\beta/\beta'$  fraction significantly increased when PknT kinase was overexpressed, respect to the kinase deletion mutant (Fig. 5.6). This is the first description of ECF  $\sigma$  factor regulation by phosphorylation, which seems to be a widespread mechanism of bacterial signal transduction, given the broad taxonomic distribution of members of ECF43 (Fig. 5.7).

The response that members of ECF43 mediate seems to be diverse. On one hand, EcfP is essential for polymyxin resistance in *V. parahaemolyticus* (Chandrashekar Iyer *et al.*, accepted, Fig. 5.5); however, EcfK from *X. citri* is required for resistance to Dictyostelium predation (Bayer-Santos *et al.*, 2018). The mechanism by which any of these stresses is signaled to their respective transmembrane STKs remains to be elucidated. Moreover, the analysis of the extracytoplasmic domains of the protein kinases associated to members of ECF43 revealed a broad array of sensing modules (Fig. 5.8), indicating either the direct sensing of different ligands or an indirect sensing mechanism involving a distinct set of intermediate proteins. This situation is reminiscent of other transmembrane components of signal transduction mechanisms, such as the methyl-accepting chemotaxis proteins (MCPs) of chemotactic systems. MCPs generally contain a ligand-binding extracytoplasmic domain, and a cytoplasmic signaling domain that transmits the binding of the ligand to downstream proteins, including the histidine kinase CheA (reviewed in (Salah Ud-Din and Roujeinikova, 2017)).

Bioinformatic analyses of the rest of the STK-associated ECF groups showed that a residue equivalent to T63 in *V. parahaemolyticus* could be phosphorylated in ECF57 and ECF217. Future analyses would test whether the putative phosphorylation of  $\sigma_{2.2}$  in ECF59 and ECF217 has the same function as in ECF43. Furthermore, residues in the non-conserved linker region could be the target of phosphorylation in the STK-associated groups ECF59, ECF217 and ECF283. Interestingly, SigH from *M. tuberculosis* – from group ECF12, which lacks microsynteny with STKs – is phosphorylated on a residue within its linker (T106) by the kinase PknB (Sang, Kang and Husson, 2008). This kinase does not share genetic neighborhood with SigH and the role of the phosphorylation of T106 is unknown, since it does not have any effect on anti- $\sigma$  factor binding (Sang, Kang and Husson, 2008). The putative phosphorylation sites predicted in the linker of groups ECF59, ECF217 and ECF283 need to be tested experimentally. One possibility is that changes in the polarity of the linker change the activity of ECF  $\sigma$  factors. The linker between  $\sigma_2$  and  $\sigma_4$  domains behaves as region 3.2 in non-ECF  $\sigma^{70}$ s (Fang *et al.*, 2019) (Section 1.2.1). This region is in intimate contact with the active cleft of the RNAP core complex (Fang *et al.*, 2019). An increased affinity of ECF  $\sigma$  factors for this region could increase ECF binding and transcription initiation rate, but an excessive binding affinity could also have detrimental effects on promoter escape or affect abortive transcription (Yan Ning Zhou, Walter and Gross, 1992; Pupov *et al.*, 2014). An alternative possibility is that phosphorylation of the linker lessens ECF affinity for RNAP. This would render the ECF inactive when phosphorylated. This hypothesis is less likely than the ECF activation by phosphorylation since groups ECF59 and ECF217

also contain a putative phosphorylated serine in their  $\sigma_{2.2}$  region, whose phosphorylation may be required for the activation of these ECFs, as in the case of members of ECF43. Therefore, it is unlikely that a putative phosphorylation of the linker has the opposite effect.

In summary, ECF phosphorylation in bacteria seems to have emerged from the combination of two signal transduction mechanisms, a STK and an ECF  $\sigma$  factor. Together they are able to regulate RNAP holoenzyme formation and direct its activity towards the transcription of alternative coding sequences.

## 6. ECF regulation based on C-terminal extensions

ECF  $\sigma$  factors are usually regulated by anti- $\sigma$  factors (114 out of the 157 ECF groups) (Section 3.4). However, the second most likely regulatory mode of ECFs is the presence of C-terminal extensions of their sequence (19 ECF groups) (Section 3.4). Members of these groups typically lack any putative anti- $\sigma$  factor encoded in their genetic neighborhood (Fig. 3.8). One exception is ECF263, which contains a C-terminal extension with five or seven transmembrane helices, depending on the subgroup, and a putative anti- $\sigma$  factor, usually encoded directly upstream of the ECF coding sequence and often featuring a DUF2007 domain and one transmembrane helix (Table S3.1). Illustrating the importance of C-terminal extension-containing groups, the second and third most abundant ECF  $\sigma$  factor groups, ECF41 and ECF42, contain C-terminal extensions (Fig. 3.8, Table S3.1). Proteins in group ECF41 contain a conserved SnoaL-like domain (Pfam: PF12680) related to epoxide hydrolases in its ~200aa extension (Wecke *et al.*, 2012). Instead, proteins from group ECF42 contain a conserved TPR in its ~120aa extension (D'Andrea and Regan, 2003; Staroń *et al.*, 2009). Given the large diversity of the domains contained in C-terminal extensions (Table S3.1), it is likely that their functional role is diverse. Indeed, while the SnoaL-like extension of ECF41 seems to have a dual role as activator and inhibitor of ECF activity (Wecke *et al.*, 2012), the TPR extension in ECF42 appears to be essential for ECF activity (Liu, Pinto and Mascher, 2018).

ECF  $\sigma$  factors from four neighboring groups – ECF41, ECF56, ECF294 and ECF294 – contain a SnoaL-like C-terminal extension (Table S3.1). Since groups ECF41 and ECF56 are clearly defined in both the original and the new ECF classification and ECF core regions were the only input of both clustering algorithms, it is possible that differences in the interaction between C-terminal extension and ECF core regions are responsible of the assignment of these proteins to different groups. These differences could be translated into a different regulatory role of the C-terminal extension over ECF activity. However, the nature of these sequence differences and their translation into a biological role has never been explored.

In this section I focus on C-terminal extensions as a source of regulation of ECF  $\sigma$  factor activity. First, I focused on deciphering the differences in the role of C-terminal extensions in ECF groups ECF41 and ECF42, which harbor different domains. Then, I targeted the largest two groups with SnoaL-like C-terminal extensions, ECF41 and ECF56, in an attempt to find differences in the interaction mode between their SnoaL-like C-terminal extensions and the core ECF regions. Results of this section critically depended on DCA predictions and are only possible for the largest ECF groups associated to C-terminal extensions. This work revealed that even the same type of C-terminal extension is likely to fulfill a different regulatory role on ECF activity, although experimental confirmation is only provided in the case of groups ECF41 and ECF42, which harbor C-terminal extensions with different domains.

**6.1. C-terminal extensions have a different role in ECF41 and ECF42**

The first part of the analysis has been published as a research article in *Molecular Microbiology* in 2019, licensed to be used in this thesis under the number 4727020486120. I contributed to this work by planning the computational experiments, writing parts of the DCA pipeline, analyzing and integrating DCA and wet-lab results and writing the text (Wu *et al.*, 2019).

# The role of C-terminal extensions in controlling ECF $\sigma$ factor activity in the widely conserved groups ECF41 and ECF42

Hao Wu,<sup>1†</sup> Qiang Liu,<sup>2,3†</sup> Delia Casas-Pastor,<sup>1†</sup> Franziska Dürr,<sup>2†</sup> Thorsten Mascher<sup>1,2</sup> and Georg Fritz<sup>1\*</sup>

<sup>1</sup>LOEWE-Center for Synthetic Microbiology (SYNMIKRO), Philipps-Universität Marburg, 35032, Marburg, Germany.

<sup>2</sup>Institute of Microbiology, Technische Universität (TU) Dresden, 01062, Dresden, Germany.

<sup>3</sup>Department Biology I, Ludwig-Maximilians-Universität München, Planegg-Martinsried, Germany.

## Summary

The activity of extracytoplasmic function  $\sigma$ -factors (ECFs) is typically regulated by anti- $\sigma$  factors. In a number of highly abundant ECF groups, including ECF41 and ECF42,  $\sigma$ -factors contain fused C-terminal protein domains, which provide the necessary regulatory function instead. Here, we identified the contact interface between the C-terminal extension and the core  $\sigma$ -factor regions required for controlling ECF activity. We applied direct coupling analysis (DCA) to infer evolutionary covariation between contacting amino acid residues for groups ECF41 and ECF42. Mapping the predicted interactions to a recently solved ECF41 structure demonstrated that DCA faithfully identified an important contact interface between the SnoaL-like extension and the linker between the  $\sigma_2$  and  $\sigma_4$  domains. Systematic alanine substitutions of contacting residues support this model and suggest that this interface stabilizes a compact conformation of ECF41 with low transcriptional activity. For group ECF42, DCA supports a structural homology model for their C-terminal tetratricopeptide repeat (TPR) domains and predicts an intimate contact between the first TPR-helix and the  $\sigma_4$  domain. Mutational analyses demonstrate the essentiality of the predicted interactions for ECF42 activity. These results indicate that

C-terminal extensions indeed bind and regulate the core ECF regions, illustrating the potential of DCA for discovering regulatory motifs in the ECF subfamily.

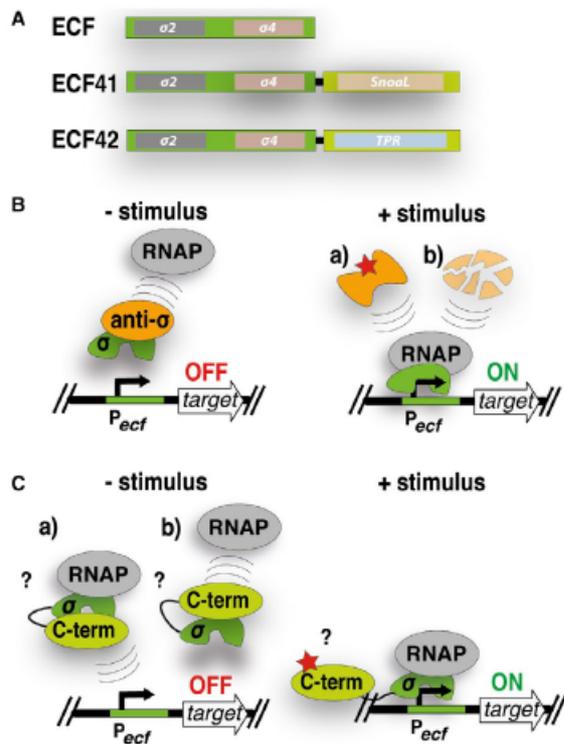
## Introduction

The communication of unicellular organisms with their extracellular environment is a key to successfully thrive in fluctuating conditions. To adapt gene expression programs to extracytoplasmic signals, bacteria harness different types of signal transduction mechanisms, of which extracytoplasmic function  $\sigma$  factors (ECFs) are among the most abundant, outnumbered only by one- and two-component systems (Staroń *et al.*, 2009). These alternative  $\sigma$  factors feature the most minimalistic domain architecture required to guide RNA polymerase (RNAP) to alternative target promoters. They only contain the conserved  $\sigma_2$  and  $\sigma_4$  domains, which are connected by a flexible, non-conserved linker (Lonetto *et al.*, 1994) (Fig. 1A). ECFs are regulated in response to a diverse array of cellular processes, including stress responses, iron uptake, differentiation, secondary metabolism, virulence and biofilm formation (Helmann, 2002; Braun *et al.*, 2003; Ho and Ellermeier, 2012; Mascher, 2013; Llamas *et al.*, 2014; Souza *et al.*, 2014). This physiological variability is reflected by their phylogenetic diversity, as evidenced by comparative genomics analyses that classified the ECF protein family into more than 90 phylogenetically distinct groups (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015; Pinto and Mascher, 2016).

The activity of ECF  $\sigma$  factors is typically controlled by anti- $\sigma$  factors encoded in the vicinity of their coding sequence, often as part of the same operon (Helmann, 2002; Staroń *et al.*, 2009; Mascher, 2013; Sineva *et al.*, 2017). Anti- $\sigma$  factors bind ECFs (Fig. 1B), thereby keeping them in an inactive conformation and – in case of membrane-anchored anti- $\sigma$  factors – often sequestering them to the proximity of the cell envelope (Mascher, 2013; Sineva *et al.*, 2017). Appropriate input signals, both from within the cytoplasm and from the extracellular space, ultimately lead to inactivating the anti- $\sigma$  factors, e.g. by conformational changes or proteolytic degradation, thereby releasing the ECF in its active conformation and

Accepted 11 April, 2019. \*For correspondence. E-mail georg.fritz@synmikro.uni-marburg.de; Tel. +49 6421 28 22582; Fax +49 6421 28 24430.

<sup>†</sup>These authors contributed equally to this work.



**Fig. 1.** Different modes of ECF regulation.

A. Domain architecture of canonical ECFs, containing only core ECF domains, and members of ECF41 and ECF42, with SnoaL-like and TPR-containing C-terminal extensions, respectively.

B. Regulation by anti- $\sigma$  factors is the most common mode of regulation of members of the ECF  $\sigma$  factor subfamily. In the absence of a stimulus the anti- $\sigma$  factor prevents contact with the RNAP. Input stimuli lead to conformational changes or degradation of the anti- $\sigma$  factor, releasing the ECF to activate transcription from its target promoters.

C. Possible  $\sigma$  factor control via C-terminal extensions. C-terminal extensions could be inhibiting the contact with the RNAP or the target promoter in the absence of a stimulus, and release their inhibition when a certain stimulus is present. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

allowing its interaction with RNAP (Ho and Ellermeier, 2012; Mascher, 2013; Sineva *et al.*, 2017). However, in addition to this archetype, there are numerous ECF groups lacking anti- $\sigma$  factors in the genomic context of the ECF, suggesting alternative modes of  $\sigma$  factor regulation (Staroń *et al.*, 2009; Mascher, 2013). In these cases, ECFs often feature group-specific microsynteny with two-component systems or protein kinases (Mascher, 2013), the former of which have proven to be the key regulators of ECF activity in these ECF groups (12, 13).

The most widely distributed theme in ECF groups lacking conserved anti- $\sigma$  factors, however, is the presence of a protein domain fused C-terminally to the  $\sigma_4$  domain of the ECF (Staroń *et al.*, 2009; Mascher, 2013; Pinto *et al.*, 2019) (Fig. 1A). Proteins in group ECF41, for instance, contain a conserved SnoaL2-like domain (Pfam: PF12680) related

to epoxide hydrolases in its ~ 200 amino acid extension (Wecke *et al.*, 2012). Proteins from group ECF42 contain a conserved tetratricopeptide repeat (TPR) related to protein-protein interactions in its ~ 120 amino acid extension (D'Andrea and Regan, 2003; Staroń *et al.*, 2009) and proteins from group ECF44 contain a conserved cysteine-rich extension of ~ 30 amino acids (Gomez-Santos *et al.*, 2011; Mascher, 2013). Additionally, a recent reclassification of the ECF  $\sigma$  factors revealed more than 15 ECF groups associated with C-terminal extensions (Casas-Pastor and Fritz, unpublished), suggesting that this is indeed a common theme in controlling ECF activity. Given the diversity of the fused protein domains, however, it is unlikely that regulation of ECFs by C-terminal extensions is exerted in a universal manner. Indeed, while the SnoaL-like extension of ECF41 seems to have a dual role as activator and inhibitor of the ECF activity (Wecke *et al.*, 2012), the TPR extension in ECF42 appears to be essential for ECF activity (Liu *et al.*, 2018). The short, cysteine-rich extension of members of ECF44 acts as an input domain that mediates the response to metal ions. It exerts a dual role in ECF activity by activating the ECF in the presence of suitable metal ions, while inhibiting ECF activity if these metals are in a different redox state (Gómez-Santos *et al.*, 2011; Marcos-Torres *et al.*, 2016).

Despite first insights into the phenomenology of  $\sigma$  factor control via C-terminal extensions, the molecular mechanisms that govern this complex regulation are largely unknown. A recently described crystal structure of SigJ from *Mycobacterium tuberculosis* revealed that the SnoaL-like extension might serve as a scaffold for the  $\sigma_2$  and  $\sigma_4$  domains in ECF41 (Goutam *et al.*, 2017), potentially sequestering the  $\sigma$  factor into an inactive conformation. But so far, there is no structural information available for ECF groups containing other C-terminal extensions that could help elucidating the details of their regulatory mechanism. The most likely hypothesis is that the regulation of  $\sigma$  factor activity results from physical contacts between the C-terminal extensions and the ECF core regions (Fig. 1C), as suggested from varying ECF-activity patterns in an array of mutant proteins with partial truncations of their C-termini (Wecke *et al.*, 2012; Liu *et al.*, 2018).

A promising approach for analyzing the putative interaction patterns between proteins is a statistical method called *direct-coupling analysis* (DCA) (Weigt *et al.*, 2009). This method is based on the fact that amino acid residues involved in intra-protein (or protein-protein) contacts tend to co-vary over the course of evolution: mutations of residues involved in critical contacts can be balanced by compensatory mutations of the interacting residues, leading to correlated variation of interacting residues in a family of homologous proteins. Additionally, the same mechanisms also apply for indirect contacts between

residues that share the same partner, but are not directly linked. In contrast to an approach that only considers mutual information between residue pairs, DCA is able to distinguish direct from indirect contacts and discards the latter (Weigt *et al.*, 2009). Another advantage of DCA is its ability of revealing the structure of unstable, transient conformations, which can be important for the interaction mechanism of the protein under study (Dago *et al.*, 2012). Since these types of conformations are normally not revealed by X-ray crystallography – as they typically capture the inactive form of the protein – DCA is a promising method for studying the molecular mechanisms underlying regulatory contacts in multi-domain proteins. Indeed, this method has been successfully applied to study the conformational changes in histidine kinases, the sensor proteins of two-component systems (Weigt *et al.*, 2009; Dago *et al.*, 2012) or the major histocompatibility complex class I in vertebrates (Dib *et al.*, 2018). A prerequisite for a good predictive power of DCA is the availability of thousands of distinct homologs of the protein under study (Weigt *et al.*, 2009). ECF41 and ECF42, the most abundant and diverse two groups in the ECF classification (Staroń *et al.*, 2009), are therefore, ideally suited for this purpose.

Here, we apply DCA to groups ECF41 and ECF42 with the aim of better understanding the regulatory role of their C-terminal extensions. As a proof of concept, we first focused on ECF41 and benchmarked the contact predictions of DCA against the crystal structure of SigJ from *M. tuberculosis* (Goutam *et al.*, 2017). We found that the regions with the highest co-variation scores reside in close proximity in the structure of SigJ and verified experimentally that these contacts are indeed important for the inactivation of Ecf41<sub>Bvj</sub> from *Bacillus licheniformis*. Next, we applied DCA to group ECF42, for which no structural information is available to date, but for which all domains can be homology-modeled based on the availability of structurally related domains. We predict that the proximal part of the TPR extension is in intimate contact with the  $\sigma_4$  domain. Strikingly, the identified point mutants in this area abolish most of the activity of Ecf42<sub>Xca</sub> of *Xanthomonas campestris*, showing the essentiality of the TPR extension. Furthermore, the DCA predictions support a homology model of the TPR domain and suggest that the conformation of the fused C-terminal extension does not clash with the interaction between ECF and core RNAP. Instead, this model suggests a possible direct or indirect contact between the C-terminal TPR domain and DNA. In summary, this work demonstrates that DCA can indeed identify important binding motifs between C-terminal extension and core of ECF groups 41 and 42, thereby guiding genetic analyses of protein function even in the absence of structural information.

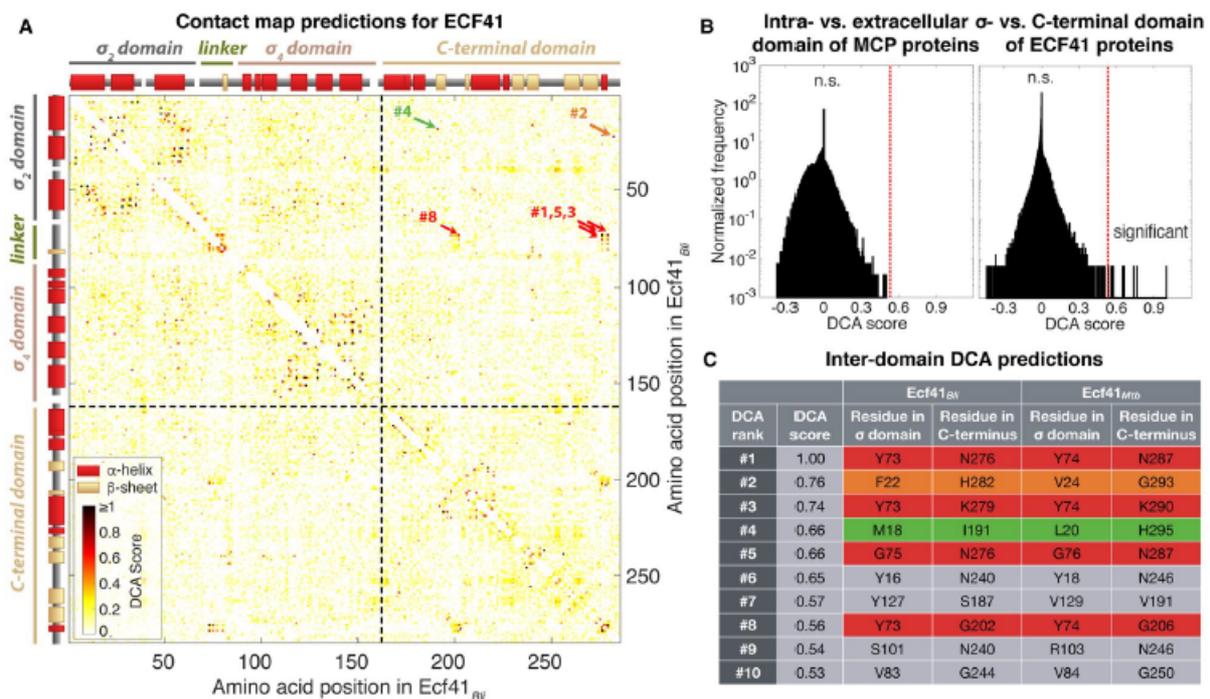
## Results

Even though most ECFs are regulated through protein–protein interactions with their cognate anti- $\sigma$  factors, some phylogenetically conserved groups of ECFs lack these and instead contain fused C-terminal extensions. A regulatory role has already been demonstrated for the C-terminal extensions of ECF41 and ECF42, representing the two largest of these ECF groups (Wecke *et al.*, 2012; Liu *et al.*, 2018). This regulatory role is likely exerted by a physical interaction of the C-termini with the ECF core. Here, we set out to applying DCA to predict the interaction interface between these regions, starting with ECF41 as the best characterized group.

### *DCA predicts contacts between the distal part of the SnoaL-like extension and the linker for ECF41*

Members of ECF41 contain a C-terminal extension of approx. 200 amino acids with a SnoaL-like domain. We applied DCA to approx. 12,000 unique sequences of ECF41 retrieved from the NCBI database. DCA assigns scores to every pair of residues according to their co-variance and, hence, the scores represent the likelihood of interaction. The contact map of ECF41 (Fig. 2A) shows high scoring pairs of residues within  $\sigma_2$  and  $\sigma_4$  domains. It also displays their secondary and tertiary structure – the anti-parallel three-helix bundle in  $\sigma_2$  and the four anti-parallel  $\alpha$ -helices of  $\sigma_4$ . Likewise, the secondary and tertiary structure of the SnoaL-like C-terminal extension was also evident from the contact map. Remarkably, the contact map shows high-scores between distant pairs of residues in the primary sequence, not only between  $\sigma_4$  and  $\sigma_2$  domain, but also between the ECF core region and the C-terminal extension, as visible in the upper-right and lower-left squares of the contact map (Fig. 2A).

First, we evaluated the significance of these results by defining a threshold score to identify and eliminate those values/pairs that are likely obtained by random chance. To this end, we applied DCA to a comparably large set of amino acid sequences from methyl-accepting chemotaxis proteins (MCP) and studied the highest DCA score observed between residues of the intra- versus extracellular domain of these transmembrane proteins, for which a direct interaction can be ruled out. The average maximum score achieved in different bootstrap samples between extracellular and intracellular regions for the MCP dataset was considered the threshold for significant co-variation (Fig. 2B). For ECF41, we found that 10 scores between ECF core domains and SnoaL-like extension were above threshold (Fig. 2B). Interestingly, four of these high-ranking predictions (rank #1, #3, #5 and #8) link a highly conserved distal part of the SnoaL-like extension



**Fig. 2.** Predicted contacts between ECF core regions and SnoL-like extension in ECF41. The colors of the contacts are maintained across the figure and when referring to predicted contacts in ECF41.

**A.** DCA-predicted contact map for ECF41. Each axis corresponds to the amino acid positions in Ecf41<sub>BW</sub>. The heatmap represents DCA scores, where darker spots correspond to higher scores. Scores > 1 are represented as 1 to allow for the observation of smaller scores. Dashed lines split the core ECF regions from the C-terminal extension. The secondary structure of Ecf41<sub>BW</sub> is represented in both axes. Significant contacts between C-terminal extension and core ECF regions are indicated by arrows.

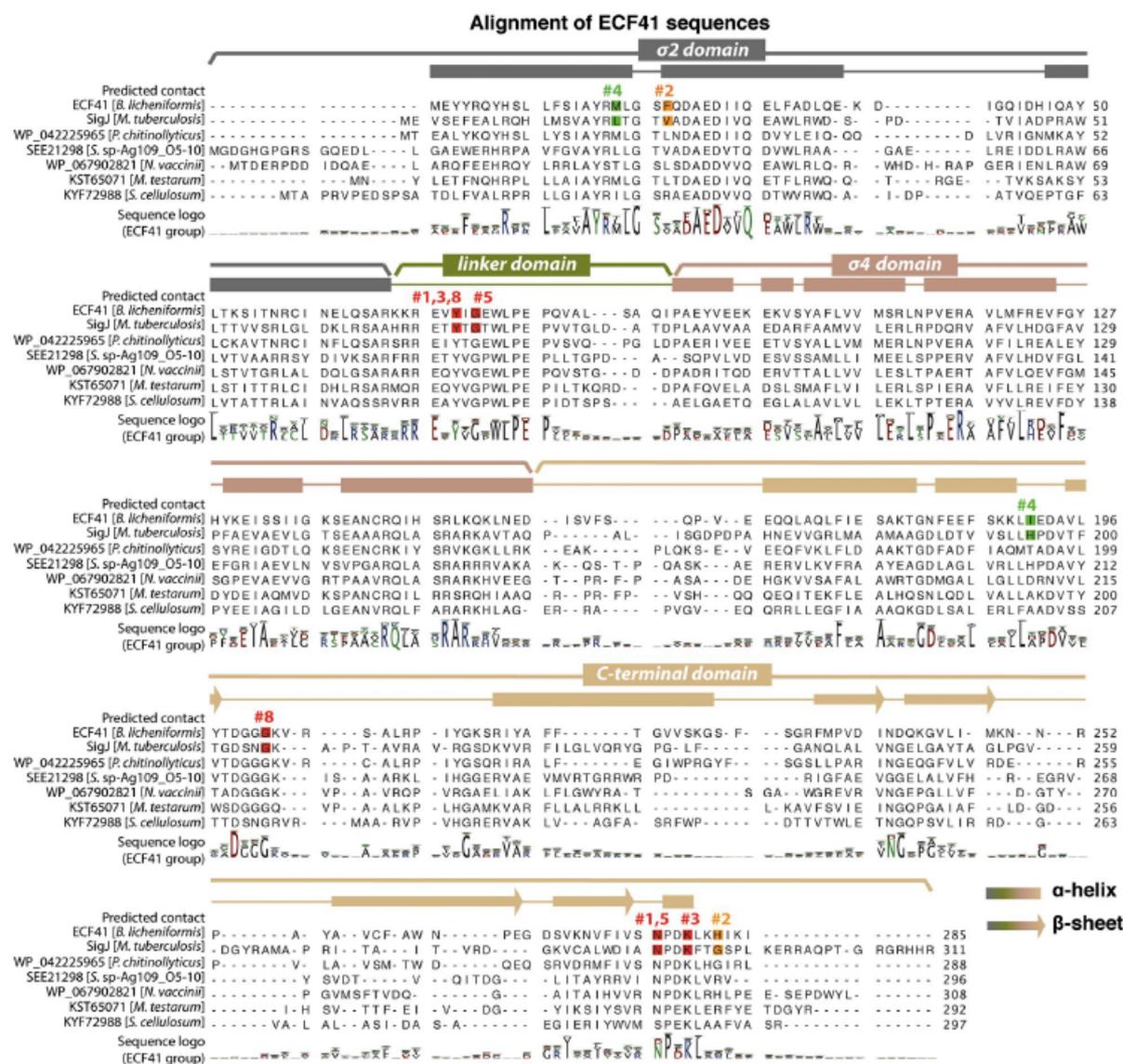
**B.** DCA score distribution for one bootstrap of MCP proteins (negative control) and for ECF41 proteins. MCP scores include only predictions between extracellular and intracellular regions, whereas ECF41 scores include only predictions between the ECF core region and the C-terminal extension. The red line, defined as the average maximum DCA score between extracellular and intracellular residues in the MCP dataset, separates significant scores from non-significant scores (threshold = 0.5318).

**C.** DCA rank and scores of the significant contacts between ECF core regions and SnoL-like extension. The residue number refers to Ecf41<sub>BW</sub> and Ecf41<sub>Mtu</sub> (also known as SigJ). The contacts selected for experimental verification were assigned to a color and correspond to the five highest ranking pairs, together with the #8 pair since the residue in the  $\sigma$  domain is shared with contacts #1 and #3. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

(consensus NPKDL) to a known ECF41-signature motif in the flexible linker between  $\sigma_2$  and  $\sigma_4$  (conserved consensus YVGPWLPEP (Wecke *et al.*, 2012)) (Figs 2C and 3; marked in red). We noted that these four predicted contact pairs mutually share at least one interacting residue. For instance, in the sequence of ECF41 from *B. licheniformis*, Ecf41<sub>BW</sub> residue Y73 in the linker region is predicted to contact G202, N276 and K279 in the C-terminus, while N276 is predicted to additionally interact with G75 – adjacent to its first contact with Y73. Besides this cluster of interactions, some high-scoring residue pairs (rank #2, #4) were also found between the C-terminal extension and the  $\sigma_2$  domain (Figs 2C and 3; marked in orange and green).

Strikingly, when mapping these co-varying residues to the three-dimensional structure of SigJ (PDB: 5XE7 (Goutam *et al.*, 2017)), we found that most of these top-scoring predictions lie in close three-dimensional

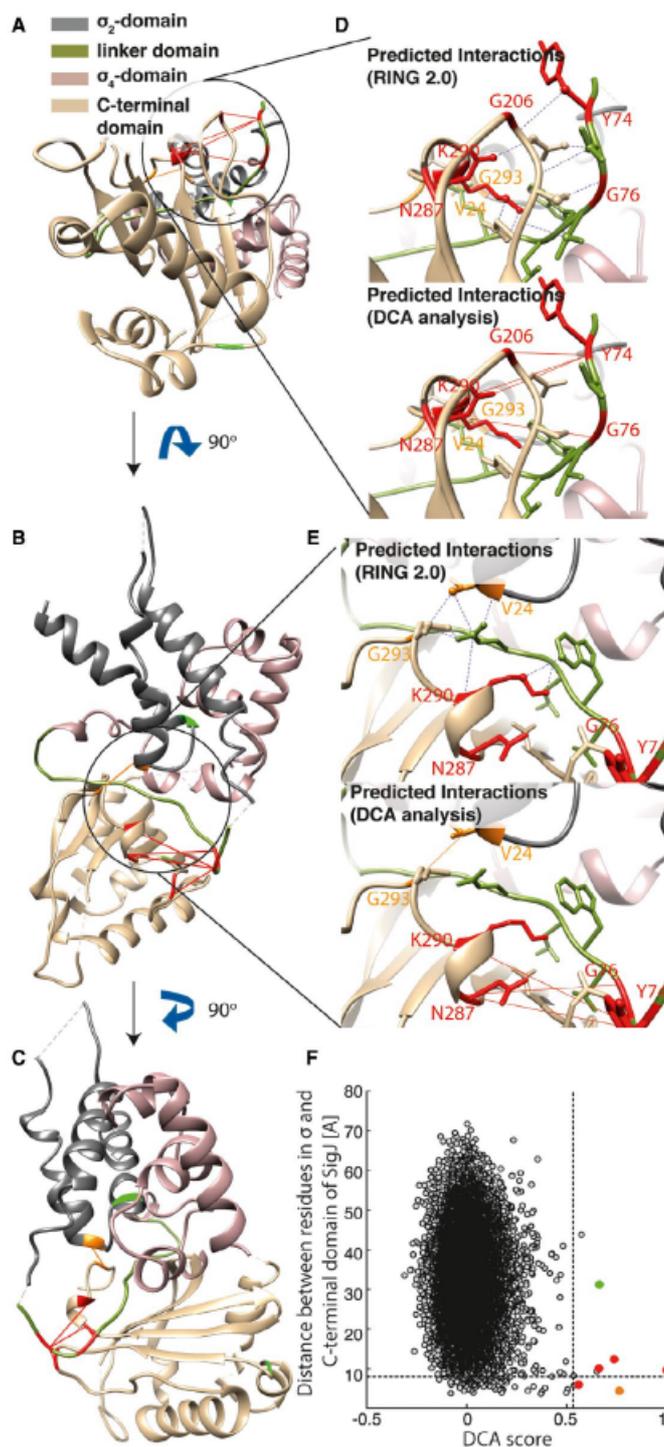
proximity – except for the pair in  $\sigma_2$  and the N-terminus of the extension (Fig. 4A–C) – supporting the idea that the predicted interactions may be in physical contact. However, when studying the correlation between the DCA score and the residue-residue distance of the C- $\alpha$  atoms in the structure of SigJ, we find that only two pairs (rank #2 and #8) are at a distance below 8 Å, which is considered to be permissive for a direct interaction (Fig. 4F). Also, despite the fact that most of the other high-scoring DCA predictions are at a distance below 12 Å in the SigJ crystal, the direct prediction of chemical bonds (using the RING 2.0 server) shows that amino acids typically contact residues immediately adjacent to the predicted DCA contact (Fig. 4D and 4). We envision four possible reasons for such slight inaccuracies: (1) The predicted residue pair could be indirectly correlated, even though DCA is designed to minimize this effect. For instance, N287 in SigJ is predicted to interact with Y74



**Fig. 3.** Mapping of DCA predictions to alignment of ECF41 sequences. The multiple-sequence alignment is shown for selected members of ECF41, together with their sequence logo derived from all 12,580 ECF41 protein sequences analyzed in this work. ECF domains, secondary structure and predicted contacting residues, together with their rank, are shown. [Colour figure can be viewed at wileyonlinelibrary.com]

and G76, but the chemical bond is formed with G206 instead, which in turn contacts Y74 (Fig. 4D). (2) The predicted contact could have an indirect role in properly adjusting the position of a highly conserved neighbor, which serves as the actual interaction partner but does not vary enough to be discovered by co-variation analysis. An example could be K290, which is predicted to interact with Y74, but it interacts instead with W78 and the backbones of L79 and P80 (Fig. 4E). (3) The *in vivo* conformation of SigJ could be slightly different from the crystal structure and more permissive for the predicted

DCA contacts. (4) The reference structure represents a single snapshot of this class of proteins, whereas DCA relies on more than 10,000 sequences. It seems plausible that if other structures were solved experimentally, the predicted contacts may be realized in those, while they may happen to not be exactly accommodated in the reference structure of SigJ. Even though we cannot discriminate between these scenarios, our results indicated that DCA is capable of identifying potentially important contacting regions, located in close physical proximity.



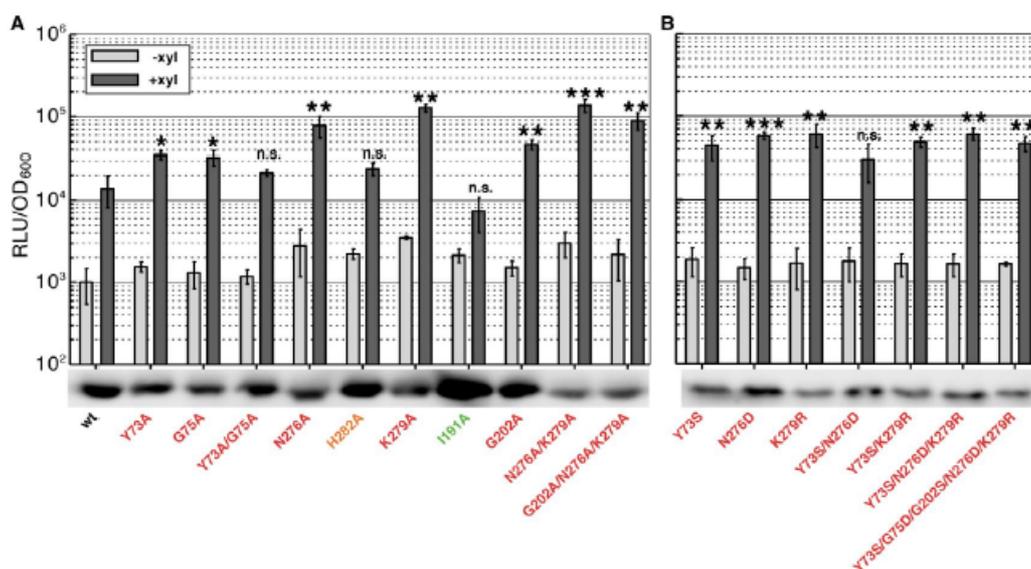
**Fig. 4.** Mapping the predicted contacts to the structure of SigJ from *M. tuberculosis* (PDB: 5XE7 (Goutam et al., 2017)). (A–C) View of the SigJ structure from different angles, colored according to domain. Contacts predicted by DCA and experimentally verified between the C-terminal extension and the core ECF regions are represented by lines. (D) and (E) Orthogonal views of the chemical interactions (dashed gray lines) as predicted by Ring 2.0 (Piovesan et al., 2016) in the environment of the predicted contacts (unbroken red lines). (F) Scatter plot of DCA score versus the distance between amino acid residues in the SigJ structure, as quantified by the distance between their C- $\alpha$  atoms. Residues with high scores are typically located at short distances, although often further apart than 8 Å. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

**Predicted contacting residues between linker and distal part of the C-terminal extension play an inhibitory role in ECF activity in members of ECF41**

To further test the functional role of putatively interacting amino acid pairs *in vivo*, we constructed an array of mutants of Ecf41<sub>BII</sub> from *B. licheniformis*, carrying single alanine substitutions of predicted interaction partners (Table S1). The functionality of these  $\sigma$  factor mutants was assessed by their ability to activate transcription from their target promoter in *B. subtilis*, leading to the transcription of a luciferase cassette as reporter. The predicted residues in the  $\sigma_2$  domain (M18 (green) and F22 (orange)) were not mutated to prevent undesired perturbation of the essential  $\sigma$  factor regions, which might lead to adverse effects in  $\sigma$  factor activity *per se*. Interestingly, mutations of the predicted residues in the linker region (Y73A and G75A, red) triggered an increased expression of the reporter (Fig. 5A). Similarly, mutations in the residues of the C-terminal extension predicted to contact the linker (G202A, N276A and K279A) led to a significantly increased expression of the reporter (Fig. 5A). Similar results were previously obtained when completely removing the C-terminal part of the SnoaL-like extension containing the NPDKL motif (Wecke *et al.*, 2012) predicted to interact with the linker. This result was confirmed in a

G202A/N276A/K279A triple mutant, but not in a Y73A/G75A double mutant (Fig. 5A), whose conformation was potentially altered in a way that it had a negative effect in ECF activity. The observation of a higher  $\sigma$  factor activity in mutants with deficient contact between linker and C-terminal extension is in line with a negative regulatory role of the C-terminal extension, inhibiting full ECF activity in response to some (unknown) physiological stimulus, as previously suggested (Wecke *et al.*, 2012). In contrast, H282A (orange) and I191A (green) mutations did not lead to a significant change in reporter expression, indicating that these putative contacts with the  $\sigma_2$  domain are not part of the inhibitory role of the C-terminal extension (Fig. 5A). Thus, together with the long distance of I191 to its predicted contacting partner (M18), our data suggests that this could be a false positive prediction, although it does not rule out the possibility that the putative contact of I191 and M18 plays a role in the stabilization of an alternative (active?) conformation between the  $\sigma$ - and C-terminal domains.

To further confirm the inhibitory interaction between the SnoaL-like extension and the linker, we next introduced compensatory mutations that aimed at restoring the interaction lost in the single amino acid substitutions. To this end, we constructed double mutants of the most important contacting pairs, Y73-N276 and Y73-K279 (red,

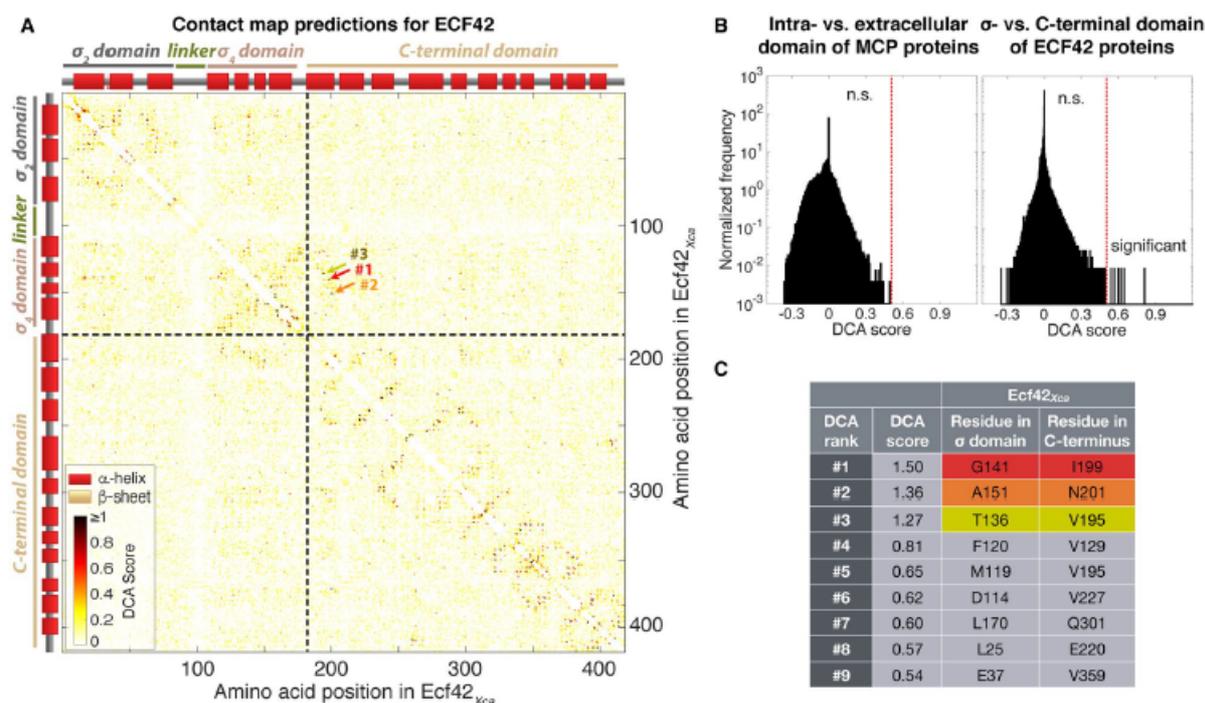


**Fig. 5.** Functional importance of top predicted amino acids in Ecf41<sub>BII</sub> in *B. subtilis*. The target promoter of Ecf41<sub>BII</sub> is controlling the expression of the luciferase cassette as a reporter of the Ecf41<sub>BII</sub> activity. The ratio between relative luminescence units (RLU) and optical density at 600nm (OD<sub>600</sub>) is shown for conditions where the expression of Ecf41<sub>BII</sub> is induced (+0.5% xyl) and uninduced (-xyl). Mean and standard deviation measured 1 h after induction in three independent replicates is shown. The significance of the change with respect to the wild type Ecf41<sub>BII</sub> is represented above the bars, where n.s. means non-significant difference, \* is *p*-value < 0.05, \*\* is *p*-value < 0.01 and \*\*\* is *p*-value < 0.001 in a two-tailed Student's *t*-test. The expression of all mutants was verified by Western blot.

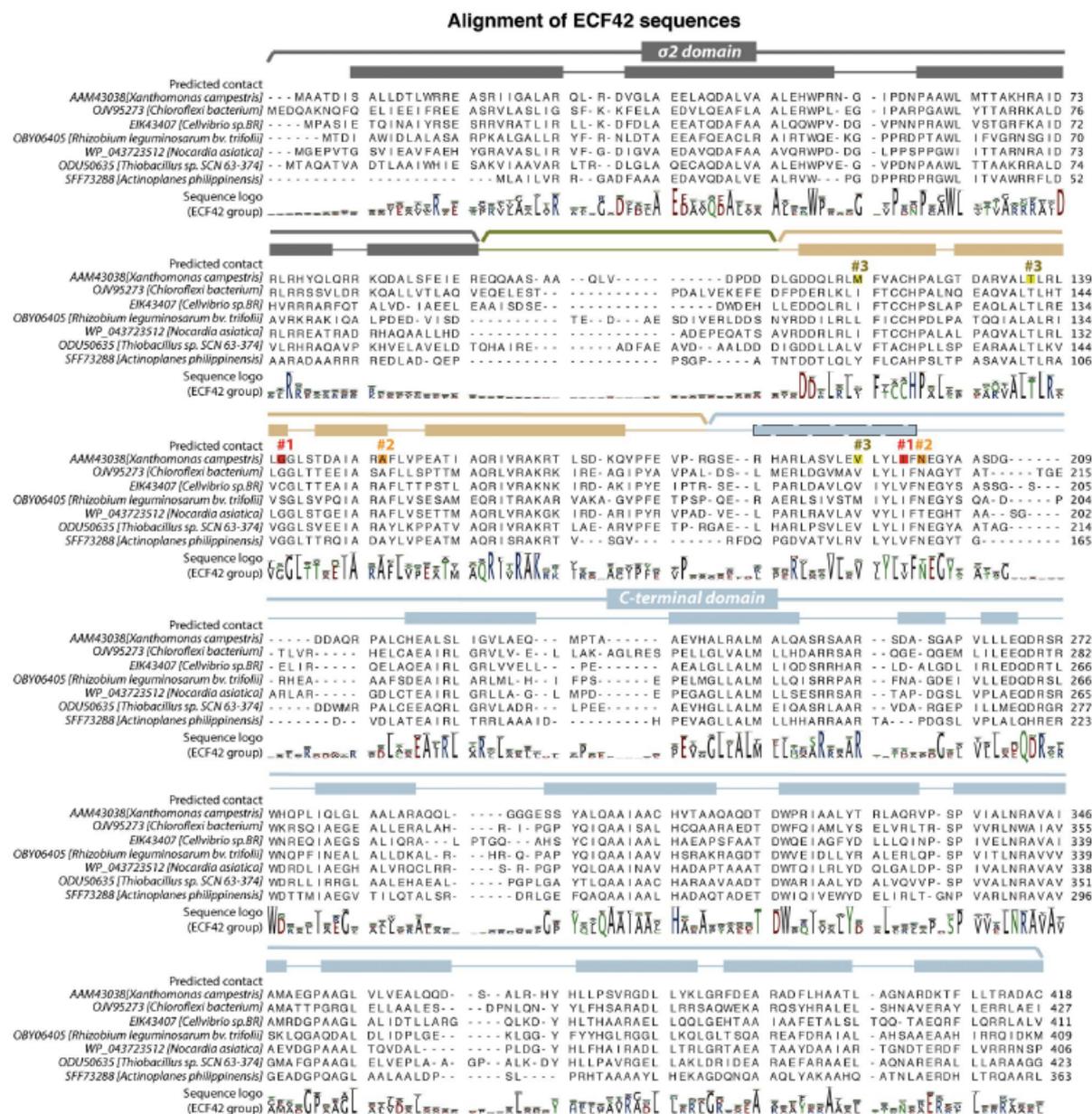
A. Assessment of the importance of the top contacts predicted by DCA (color code of the mutants is identical to the one used in Figs. 2–4). Residues of the C-terminal extension predicted to contact the linker between  $\sigma_2$  and  $\sigma_4$  exert a negative regulation over ECF activity. B. Compensatory mutations introduced in a selected pair of interacting residues with increased activity in the alanine mutant. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

rank #1 and #3 according to DCA score). We substituted these residues to the second most common amino acid pair in these positions and assessed whether the wild type levels of reporter activity were restored, this is, the interaction is functional. However, while the single amino acid substitutions Y73S, N276D and K279R led to de-repressed  $\sigma$  factor activity (Fig. 5B) similar to the substitutions with alanine (Fig. 5A), the Y73S/N276D double mutant shows an only slightly reduced target promoter activity that did not significantly differ from the wild type (Fig. 5B). Moreover, the Y73S/K279R mutant displays a significantly higher promoter output (Fig. 5B), as in the case of the single mutants, indicating that the interaction between the linker and the *SnoaL*-like extension was not recovered. This result was confirmed in the Y73S/N276D/K279R triple mutant (Fig. 5B). A possible reason is that K279 also interacts with G75 (red, rank #5 in the DCA scores), and this contact is abolished in the above-mentioned double and triple mutants. This is also the case

for Y73S, which is predicted to interact with G202 (red, rank #8 in the MSA scores). Therefore, we constructed a mutant where the complete interaction web of the five residues between the linker and the C-terminal extension was exchanged to the second most common residues, which should interact and reverse the transcription of the reporter to wild type level. Instead, the increased activity of Ecf41<sub>Blr</sub> was maintained (Fig. 5B), suggesting that the contact between the linker and extension was not recovered. This indicates that even if the predicted interactions were in true physical contact *in vivo*, the compensatory mutations introduced by our approach may be too invasive to successfully restore the conformation in which the C-terminal *SnoaL* domain binds to the linker region. Taken together, the application of DCA to Ecf41<sub>Blr</sub> confirmed its usefulness for predicting important inter-domain contacts, but clearly shows that additional factors, such as adjacently conserved motifs, have to be taken into account to gain a full understanding of the interaction interface.



**Fig. 6.** Predicted contacts between ECF core regions and TPR-containing extension in ECF42. The colors of the contacts are maintained across the figure and when referring to predicted contacts in ECF42. **A.** DCA-predicted contact map for ECF42. Each axis corresponds to the amino acids of Ecf42<sub>Xca</sub> and the heat map represents the DCA scores, where darker spots correspond to higher scores. Scores > 1 are set to 1 in order to allow for the observation of smaller scores. Dashed lines split the core ECF regions from the C-terminal extension. The predicted secondary structure of Ecf42<sub>Xca</sub> is represented in both axes. Significant contacts between C-terminal extension and core ECF regions are indicated by arrows. **B.** DCA score distribution for a selected bootstrap of the MCP dataset and for ECF42. MCP scores include only predictions between extracellular and intracellular regions, whereas ECF42 scores include only predictions between the ECF core region and the C-terminal extension. The red line, defined as the average maximum DCA score between extracellular and intracellular residues in the MCP dataset, separates significant scores from non-significant scores (threshold = 0.5048). **C.** DCA rank and scores of the significant contacts between ECF core regions and TPR-containing extensions. The residue number refers to Ecf42<sub>Xca</sub>. The top three contacts were selected for experimental verification. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]



**Fig. 7.** Mapping of DCA predictions to alignment of ECF42 sequences. The multiple-sequence alignment is shown for selected members of ECF42, together with their sequence logo derived from all 10,094 ECF42 protein sequences analyzed in this work. ECF domains, secondary structure and predicted contacting residues, together with their rank, are shown. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

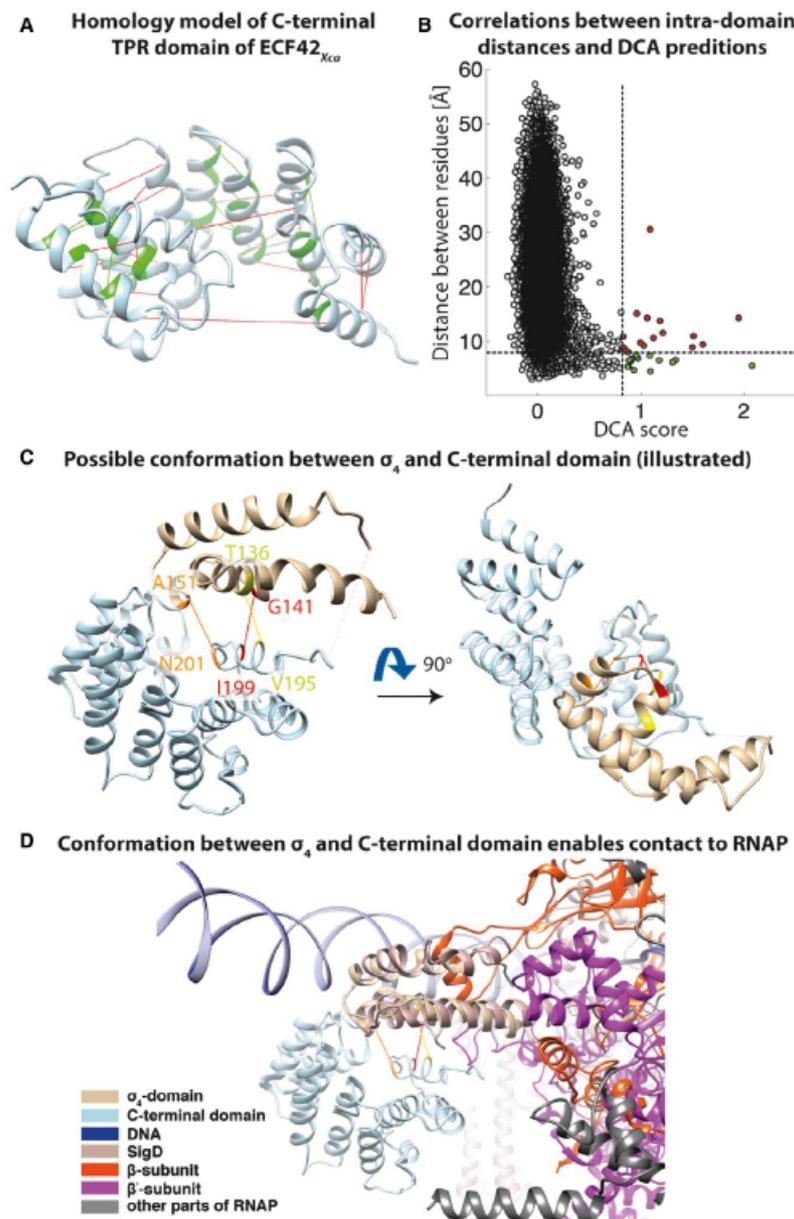
*TPR domains in C-terminal extensions of ECF42 are intimately connected to the  $\sigma_4$  region*

Next, we applied DCA to a recently studied novel ECF group, ECF42 (Liu *et al.*, 2018). Members of ECF42 are characterized by the lack of an anti- $\sigma$  factor in their genetic neighborhood and the presence of a C-terminal extension

of ~120 amino acids containing a tetratricopeptide repeat (TPR) domain. TPR domains are known to mediate protein–protein interactions (D’Andrea and Regan, 2003). We applied DCA to the approx. 10,000 unique members of ECF42 extracted from the NCBI database. The resulting contact map resembled the one for ECF41 in the  $\sigma$  factor core region, with high covariation scores within

the  $\sigma_2$  and  $\sigma_4$  domains and lower scores between the two domains (Fig. 6A). However, the contact map of the C-terminal extension appeared radically different. While

the stretches of diagonal and anti-diagonal high-scoring pairs are relatively short in the SnoaL domain of ECF41, accounting for the  $\beta$ -sheets of its secondary structure,



**Fig. 8.** Mapping the predicted contacts to the homology model of Ecf42<sub>Xca</sub>.  
 A. Mapping of the highest scoring pairs of residues to a homology model of the C-terminal extension of Ecf42<sub>Xca</sub>. Red lines correspond to predicted contacts with a distance > 8 Å in the homology model, whereas green contacts have shorter distances.  
 B. Scatterplot of DCA score versus distance between the C- $\alpha$  atoms of amino acid residues in the homology-modeled structure of the C-terminus of Ecf42<sub>Xca</sub>. The vertical line indicates the thresholding DCA score used for selecting the contacts plotted in panel (A). A distance of 8 Å was defined as the maximum residue–residue distance enabling physical contact.  
 C. Orientation of the homology-modeled structures of the C-terminal extension (blue) and the  $\sigma_4$  domain (beige) of Ecf42<sub>Xca</sub> that minimized the distance between the top three predicted residues (manual arrangement).  
 D. Structural alignment between the  $\sigma_4$  domain of Ecf42<sub>Xca</sub> and SigD from *E. coli* in the RNAP open complex (PDB: 6B6H (Liu et al., 2017)). The model of the C-terminal extension of Ecf42<sub>Xca</sub> was included in the same orientation as in panel (C), suggesting that the C-terminal extension is unlikely to interfere with the  $\sigma$  factor/RNAP binding interface, and instead suggests that it could be part of the DNA contact surface. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

these stretches are longer and richer in high-scoring contacts in the TPR extension of ECF42, indicating bundles of long, anti-parallel  $\alpha$ -helices, as further discussed below. The most striking difference is the long interaction interface between the proximal part of the C-terminal extension and the  $\sigma_4$  domain (Fig. 6A, arrows). The accumulation of high scores in this region indicates that the C-terminal extension, or at least its proximal part, is integrated into the tertiary structure of the  $\sigma_4$  domain. Using the same strategy to compute the DCA score threshold as in the case of ECF41, we found significant co-variation scores for nine putative inter-domain contacts between C-terminal extension and the ECF core regions (Fig. 6B). Interestingly, these candidate pairs of residues are located exclusively in the  $\sigma_4$  domain and in the proximal part of the TPR extension featuring the conserved consensus sequence VLYLVFNEG (Figs 6 and 7, colored in red, orange and yellow for the #1, #2 and #3 ranking predicted pairs). Given the lack of a resolved crystal structure for any member of ECF42, we individually modeled the three-dimensional structure of the core ECF domains and the TPR extension of Ecf42<sub>Xca</sub> using the I-TASSER server (Yang *et al.*, 2015). As a first test of accuracy of the DCA predictions and the modeled structure, we mapped the predicted intra-domain DCA contacts to the modeled structure of the C-terminal extension alone (Fig. 8A; residues linked with red or green lines). As in the case of ECF41, there is a negative correlation between distance in the homology model and DCA score (Fig. 8B). In some predicted pairs the distance is larger than 8 Å (Fig. 8A; red lines), but in general there is a good agreement between the modeled structure and the DCA prediction (Fig. 8A; green lines). For the inter-domain interaction between the C-terminal extension and the  $\sigma$  factor the highest scoring residues are located in the N-terminal part of the C-terminal extension, which represents an unstructured region according to I-TASSER (Figs 6C and 7). Nevertheless, its modeled structure resembles an  $\alpha$ -helix (Fig. 8A), as confirmed by a prediction of the secondary structure using J-Pred 4 (Drozdetskiy *et al.*, 2015). Applying the experience derived from the DCA-analysis of ECF41, these results indicate that the first  $\alpha$ -helix of the TPR-like C-terminal extension is integrated into the structure of the  $\sigma_4$  domain due to its interaction with the second and third  $\alpha$ -helices of the  $\sigma_4$  domain (Fig. 8C). Indeed, the predicted covarying residues are located at the same face of their corresponding  $\alpha$ -helices (Fig. 8C), further suggesting an intimate and long-ranging contact between  $\sigma_4$  and the TPR extension.

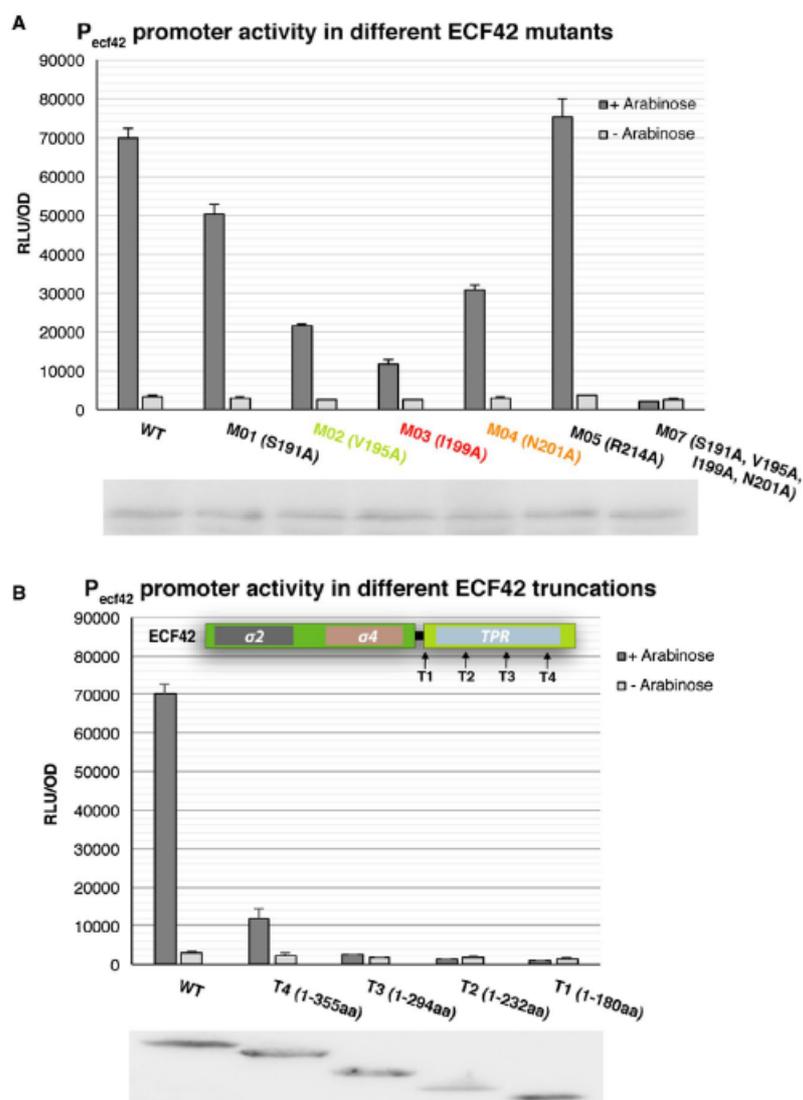
We wondered if such a persistent contact would still warrant ECF42  $\sigma$  factor activity or whether the TPR extension could be sterically clashing with the binding of the  $\sigma$  factor to the RNAP and/or promoter DNA. To investigate this hypothesis, we overlaid the modeled  $\sigma_4$  domain of

Ecf42<sub>Xca</sub> in the same orientation as the  $\sigma_4$  domain of the housekeeping  $\sigma$  factor in the structure of the RNAP open complex in *E. coli* (PDB: 6B6H (Liu *et al.*, 2017)). We manually added the modeled structure of the TPR extension in the orientation that minimized the distance between the predicted contacting residues (Fig. 8D). In this conformation, the C-terminal extension is neither interfering with the core RNAP nor the promoter. Therefore, the TPR extension of members of ECF42 might be interacting with the  $\sigma_4$  domain through its first  $\alpha$ -helix and this contact seems to be compatible with simultaneous binding of the ECF to RNAP core as well as the promoter DNA.

#### *Mutations in the residues predicted by DCA in the proximal helix and truncation of the C-terminal extension of Ecf42<sub>Xca</sub> abolish its activity*

To assign a functional role to the predicted contacting residues in the proximal helix of the TPR extension, we constructed mutants of Ecf42<sub>Xca</sub> carrying single alanine substitutions in the top predicted residues of the C-terminal extension, this is, I199A (red), N201A (orange) and V195A (yellow) (Fig. 9A), assuming that alanine interferes with the contact to the  $\sigma_4$  domain. As controls, we constructed alanine mutants in residues S191 and R214, which are also located in the proximal region of the C-terminal extension but have lower co-variation scores (ranks #24 and #344 respectively). We assessed the functionality of Ecf42 mutants according to their ability to induce transcription of the luciferase cassette when heterologously expressed in *Escherichia coli* (Fig. 9A). The results demonstrate that mutations in the three residues predicted by the DCA (I199A, N201A and V195A) led to more than 50% reduction of Ecf42<sub>Xca</sub> activity, with the strongest effect seen for the mutation of the top ranking DCA prediction (I199A), retaining only ~ 14% of the wild type activity (Fig. 9A). In contrast, the S191A mutation resulted in a much smaller loss of activity, in line with its lower covariation score, while the activity of the R214A mutant resembles, and even exceeds, that of the wild type (Fig. 9A). We next constructed an Ecf42<sub>Xca</sub> allele that combined all four mutations in the first  $\alpha$ -helix of the C-terminal extension (M07). Strikingly, this led to a complete loss of Ecf42<sub>Xca</sub> activity to background levels, indicating a cumulative effect of the point mutations (Fig. 9A). Therefore, we conclude that the predicted residues in the proximal  $\alpha$ -helix of the C-terminal extension are essential for ECF activity, in agreement with the tight contact with the  $\sigma_4$  domain that the DCA and the modeled structure of Ecf42<sub>Xca</sub> predict.

The TPR domain is located in the distal part of the C-terminal extension and previous observations demonstrated that even minor truncations of the C-terminal



**Fig. 9.** Experimental verification of DCA-predicted contacts for Ecf42<sub>Xca</sub> in *E. coli*. Activity of different alanine substitutions (A) and C-terminal truncations (B) of Ecf42<sub>Xca</sub> heterologously expressed in *E. coli*. The color code of the mutants (A) is identical to the one used in Figs. 6–8. The target promoter of Ecf42<sub>Xca</sub> controls expression of a luciferase cassette, used as a reporter for the activity of the Ecf42<sub>Xca</sub>. The relative luminescence activity, RLU/OD<sub>600</sub>, was measured 2 h after the induction of the expression of the Ecf42<sub>Xca</sub> with arabinose, and in the absence of inducer. The bars show average and standard deviation of three biological replicates. The expression level of Ecf42<sub>Xca</sub> variants and wild type was confirmed by Western blot. [Colour figure can be viewed at [wileyonlinelibrary.com](http://wileyonlinelibrary.com)]

extension lead to loss of activity in members of ECF42 in *Streptomyces venezuelae* (Liu *et al.*, 2018). We therefore generated similar truncations, T1 to T4, of Ecf42<sub>Xca</sub> (Fig. 9B). The shortest truncation T1 (1–180 aa) only contains the core  $\sigma$  regions, T2 (1–232 aa) contains the two first helices of the C-terminal extension, T3 (1–294 aa) contains about half of the C-terminal extension, and T4 (1–355 aa) only lacks the last four helices of the C-terminal extension. Transcriptional assays in *E. coli* showed that even the shortest truncation, T4, loses almost its entire activity

(~15% of the wild type), while all longer truncations are unable to induce transcription at all (Fig. 9B). These results are in perfect agreement with the previous findings for ECF42 in *S. venezuelae* (Liu *et al.*, 2018), indicating that the full C-terminal extension, and not only its proximal part, is an essential component of the ECF42  $\sigma$  factors that is required for its transcriptional activity. Whether TPR extensions undergo conformational changes as part of a modulation of the ECF activity, given the lack of anti- $\sigma$  factors in group ECF42, remains to be investigated.

## Discussion

According to our current census, tens of thousands of ECFs from at least 15 phylogenetically conserved ECF groups annotated in current databases lack obvious anti- $\sigma$  factors in their genetic context. Instead, they feature conserved protein domains fused to their C-termini. Even though their abundance and wide distribution suggest physiological importance in the bacterial world, these signaling systems have received little attention in the past and neither their inducing signals nor their cellular role have so far been identified. Accordingly, their regulatory roles as well as the mechanisms of regulation remained largely elusive. In this study, we demonstrated that computational analyses substantially help to elicit hidden information from the coevolution of contacting amino acid residues – thereby guiding the experimental design to identifying the relevant contact interfaces. Our combined theoretical and experimental results depict distinct roles for the C-terminal extensions in members of ECF41 and ECF42, which contain SnoaL-like and TPR-like protein domains, respectively, as will be discussed below.

In the case of ECF41, the distal part of the SnoaL-like extensions exhibits an inhibitory role. This inhibition seems to be a consequence of the contact between the conserved NPDKL motif of the distal part of the extension and the conserved YVGPWLPEP ECF41-signature sequence of the linker (Fig. 2; *red*), as the deletion of the contacting residues increases the transcriptional activity of Ecf41<sub>Br</sub> (Fig. 5). Previous work showed that deletion of the NPDKL motif, but not of truncations including the proximal part of the SnoaL-like extension, leads to higher ECF activity in Ecf41 from *B. licheniformis* and *Rhodobacter sphaeroides* (Wecke *et al.*, 2012). One reason for this increased ECF activity could be that the contact between the distal part of the C-terminal extension and the linker is occluding binding determinants to the RNAP or the promoter. Distal, hyperactive mutants of the SnoaL-like extension bind the core RNAP less efficiently (Wecke *et al.*, 2012), arguing in favor of a diminished promoter recognition under conditions when the linker binds the C-terminal extension. But this hypothesis requires further validation. If true, the NPDKL motif of the SnoaL-like extension could act as a switch that turns the ECF hyperactive under conditions that avoid its contact with the linker. The mechanism of ECF control described above is reminiscent of group ECF44, in which a short cysteine-rich C-terminal extension and the linker are responsible for metal binding and ECF activation (Marcos-Torres *et al.*, 2016). Unfortunately, the physiological role of ECF41 is still unknown, despite the wide distribution and abundance of this ECF group, particularly in the Actinobacteria. As a consequence, inducing conditions have not been identified so far.

Aside from the terminal part of the C-terminal extension, DCA predicted one contact between its proximal part and the  $\sigma_2$  domain (Fig. 2; *green*). This is the only predicted contact that, when disrupted, slightly decreases the activity of Ecf41<sub>Br</sub> (Fig. 5A). While its distance is too far for a direct contact, at least according to the resolved structure of SigJ from *M. tuberculosis* (Fig. 4, PDB: 5XE7 (Goutam *et al.*, 2017)), the negative effect of its deletion agrees well with the reduced activity of mutants lacking the proximal part of the C-terminal extension in Ecf41 from *B. licheniformis* and *R. sphaeroides* (Wecke *et al.*, 2012). This observation was explained by their reduced affinity of such truncated ECF alleles to the core RNAP (Wecke *et al.*, 2012). Indeed, the C-terminal extension seems to be required for the compactness of the core ECF regions of group ECF41 (Goutam *et al.*, 2017). In addition to this well documented, but still poorly understood positive role of the proximal part of the C-terminal extension, the combined results from this and previous studies suggest that the contact between the NPDKL motif in the distal part of the SnoaL-like extension and the YVGPWLPEP motif in the linker, exert a negative regulatory role over members of ECF41.

In the case of group ECF42, the complete TPR-containing C-terminal extension is required for the activity of these ECFs. DCA indicates that only the proximal part of the C-terminal extension is in charge of the interaction with the  $\sigma_4$  domain, and point mutations in the top amino acids predicted by DCA are enough to abolish most of the ECF activity (Fig. 9A). This is consistent with the lack of an unstructured, flexible linker between  $\sigma_4$  domain and C-terminal extension, as revealed by the DCA contact map (Fig. 6A). Nevertheless, the rest of the TPR extension is required for ECF activity, as revealed by truncations of Ecf42<sub>Xca</sub> (Fig. 9B) and an ECF42 protein from *S. venezuelae* (Liu *et al.*, 2018). One possible explanation is that the distal parts of the extension help to recruit the ECF to the promoter region. If true, any significant distortion and/or truncation of the TPR domain would abolish  $\sigma$  factor activity. In support of this hypothesis, the putative conformation of the TPR domain relative to  $\sigma_4$  (Fig. 8D) seems to be permissive for a direct or indirect contact between the TPR domain and the promoter-proximal DNA. However, further studies will be required to unveil the importance of the distal parts of TPR extensions in members of ECF42.

Another question that remains unsolved is the role of the TPR in the regulatory mechanism of ECF42. Both our own and previous results (Liu *et al.*, 2018) demonstrated that the TPR extension is required for ECF activity and that it is tightly linked to the  $\sigma_4$  domain. Still, it seems unlikely that it acts as an anti- $\sigma$  factor *per se*. One possibility is that changes in the distal part of the extension are transmitted to the proximal part of the extension to turn off the ECF under non-inducing conditions. Alternatively, binding

of other (unknown) factors to the TPR domain could abolish its stimulating role in transcriptional regulation. If this were the case, inducing signals would inactivate ECF42 proteins, instead of turning them on as in the rest of the ECFs studied to date. Again, this assumption will require more detailed follow-up studies to be elucidated.

Previously, DCA had already been successfully established to predict direct contact interfaces between interacting protein pairs such as two-component systems in bacteria (Weigt *et al.*, 2009) or the major histocompatibility complex in vertebrates (Dib *et al.*, 2018). Also, it has been demonstrated that DCA is capable of identifying functionally important contacts between the HisKA and HATPase\_c domains of sensor histidine kinases (Dago *et al.*, 2012). Our work substantiates that DCA can serve as a powerful tool for predicting functionally critical contacting regions between two distinct domains within the same protein – in our case between the core region and C-terminal extensions of members of groups ECF41 and ECF42. The success of DCA was possible due to the large abundance and diversity of members from both groups, with ~ 12,000 proteins available for ECF41 and ~ 10,000 for ECF42. In the future, as more and more sequences will become available also for other ECF groups, we expect that similar interdisciplinary approaches can help shedding light on their regulation mechanisms.

## Experimental procedures

### Bioinformatic procedures

Protein sequences of members of ECF41 and ECF42 were retrieved from the NCBI database and homology to ECF41 and ECF42 was identified in a recent re-classification effort of ECF  $\sigma$  factors (Casas-Pastor and Fritz, unpublished). This led to 12,580 proteins with homology to ECF41 and 10,094 proteins with homology to ECF42 for multiple-sequence alignment for DCA. These multiple-sequence alignments were computed using Clustal Omega 1.2.0 (Sievers *et al.*, 2011) with default parameters. We applied Gaussian DCA with default parameters (Baldassi *et al.*, 2014), which led to an effective number of independent sequences,  $M_{\text{eff}}$  (defined in (Baldassi *et al.*, 2014)), of 294.1 for the ECF41 dataset and 262.9 for the ECF42 dataset. The number of columns was 1071 for ECF41 and 1171 for ECF42. In the DCA, we only included columns in the alignments with < 90% gaps. DCA results were mapped to the reference proteins Ecf42<sub>Xca</sub> from *X. campestris* (GenBank AAM43038.1) and Ecf41<sub>Bli</sub> from *B. licheniformis* (GenBank AAU43179.1). Positions of the alignment that correspond to a gap in the reference proteins were ignored.

In order to evaluate the significance of the DCA scores, we applied DCA to the methyl-accepting chemotaxis proteins (MCP). These proteins were extracted from a HMMER search (Finn *et al.*, 2011) using MCP1 from *E. coli* (GenBank NP\_418775.1) as the query sequence and restricted the dataset to sequences with an *E*-value < 2e-11. The search

yielded 1,017 proteins with a variable range of homology with respect to the query. The presence of two separate regions that cannot be in direct contact, this is, an intracellular and an extracellular domain of the MCP, allowed us to establish a maximum DCA score obtained by random chance. We took three random bootstraps from the MCP dataset in order to match the ratio between number of columns in the alignment and  $M_{\text{eff}}$  of the ECF41 and ECF42 datasets. The average maximum score for intracellular-extracellular scores in the three bootstraps was defined as the significance threshold. This score was 0.5318 for ECF41 and 0.5048 for ECF42.

We used the RING 2.0 server (Piovesan *et al.*, 2016) to obtain predictions for the direct chemical interactions in the crystal structure of SigJ from *M. tuberculosis* (PDB: 5XE7 (Goutam *et al.*, 2017)). The missing side chain atoms were reconstructed using the PDB Hydro server (Azuara *et al.*, 2006). The three-dimensional structure of Ecf42<sub>Xca</sub> was modeled using the I-TASSER server with default parameters (Yang *et al.*, 2015), modeling C-terminal extension and ECF core regions separately.

### Bacterial strains and growth conditions

*E. coli* DH10 $\beta$  was used for plasmid propagation. *E. coli* MK01 (Kogenaru and Tans, 2014) was used for luminescence assays. *E. coli* cells were grown at 37°C in LB broth supplemented with spectinomycin (100  $\mu\text{g ml}^{-1}$ ), ampicillin (100  $\mu\text{g ml}^{-1}$ ), kanamycin (50  $\mu\text{g ml}^{-1}$ ) or chloramphenicol (35  $\mu\text{g ml}^{-1}$ ), when appropriate. *B. subtilis* was grown in MNGE medium for transformation and the luminescence assay was carried out in MCSE medium at 37°C with agitation (Radeck *et al.*, 2013). MCSE was supplemented with selective antibiotics chloramphenicol 5  $\mu\text{g ml}^{-1}$ , erythromycin 1  $\mu\text{g ml}^{-1}$  and lincomycin 25  $\mu\text{g ml}^{-1}$  when appropriate.

### Plasmid and strain construction

For ECF41 assays, all plasmids were constructed in *E. coli* following standard cloning methods (Sambrook and Russell, 2001). The promoter-reporter plasmid was created by integrating the Ecf41<sub>Bli</sub> (GenBank AAU43179.1) target promoter,  $P_{ydfG(-146-54)}$ , into the pBS3C/*lux* reporter vector using the restriction endonucleases EcoRI and SpeI (New England BioLabs, Ipswich, MA, USA). The expression of the *flag-ecf41<sub>Bli</sub>* and its mutated alleles were under the control of the xylose inducible promoter  $P_{xyIA}$  encoded on the pBS2EP<sub>xyIA</sub> vector. The *flag-ecf41* genes were cloned into this vector using the restriction enzymes EcoRI and SpeI. Primers to amplify each construct are listed in Table S1. In order to build the final *B. subtilis* strains, the promoter-reporter plasmid was introduced into *B. subtilis* following the protocol for natural competence described elsewhere (Radeck *et al.*, 2013). Subsequently, this strain was used as host to transform the individual expression plasmids carrying the different *flag-ecf41<sub>Bli</sub>* alleles. The activity of the Flag-Ecf41<sub>Bli</sub> variants was measured as luminescence output generated by the Ecf41<sub>Bli</sub> target promoter in a luciferase assay.

For ECF42 assays, vectors from the MoClo system were used for general cloning and assemblage of transcriptional units (Werner *et al.*, 2014) using the protocol described elsewhere (Pinto *et al.*, 2018). MoClo-adapted level 0 parts were obtained from (Pinto *et al.*, 2018) or adapted from (Rhodius *et al.*, 2013) in the case of Ecf42<sub>Xca</sub> from *X. campestris* (GenBank AAM43038.1). The different truncations of the C-terminal extension of Ecf42<sub>Xca</sub>, T1 (1-180aa), T2 (1-232aa), T3 (1-294aa) and T4 (1-355aa), were obtained by PCR amplification from pVRa42\_4454 (Rhodius *et al.*, 2013) with the primers listed in Table S1 (plasmid series 0-15\_ecf42\_4454\_T). The single or multiple site mutations in Ecf42<sub>Xca</sub> were performed with fusion PCR using the Phusion High-Fidelity PCR Kit (NEB) with the primers listed in Table S1 (plasmid series 0-15\_ecf42\_4454\_M). All the alleles of Ecf42<sub>Xca</sub> contained an N-terminal 3\*FLAG-tag (Table S1). We assembled the different Ecf42<sub>Xca</sub> alleles into transcriptional units under the control of the arabinose-inducible promoter (plasmid series 1-1R). Reporter transcriptional units included the target promoter of Ecf42<sub>Xca</sub> controlling the expression of the luciferase cassette (plasmid 1-3L). 1-1R and 1-3L transcriptional units were combined in a divergent configuration and separated by insulators to avoid crosstalk (plasmid series M). The generated level M plasmids were transformed into *E. coli* MK01 for luminescence assay.

#### Luminescence assays

In order to measure the expression of the luciferase cassette in Ecf41<sub>Bli</sub> mutants, *B. subtilis* strains were inoculated 1:500 from overnight cultures into 10 ml of MCSE medium (Radeck *et al.*, 2013) without the addition of selective antibiotics. The cells were grown to an OD<sub>600</sub> = 0.2 at 37°C and then diluted to 0.05. Following this, 100 µl of each dilution were distributed into opaque 96-well plates and their growth as well as luminescence output were monitored every 5 min in a Synergy™ 2 microplate reader (Biotek, Winooski, VT, USA) controlled by the Gen5 software. After 1 h, the strains were induced with 0.5% xylose and luminescence and cell growth (OD<sub>600</sub>) were measured 1 h after induction and followed for at least 2 h.

For measuring the activation of the luciferase cassette by ECF42<sub>Xca</sub> mutants in *E. coli*, the same protocol employed for *B. subtilis* was used with minor modifications. In brief, an overnight culture was diluted 1:100 in fresh LB without antibiotics, grown at 37°C for 2 h, and diluted to OD<sub>600</sub> = 0.01 with fresh LB medium. 100 µl of this culture were then transferred to opaque 96-well plates. Cultures were induced with 0.2% arabinose after 1 h at 37°C. Luminescence and cell growth (OD<sub>600</sub>) were measured 2 h after induction and followed for 6 h.

After the experiment, the raw luminescence output (relative luminescence units, RLU) was normalized to cell density (RLU/OD<sub>600</sub>). The OD<sub>600</sub> and RLU values were background corrected by subtracting the respective values for wells containing only medium. Both mean and standard deviation of RLU/OD<sub>600</sub> values were determined from three biological replicates. The significance of the difference between different variant was evaluated by a two-tailed Student's t-test function in GraphPad PRISM 5.

#### Western Blot analysis

We verified the expression levels of the individual Flag-Ecf41<sub>Bli</sub> and Flag-Ecf42<sub>Xca</sub> variants in order to rule out any differences in activity due to their different expression levels. In the case of *B. subtilis*, strains with FLAG-Ecf41<sub>Bli</sub> were inoculated 1:500 from overnight cultures with antibiotic selection into 25 ml of MCSE medium without antibiotics. The cultures were grown until OD<sub>600</sub> = 0.4, induced with 0.5% xylose and harvested after 1 h of incubation at 37°C. The pellets were resuspended in 400 µl buffer (20 mM Tris pH 7.5, 0.5 mM EDTA, 150 mM NaCl, 1 mM PMSF, 5% Glycerol) and disrupted by sonication. In the case of *E. coli*, strains expressing variants of FLAG-Ecf42<sub>Xca</sub> were inoculated 1:100 from overnight cultures in 10 ml of LB medium and grown at 37°C until OD<sub>600</sub> = 0.5. Cells were harvested after 2 h of induction with 0.2% arabinose and resuspended in 1.5 ml of lysis buffer (100 mM Tris-HCL PH8.0, 100 mM NaCl) for sonication. In both cases, after separating the cell debris from the soluble fraction by centrifugation, the resulting supernatants were diluted in lysis buffer to reach equal concentrations. The protein solutions were separated on a 12.5% SDS-PAGE at 200V. Subsequently, the proteins were transferred onto a polyvinylidene difluoride (PVDF) membrane (Carl Roth GmbH + Co. KG, Karlsruhe, Germany) in Towbin buffer (2.5 mM Tris pH 8.3, 19.2 mM glycine) using a PerfectBlue™ Semi-Dry Electro Blotter from Peqlab (in the case of FLAG-Ecf41<sub>Bli</sub>) or a Mini Trans-Blot® cell from Bio-Rad (in the case of FLAG-Ecf42<sub>Xca</sub>). The membrane was blocked overnight with Blotto (2.5% skim milk powder in 1x TBS (50 mM Tris-HCl pH7.6, 150 mM NaCl)) and the primary antibody (ANTI-FLAG® M2 antibody, Merck KGaA, Darmstadt, Germany) was applied for 1 h at room temperature. After four consecutive, 10 min washing steps with Blotto, the secondary antibody (Anti-Mouse IgG, HRP conjugate, Promega, Mannheim, Germany) was added and incubated for 1 h, followed by another four 10-min washing steps. Finally, the membrane was washed for 5 min in 1x TBS and the bands could be detected using the AceGlow™ chemiluminescence substrate solution (VWR International GmbH, Darmstadt, Germany) in a FluorChem™ Imager (Alpha Innotec, Kasendorf, Germany).

#### Acknowledgements

We thank Daniela Pinto for fruitful discussions and the anonymous reviewers for constructive suggestions to improve the manuscript. This work was supported by the Graduate Academy of the TU Dresden (scholarship to F.D.), the LOEWE Program of the State of Hesse (SYNMIKRO), the German Federal Ministry of Education and Research [BMBF grant 031L0010A to T.M., 031L0010B to G.F.] funded in the framework of the ERASynBio initiative and the Deutsche Forschungsgemeinschaft (DFG grant MA2837/2-2 to T.M.).

#### Data availability statement

The data that support the findings of this study are available from the corresponding author upon reasonable request.

## References

- Azuara, C., Lindahl, E., Koehl, P., Orland, H. and Delarue, M. (2006) PDB\_Hydro: incorporating dipolar solvents with variable density in the Poisson-Boltzmann treatment of macromolecule electrostatics. *Nucleic Acids Research*, **34**, W38–W42.
- Baldassi, C., Zamparo, M., Feinauer, C., Procaccini, A., Zecchina, R., Weigt, M., *et al.* (2014) Fast and accurate multivariate Gaussian modeling of protein families: predicting residue contacts and protein-interaction partners. *PLoS ONE*, **9**, e92721.
- Braun, V., Mahren, S. and Ogierman, M. (2003) Regulation of the Fecl-type ECF sigma factor by transmembrane signalling. *Current Opinion in Microbiology*, **6**, 173–180.
- Dago, A.E., Schug, A., Procaccini, A., Hoch, J.A., Weigt, M. and Szurmant, H. (2012) Structural basis of histidine kinase autophosphorylation deduced by integrating genomics, molecular dynamics, and mutagenesis. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, E1733–E1742.
- D'Andrea, L.D. and Regan, L. (2003) TPR proteins: the versatile helix. *Trends in Biochemical Sciences*, **28**, 655–662.
- Dib, L., Salamin, N. and Gfeller, D. (2018) Polymorphic sites preferentially avoid co-evolving residues in MHC class I proteins. *PLoS Computational Biology*, **14**, e1006188.
- Drozdetskiy, A., Cole, C., Procter, J. and Barton, G.J. (2015) JPred4: a protein secondary structure prediction server. *Nucleic Acids Research*, **43**, W389–W394.
- Finn, R.D., Clements, J. and Eddy, S.R. (2011) HMMER web server: interactive sequence similarity searching. *Nucleic Acids Research*, **39**, W29–W37.
- Gómez-Santos, N., Pérez, J., Sánchez-Sutil, M.C., Moraleda-Muñoz, A. and Muñoz-Dorado, J. (2011) CorE from *Myxococcus xanthus* is a copper-dependent RNA polymerase sigma factor. *PLoS Genetics*, **7**, e1002106.
- Goutam, K., Gupta, A.K. and Gopal, B. (2017) The fused SnaL\_2 domain in the *Mycobacterium tuberculosis* sigma factor  $\sigma^J$  modulates promoter recognition. *Nucleic Acids Research*, **45**, 9760–9772.
- Helmann, J.D. (2002) The extracytoplasmic function (ECF) sigma factors. *Advances in Microbial Physiology*, **46**, 47–110.
- Ho, T.D. and Ellermeier, C.D. (2012) Extra cytoplasmic function  $\sigma$  factor activation. *Current Opinion in Microbiology*, **15**, 182–188.
- Huang, X., Pinto, D., Fritz, G. and Mascher, T. (2015) Environmental sensing in Actinobacteria: a comprehensive survey on the signaling capacity of this phylum. *Journal of Bacteriology*, **197**, 2517–2535.
- Jogler, C., Waldmann, J., Huang, X., Jogler, M., Glockner, F.O., Mascher, T., *et al.* (2012) Identification of proteins likely to be involved in morphogenesis, cell division, and signal transduction in *Planctomycetes* by comparative genomics. *Journal of Bacteriology*, **194**, 6419–6430.
- Kogenaru, M. and Tans, S.J. (2014) An improved *Escherichia coli* strain to host gene regulatory networks involving both the AraC and LacI inducible transcription factors. *Journal of Biological Engineering*, **8**, 2.
- Liu, B., Hong, C., Huang, R.K., Yu, Z. and Steitz, T.A. (2017) Structural basis of bacterial transcription activation. *Science*, **358**, 947–951.
- Liu, Q., Pinto, D. and Mascher, T. (2018) Characterization of the widely distributed novel ECF42 group of extracytoplasmic function  $\sigma$  factors in *Streptomyces venezuelae*. *Journal of Bacteriology*. <https://doi.org/10.1128/JB.00437-18>
- Llamas, M.A., Imperi, F., Visca, P. and Lamont, I.L. (2014) Cell-surface signaling in *Pseudomonas*: stress responses, iron transport, and pathogenicity. *FEMS Microbiology Reviews*, **38**, 569–597.
- Lonetto, M.A., Brown, K.L., Rudd, K.E. and Buttner, M.J. (1994) Analysis of the *Streptomyces coelicolor sigE* gene reveals the existence of a subfamily of eubacterial RNA polymerase sigma factors involved in the regulation of extracytoplasmic functions. *PNAS*, **91**, 7573–7577.
- Marcos-Torres, F.J., Pérez, J., Gómez-Santos, N., Moraleda-Muñoz, A. and Muñoz-Dorado, J. (2016) In depth analysis of the mechanism of action of metal-dependent sigma factors: characterization of CorE2 from *Myxococcus xanthus*. *Nucleic Acids Research*, **44**, 5571–5584.
- Mascher, T. (2013) Signaling diversity and evolution of extracytoplasmic function (ECF)  $\sigma$  factors. *Current Opinion in Microbiology*, **16**, 148–155.
- Pinto, D., Liu, Q. and Mascher, T. (2019) ECF  $\sigma$  factors with regulatory extensions: the one-component systems of the  $\sigma$  universe. *Molecular Microbiology*. (accepted).
- Pinto, D. and Mascher, T. (2016) The ECF classification: a phylogenetic reflection of the regulatory diversity in the extracytoplasmic function  $\sigma$  factor protein family. In: *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*. de Bruijn F.J. (ed). Hoboken, NJ: John Wiley & Sons Inc, pp. 64–96.
- Pinto, D., Vecchione, S., Wu, H., Mauri, M., Mascher, T. and Fritz, G. (2018) Engineering orthogonal synthetic timer circuits based on extracytoplasmic function  $\sigma$  factors. *Nucleic Acids Research*, **296**, 1466.
- Piovesan, D., Minervini, G. and Tosatto, S.C.E. (2016) The RING 2.0 web server for high quality residue interaction networks. *Nucleic Acids Research*, **44**, W367–W374.
- Radeck, J., Kraft, K., Bartels, J., Cikovic, T., Dürr, F., Emenegger, J., *et al.* (2013) The *Bacillus* BioBrick Box: generation and evaluation of essential genetic building blocks for standardized work with *Bacillus subtilis*. *Journal of Biological Engineering*, **7**, 29.
- Rhodiou, V.A., Segall-Shapiro, T.H., Sharon, B.D., Ghodasara, A., Orlova, E., Tabakh, H., *et al.* (2013) Design of orthogonal genetic switches based on a crosstalk map of  $\sigma$ s, anti- $\sigma$ s, and promoters. *Molecular Systems Biology*, **9**, 702–702.
- Sambrook, J. and Russell, D.W. (ed.) (2001) *Molecular Cloning: A Laboratory Manual*. New York: Cold Spring Harbor Laboratory Press.
- Sievers, F., Wilm, A., Dineen, D., Gibson, T.J., Karplus, K., Li, W., *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Molecular Systems Biology*, **7**, 539–539.
- Sineva, E., Savkina, M. and Ades, S.E. (2017) Themes and variations in gene regulation by extracytoplasmic function (ECF) sigma factors. *Current Opinion in Microbiology*, **36**, 128–137.
- Souza, B.M., de Castro, T.L.P., Carvalho, R.D.O., Seyffert, N., Silva, A., Miyoshi, A., *et al.* (2014) ECF  $\sigma$  factors of gram-positive bacteria. *Virulence*, **5**, 587–600.
- Staroń, A., Sofia, H.J., Dietrich, S., Ulrich, L.E., Liesegang, H. and Mascher, T. (2009) The third pillar of bacterial signal

514 H. Wu et al.

- transduction: classification of the extracytoplasmic function (ECF)  $\sigma$  factor protein family. *Molecular Microbiology*, **74**, 557–581.
- Wecke, T., Halang, P., Staroń, A., Dufour, Y.S., Donohue, T.J. and Mascher, T. (2012) Extracytoplasmic function  $\sigma$  factors of the widely distributed group ECF41 contain a fused regulatory domain. *Microbiologyopen*, **1**, 194–213.
- Weigt, M., White, R.A., Szurmant, H., Hoch, J.A. and Hwa, T. (2009) Identification of direct residue contacts in protein-protein interaction by message passing. *Proceedings of the National Academy of Sciences of the United States of America*, **106**, 67–72.
- Werner, S., Engler, C., Weber, E., Gruetzner, R. and Marillonnet, S. (2014) Fast track assembly of multigene constructs using Golden Gate cloning and the MoClo system. *Bioengineered*, **3**, 38–43.
- Yang, J., Yan, R., Roy, A., Xu, D., Poisson, J. and Zhang, Y. (2015) The I-TASSER suite: protein structure and function prediction. *Nature Methods*, **12**, 7–8.

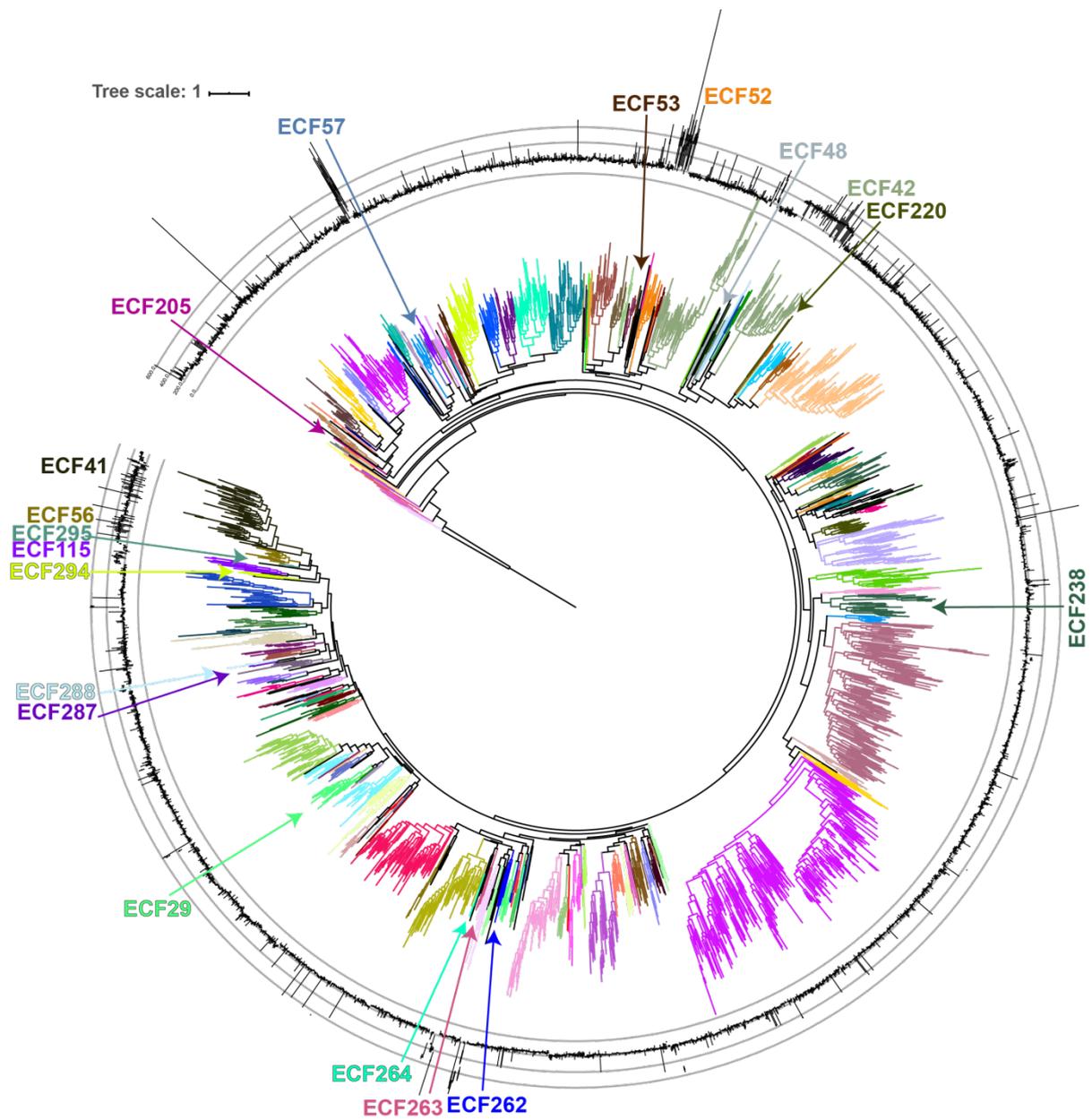
### Supporting Information

Additional supporting information may be found online in the Supporting Information section at the end of the article.

### 6.2. SnoaL-like extensions differ across ECF groups

Work by Wu and colleagues resolved the interaction map between C-terminal extensions and core regions for the largest ECF groups associated to C-terminal extensions, ECF41 and ECF42 (Wu *et al.*, 2019). This work also found that these interactions define two types of interactions, inhibitory in ECF41 and essential for activity in ECF42 (Wu *et al.*, 2019). The phylogenetic tree reflects the presence of groups with C-terminal extensions across its topology (Fig. 6.1). However, groups ECF41, ECF56, ECF295 and ECF294 share the same area of the tree and they contain a C-terminal extension with a SnoaL-like domain (Pfam: PF12680). The clustering of these groups together in the same area of the phylogenetic tree indicates that the core ECF regions experience similar modifications of their structure across SnoaL-containing groups. However, the comparison of their protein sequences showed some clear differences, which ultimately lead to the classification in four distinct groups (Fig 6.2). The most important difference from ECF41 is the lack of the NPDKL and YVGPWLPEP motifs in C-terminus and linker, respectively (Fig 6.2, grey boxes). Since these motifs interact with each other in ECF41 (Wu *et al.*, 2019), their absence in other SnoaL-containing ECF groups suggests that the function of the extension differs from ECF41. This leads to question whether the differences observed across groups are related to a different binding of the SnoaL-like C-terminal extension to the core ECF regions.

Given that the only groups with a sufficient number of non-redundant proteins are ECF41 and ECF56 (12,157 and 3,586, respectively), I applied Gaussian DCA (Baldassi *et al.*, 2014) to ECF56 and compared the results against the published data from ECF41 (Wu *et al.*, 2019). The results are reflected in a contact map, which shows the direct information, this is, the direct covariation score, between any pair of columns in the MSA of members of ECF56 (Fig. 6.3) (see Section 8.14 and (Martin Weigt *et al.*, 2009) for a broader explanation). In this contact map, the pattern of high direct information scores within the core ECF domains is similar to the published contact map of ECF41 (Wu *et al.*, 2019), although its resolution is limited, potentially due to the smaller number of proteins in ECF56 (Fig. 6.3). The pattern of high scores within the C-terminal extension in members of ECF56 is similar to the one in ECF41 (Wu *et al.*, 2019), although in ECF56 the DCA scores tend to be larger and more extended, especially in the second half of the extension (Fig. 6.3). It is unclear whether the slight differences observed between the contact maps of ECF41 and ECF56 C-terminal extensions are translated into differences in their ternary structure.



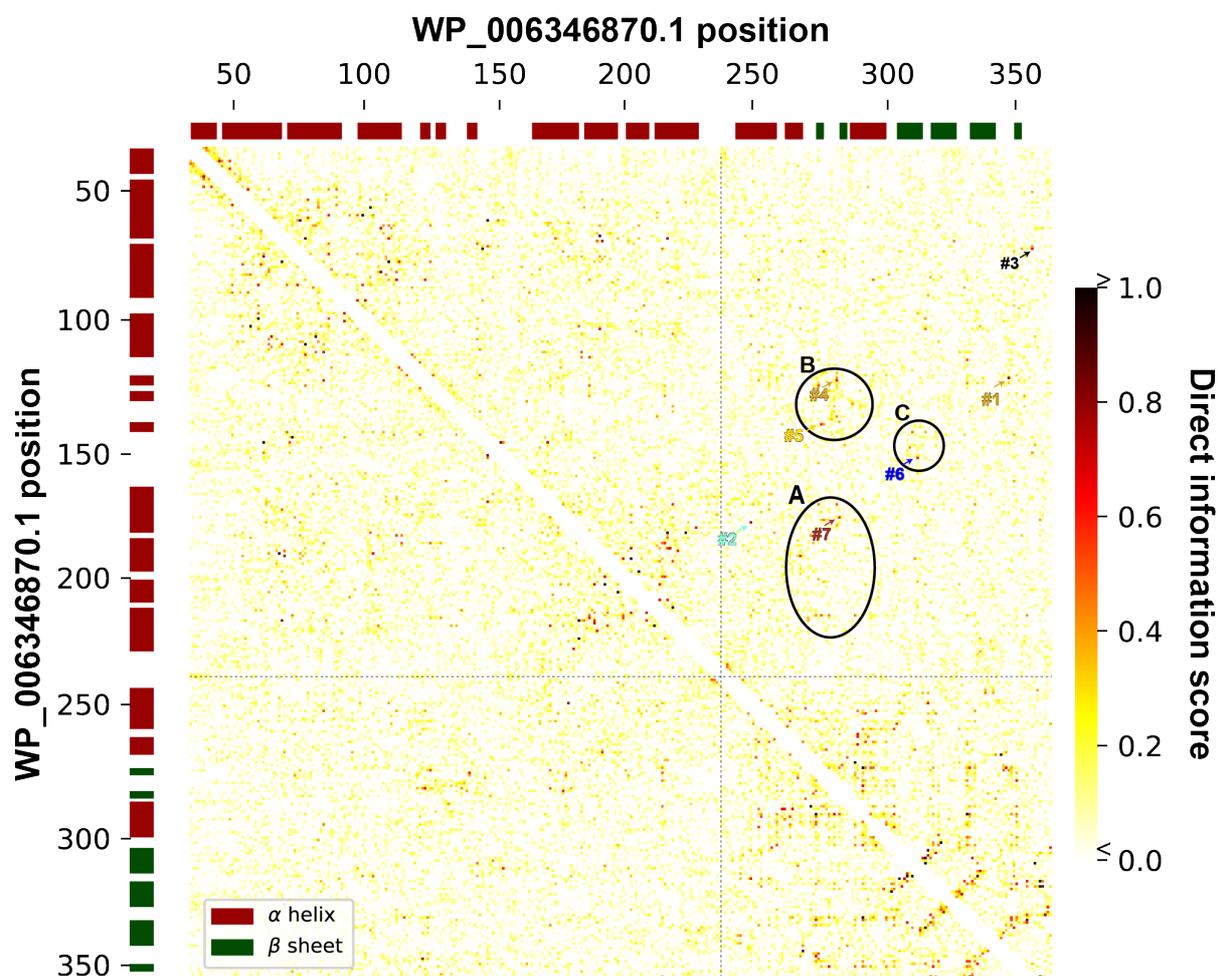
**Figure 6.1. Phylogenetic tree of ECF  $\sigma$  factors subgroups.** The length distribution of the proteins included in each subgroup is shown as a small boxplot in the ring around the tree. Groups with C-terminal extensions, supported by longer protein sequences, are labeled.



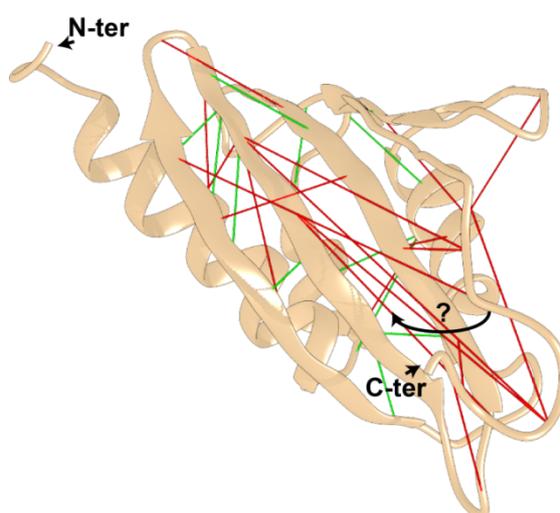
**Figure 6.2. Multiple-sequence alignment of the protein sequences from groups with a SnoaL-like C-terminal extension.** Amino acids follow RasMol color code. Two random members of each group are shown and labelled with their NCBI identifier, followed by an abbreviation of the organism of origin. Logos underneath depict the full sequence alignment of each group. Groups are sorted by number of sequences, where ECF41 is the largest group and ECF295 the smallest. The NPDKL motif, on the distal part of the C-terminal extension, and the YVGPWLPEP motif, on the linker between  $\sigma_2$  and  $\sigma_4$  domains, interact with each other in ECF41 and are marked with gray boxes. The secondary structure, plotted with green ( $\beta$  sheet) and brown ( $\alpha$  helix) boxes, is drawn according to the crystal structure of  $\sigma^I$  from *M. tuberculosis* (PDB: 5XE7 (Goutam, Gupta and Gopal, 2017)) from ECF41. The legend shows the name code of the organisms from which the protein sequences are drawn.

I tested the feasibility of the DCA results in the C-terminal region of ECF56 plotting the 30 highest scoring predicted interactions in a homology model of the structure of the C-terminal extension of WP\_006346870 from *Streptomyces tsukubaensis* (locus STSU\_11560) (Fig. 6.4). This model confirmed that the DCA results for the SnoaL-like extension of ECF56 are close to space (Fig. 6.4). However, in some cases the predictions were at a distance between  $\alpha$  carbons  $>8 \text{ \AA}$ , considered to be a largest distance to allow for a contact between amino acids, indicating that the real structure of members of ECF56 slightly differs from this structure prediction (Fig. 6.4). This is more clear at the very C-terminus of the extension, which binds to  $\sigma_4$  domain in ECF41 (Wu *et al.*, 2019). According to

the DCA results on ECF56, this C-terminus seems to bind directly on top of the  $\beta$  sheet surface of the SnoaL-like extension in ECF56 (Fig. 6.4).



**Figure 6.3. Contact map of ECF56.** Darker spots indicate larger direct information and higher probability of interaction. The axes show positions in the protein sequence WP\_006346870.1 from *S. tsukubaensis* (locus STSU\_11560). Dashed lines separate core ECF regions from C-terminus. Axes indicate the presence of  $\alpha$  helices (red) or  $\beta$  sheets (green) in the structure of WP\_006346870.1, as predicted by I-TASSER (Yang *et al.*, 2014). Areas with clusters of predictions are bound by circles (A-C). Top 7 predictions are depicted by arrows and (#1-#7). The amount of predicted contacts is larger within  $\sigma_2$ ,  $\sigma_4$  and C-terminus than between C-terminal extension and core region.



**Figure 6.4. Top 30 DCA predictions within the C-terminal extension of ECF56** plotted in the homology model of the C-terminal extension of WP\_006346870 from *S. tsukubaensis*, which was calculated with I-TASSER (Yang *et al.*, 2014). Green links show predicted

interactions with distance  $\leq 8\text{\AA}$ . Most of the predictions separated by  $>8\text{\AA}$  (red lines) seem still feasible with a slightly different configuration. The possible alternative configuration of the C-terminus is indicated by an arrow.

I found clear differences between ECF41 and ECF56 when focusing in the predicted contacts between the SnoaL-like extension and the ECF core domains. While ECF41 has clear contacts between the distal part of the extension and the linker between  $\sigma_2$  and  $\sigma_4$ , these are not present in ECF56 (Fig. 6.3). In ECF56,  $\sigma_4$  domain is predicted to be in contact with the N-terminal part of the extension (Fig. 6.3, circle A). The linker is also involved in the interaction with the extension in two positions, one equivalent to the one that contacts  $\sigma_4$  (Fig. 6.3, circle B) and another C-terminally from there (Fig. 6.3, circle C). In contrast with  $\sigma_4$  domain,  $\sigma_2$  is relatively scarce in contacts with the C-terminal extension, and the existing ones are scattered (Fig. 6.3). I plotted the top 7 predicted contacts in the homology-modelled structures of the core ECF and the C-terminal extension of WP\_006346870.1 from *S. tsukubaensis*.

**Table 6.1.** Position of the top 7 predictions in WP\_006346870.1 from *S. tsukubaensis* with their color in Fig. 6.5, their region of the ECF core where they contact and their DCA score.

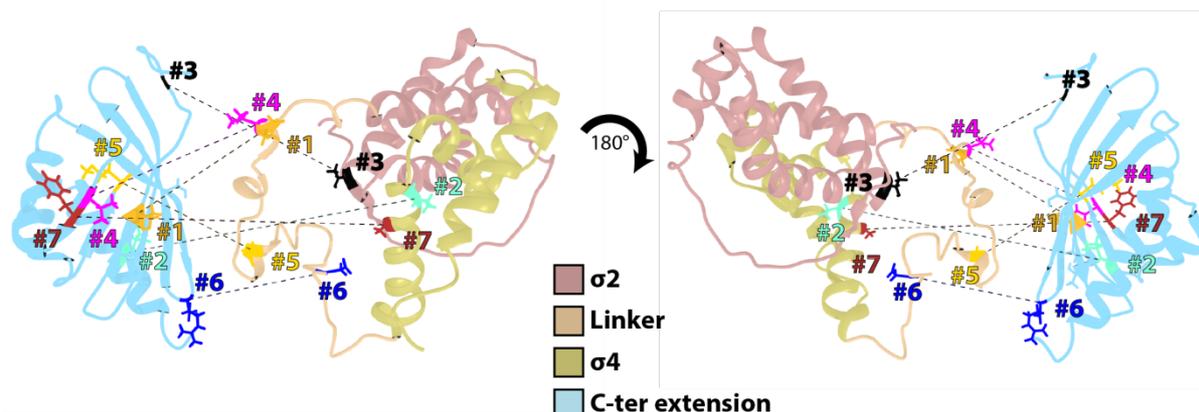
Rank	Core ECF (aa)	Extension (aa)	Color	Region	DCA score
1	123	353	goldenrod	Linker	0.78948
2	181	252	aquamarine	$\sigma_{4.1}$	0.77714
3	73	362	black	$\sigma_2$	0.65081
4	124	285	magenta	Linker	0.63232
5	141	279	gold	Linker	0.61234
6	154	318	blue	Linker	0.60816
7	179	286	brown	$\sigma_{4.1}$	0.60095

I plotted the top 7 predicted contacts (Table 6.1) in the homology-modelled structures of the core ECF regions and the C-terminal extension of WP\_006346870.1 from *S. tsukubaensis* (Fig 6.5). Most of these predictions face each other and are feasible in a more compact core ECF structure. The exceptions are predictions #2 and #5, which face the core of the C-terminal extension instead of its surface (Fig 6.5). However, most of the predictions in the interface between C-terminus, linker and  $\sigma_2$  and  $\sigma_4$  domains could be possible according to these homology models (Fig. 6.5). The equivalent region to ECF41's YVGPWLPEP linker motif (Fig. 6.1), which is essential for the contact with the distal part of the C-terminal extension in ECF41 (Wu *et al.*, 2019), is predicted to contact the central area of ECF56's C-terminal extension (Fig. 6.3 circle C, Fig. 6.5 prediction #6). This contrasts with ECF41, where the NPDKL motif, in the distal part the C-terminal extension, performs this function (Wu *et al.*, 2019). Instead of the NPDKL motif of members of ECF41, the last part of the C-terminal extension of ECF56 contains FGLP (Fig. 6.2), which is predicted to contact the residue before the DAED motif of  $\sigma_{2.2}$  region (Fig. 6.3, Fig. 6.5, prediction #3). It is possible that mutations of these 7 residues could change the activity of WP\_006346870.1. Mutations of these residues, including as negative controls mutations in other residues with low DCA score and low degree of conservation

across ECF56 (Table 6.2) are being tested in by Rute Oliveira and Dr. Marta V. Mendes, from Instituto de Investigação e Inovação em Saúde at the University of Porto.

**Table 6.2. Negative controls** for testing changes of the activity of WP\_006346870.1 when its C-terminal extension is mutated.

Position	DCA rank
S324	113
T343	171
E351	125



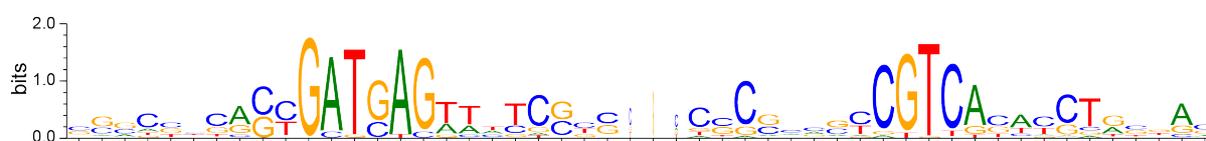
**Figure 6.5. Top 7 predicted contacts between C-terminal extension and ECF core regions of ECF56.** Predictions were plotted on the modeled structures of the core region and the C-terminal extension of WP\_006346870.1 from *S. tsukubaensis*. Predicted contacts are connected by dashed lines and labelled with the same number, which refers to their rank of DCA prediction (Table 6.1).

In conclusion, members of ECF56 seem to have a SnoaL-containing C-terminal extension with a similar fold as ECF41, although with slight changes in its most distal part, which is also divergent in sequence (Fig. 6.2, Fig. 6.4). Moreover, the inhibitory role of the SnoaL-like C-terminal extension of ECF41 over ECF activity is not so clear in ECF56, given the lack of both the predicted contacts and the amino acid sequence of its distal part and its linker (Fig. 6.2, Fig. 6.3, Fig. 6.5). Instead, most of the predicted contacts between ECF56 core and C-terminal extension occur in more proximal regions of the C-terminal extension, which are linked to several scattered areas of the ECF core region (Fig. 6.5). Interestingly, the region equivalent to the distal part of ECF41's conserved motif (consensus NPDKL) is also conserved in ECF56, although with a consensus FGLP (Fig. 5.2). This area is predicted to bind  $\sigma_2$  domain in close proximity to a region in charge of RNAP core binding (L. Li *et al.*, 2019) and with consensus DAED (Fig. 6.5, Table 6.1, Fig 5.10).

### 6.3. Finding a target promoter motif for STSU\_11560 from *Streptomyces tsukubaensis*

As part of a collaboration with the group of Rute Oliveira and Dr. Marta V. Mendes, from the University of Porto, I focused on finding the target promoter of STSU\_11560 from *S. tsukubaensis* (NCBI RefSeq WP\_006346870.1). *S. tsukubaensis* has four ECFs with a SnoaL-like extension. Two of these are from ECF41 (subgroups s2 and s21) and two from ECF56 (subgroups s2 and s3). The target of this study, STSU\_11560, is part of ECF56s3.

The putative regulon of STSU\_11560 could be found looking for its target promoter motifs in *S. tsukubaensis* genome. Even though the target promoter motif of STSU\_11560 is not defined, the ECF classification provided predictions of target promoters in ECF subgroups (Section 8.6); however, these predictions rely on ECFs being autoregulated (Staroń *et al.*, 2009; Rhodius *et al.*, 2013). In the case of ECF56s3, the promoter motif seems well defined (Fig. 6.6), which indicates that either 1) most of the members of this subgroup are autoregulated, or 2) members of this subgroup are often regulated by the same DNA binding protein. I searched for regions of the genome of *S. tsukubaensis* with similarity to ECF56s3's predicted binding motif using online Virtual Footprint (Münch *et al.*, 2005). This resulted in 2,261 hits. Of those, 183 were at a distance  $\leq 300$ bp from a coding sequence in the same direction. I derived the Pfam domains contained in the proteins encoded in the operons directly downstream of these 183 promoters and I analyzed the GO terms associated to these Pfam domains (Section 8.15). As a result, 7.45% of the proteins from *S. tsukubaensis* that were associated to ECF56s3-like promoters are linked to the GO term "DNA binding". This GO term was closely followed by "oxidation-reduction process", "regulation of transcription, DNA-templated" and "membrane", which appeared in  $\sim 6\%$  of the proteins associated to ECF56s3-like promoters. These results agree with the only member of ECF56 that has been functionally described, SigG from *M. tuberculosis*. SigG is upregulated during macrophage infection and also as part of the RecA independent DNA damage response (Gaudion *et al.*, 2013). However, SigG only targets directly two coding sequence of glyoxalases, which could be related to methylglyoxal detoxification (Gaudion *et al.*, 2013). Interestingly, STSU\_11560 does not seem part of the set of genes regulated by ECF56s3 promoter, indicating the lack of autoregulation of this ECF  $\sigma$  factor or the lack of regulation by the same transcriptional regulator as other members of ECF56s3. Supporting these data, SigG (member of ECF56s1) is not autoregulated (Gaudion *et al.*, 2013).



**Figure 6.6.** Predicted target promoter motif of members of ECF56s3. Putative -35 and -10 promoter elements are conserved.

Rute Oliveira and Dr. Marta V. Mendes, from the University of Porto, performed ChIP-seq to detect the regions of the genome directly targeted by STSU\_11560 (Oliveira *et al.*, unpublished). For ChIP-seq, STSU\_11560 and its target DNA were cross-linked and immunoprecipitated with antibodies specific of this ECF (Oliveira *et al.*, unpublished). The DNA attached to STSU\_11560 was sequenced (Johnson *et al.*, 2007). In order to find a common binding site for the areas of the genome identified by ChIP-seq, I first checked whether there is any similarity to the target predicted promoter motif of ECF56s3. An alignment of the regions upstream of transcription start site (TSS) in ChIP-seq positives and the predicted promoters for members of ECF56s3 revealed the lack of an ECF56-like promoter in

these regions (data not shown), confirming that STSU\_11560 does not target ECF56s3-like promoters. In the absence of autoregulation, the conservation of ECF56s3-like motifs upstream of members of ECF56s3 could indicate that most of the members of this subgroup are regulated by a common alternative  $\sigma$  factor or transcriptional regulator. Searching for transcriptional regulators that could bind to either the GATGAG motif (the predicted -35 element) and CGTCA motif (the predicted -10 element) of ECF56s3 promoter prediction (Fig. 6.6) using TomTom from MEME suite (Gupta *et al.*, 2007), I found that the -35 element was similar to the binding site of the transcriptional activator MalT from *E. coli* (DPInteract database: <http://arep.med.harvard.edu/dpinteract/>), and the -10 element is similar to the PhhR transcriptional activator from *S. aeruginosa* (PRODORIC 8.9 (Münch *et al.*, 2005)). Given the differences in organism and the fact that members of ECF56s3 are only found in actinobacterial genomes (Table S3.1), it is more likely that these regions are indeed -35 and -10 elements targeted by an unknown  $\sigma$  factor, or another uncharacterized transcriptional regulator. Moreover, an ECF  $\sigma$  factor could target this region since 18 of the 157 ECF groups contain CGTC in their predicted -10 element (Table S3.1).

In conclusion, members of ECF56s3 do not seem to be autoregulated, as in the case of STSU\_11560 (Oliveira *et al.*, unpublished) and SigG (Gaudion *et al.*, 2013). More analyses on ChIP-seq results would reveal the target promoter motif of STSU\_11560.

#### 6.4. Finding a conserved transcriptional response for members of ECF56s3

Given that STSU\_11560 does not seem to target the promoter predicted for its subgroup (ECF56s3), I wondered whether the target response triggered by members of ECF56s3 is conserved. For this, I assumed members of ECF56s3 target ECF56s3-like promoters. Even if member of ECF56s3 do not target this promoter motif, it appears upstream of 55 out of the 102 members of ECF56s3 from genomes tagged as “reference” or “representative”, excluding GenBank assemblies when both RefSeq and GenBank records are available. This overrepresentation indicates a common regulator that targets the so-called ECF56s3-like promoters for ~50% of the members of ECF56s3. A total of 79 reference/representative genomes contain members of ECF56s3. I searched on those genomes for ECF56s3-like sequences using Virtual Footprint (Münch *et al.*, 2005). I further considered only the 24,339 promoters at a distance  $\leq 300$ bp to an ORF in the same direction. I annotated the first ORF downstream of these promoters in order to avoid biases due to proteins with a similar function in the same operon. This annotation was done using EggNOG mapper against the library of non-supervised orthologous groups (NOGs) from Actinobacteria (Huerta-Cepas *et al.*, 2016, 2017). The most common NOGs found to be downstream of an ECF56s3-like promoter in genomes with members of ECF56s3 were:

- 1) 00BV3: featuring ATP synthase D (ATPD) found in 58% of the genomes
- 2) 00DB6: featuring a cysteine desulfurase (57% of the genomes)
- 3) 00BI8: an inner-membrane translocator (48%)

RNA-seq results would reveal whether any of these genes is differentially regulated in  $\Delta$ STSU\_11560 deletion mutant respect to the wild-type strain, confirming these results. To be true, members of ECF56s3 would be part of the same response as the genes encoded downstream of ECF56s3-like promoters conserved across genomes with members of ECF56s3.

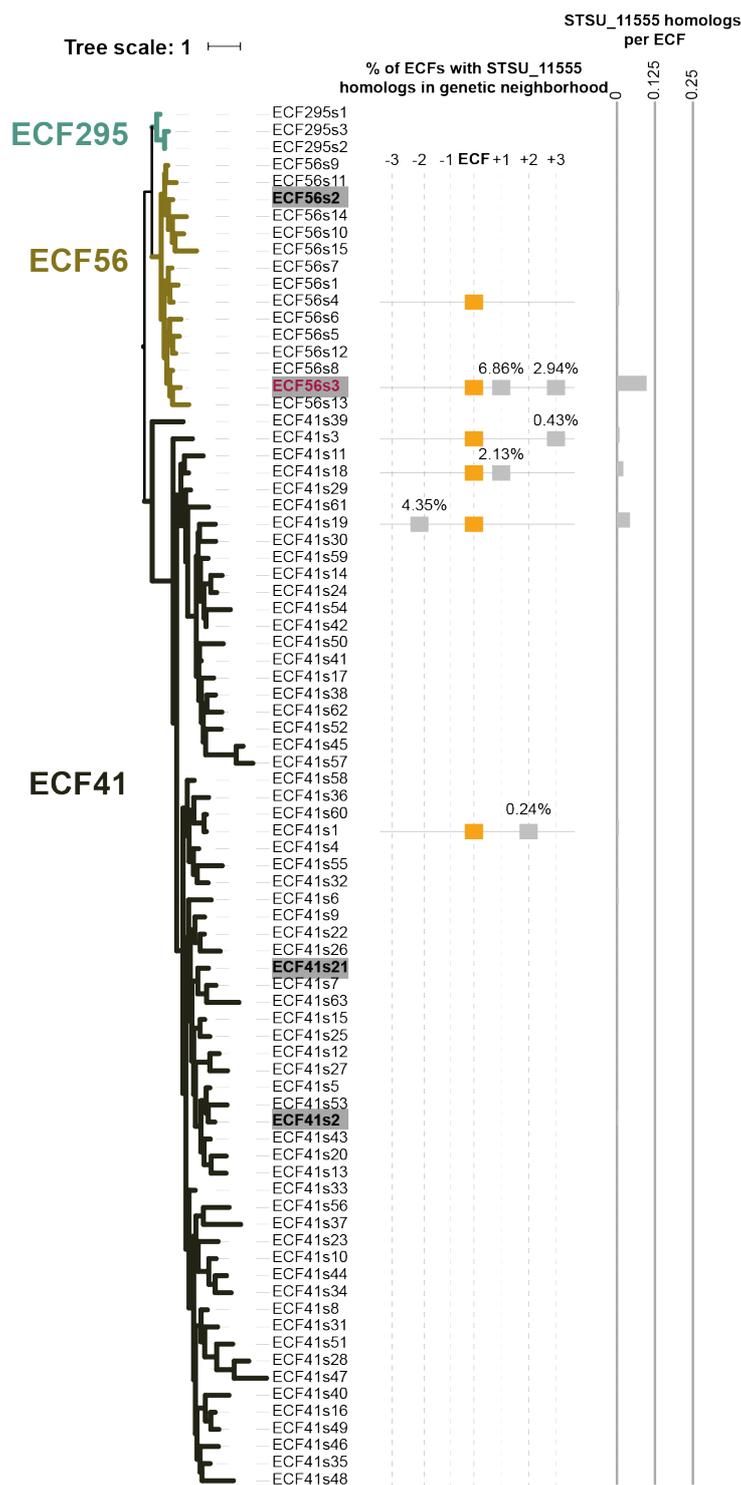
#### 6.5. Studies on the association of STSU\_11560 to the putative anti- $\sigma$ factor STSU\_11555

STSU\_11555 is encoded directly downstream of STSU\_11560 and in the same direction. The idea that STSU\_11555 could function as an anti- $\sigma$  factor emerged studies that showed the interaction between the two proteins, and from the opposite phenotype that the deletion of each gene causes (Oliveira *et al.*, unpublished).

STSU\_11555 is a hypothetical cytoplasmic protein of 154 residues with a DUF5640 domain. DUF5640 appears mostly in Firmicutes according to Pfam database (El-Gebali *et al.*, 2019). A BLAST search (Altschul *et al.*, 1990) of this protein revealed a ~27% identity and a 83% coverage of an  $\alpha/\beta$  hydrolase from *Methanosarcina* sp. MSH10X1.

I built an HMM using the alignment of the top 100 hits retrieved searching from STSU\_11555 using HHblits (Zimmermann *et al.*, 2018). This HMM was used for searching over the genetic neighborhoods of all the ECF  $\sigma$  factors in the new classification (Finn, Clements and Eddy, 2011) (see Section 8.17 for details). As a result, I found proteins with sequence similarity to STSU\_11555 in 1% of the genetic neighborhoods of members of ECF56 and 10% of the members of ECF56s3 (Fig. 6.7). In 70% of the members of ECF56s3 that are associated to STSU\_11555, this protein is encoded directly downstream of the ECF coding sequence (position +1) (Fig. 6.7). These data show that proteins similar to STSU\_11555 are rare in genetic neighborhoods of ECFs, also when looking at ECF56s3 exclusively, arguing against a conserved functional role of STSU\_11555 over ECF activity. To further analyze a potential role of STSU\_11555 as anti- $\sigma$  factor, I predicted its structure using online I-TASSER (Yang *et al.*, 2014) with default options. All the ECF anti- $\sigma$  factors crystalized so far have an  $\alpha$ -helical structure (Sineva, Savkina and Ades, 2017; Schumacher *et al.*, 2018). The structural prediction of STSU\_11555 (Yang *et al.*, 2014) showed that its structure is likely a  $\beta$  barrel (Fig. 6.10). Interestingly, the top two proteins structurally close to STSU\_11555 (PDB: 2FWVA and 2FR2A) are fatty acid-binding protein-like proteins from *M. tuberculosis*, involved in the transport and storage of small hydrophobic molecules, usually from the cell envelope (Shepard *et al.*, 2007).

These results show that most of the members of ECF56 and ECF56s3 do not contain a conserved STSU\_11555-like protein in their genetic neighborhood, indicating that the functional role of this protein is limited to few members of ECF56. Given its predicted  $\beta$ -barrel tertiary structure, STSU\_11555 could be a new type of anti- $\sigma$  factor. Future structural analysis of STSU\_11555 in complex with STSU\_11560 would shed light into the structure of this complex.



**Figure 6.7. Clade composed of ECF295, ECF56 and ECF41 extracted from the ECF phylogenetic tree (Fig 3.5).** The presence of proteins similar to the putative anti- $\sigma$  factor STSU\_11555 in the genetic neighborhood of ECFs from representative/reference organisms from each subgroup is represented by the gray boxes. ECFs are represented by orange boxes and are always depicted in the forward orientation. The average percentage of ECFs encoding STSU\_11555-like proteins in a specific position is shown. Bar plots on the right represents the average number of STSU\_11555-like proteins per ECF in the different subgroups. Subgroups that contain proteins from *S. tsukubaensis* are shaded in gray. The subgroup of STSU\_11560 is labelled in red.

## 6.6. Discussion and summary

In this section I addressed the functional role of C-terminal extensions in ECF  $\sigma$  factors. These extensions are divergent in sequence and contain different types of domains. Here, I apply DCA to

predict contacts between the C-terminal extension and the core ECF regions in members of ECF41, ECF42 and ECF56. While members of ECF42 contain tetratricopeptide repeats in their C-terminal extension, both members of ECF41 and ECF56 contain a SnoaL-like domain. In collaboration with Dr. Qiang Liu, Franziska Dürr and Prof. Thorsten Mascher from the Technische Universität Dresden, and Hao Wu, from the Philipps Universität Marburg, we revealed that the predicted contacts between the C-terminal extensions and the core domains of members of ECF41 and ECF42 have opposite effects in ECF activity (Wu *et al.*, 2019). While the distal part of the C-terminal extension of ECF41 inhibits activity, the full C-terminal extension is required by ECF function in ECF42 (Wu *et al.*, 2019). I further expanded this analysis to members of ECF56, where I predict that the SnoaL-like C-terminal extension contacts the core ECF regions differently from ECF41. Moreover, a member of ECF56, STSU\_11560 from *S. tsukubaensis*, seems to be regulated by a non-conserved anti- $\sigma$  factor, STSU\_11555, which might be a new anti- $\sigma$  factor class (Oliveira et al, unpublished). This section sheds light into the regulatory role of C-terminal extensions and suggests a multi-layered regulation involving anti- $\sigma$  factors in the case of some members of ECF56 (Oliveira et al, unpublished), as discussed in Section 7.3.

DCA results predicted an extensive contact between the proximal part of the TPR-containing C-terminal extension and the  $\sigma_4$  domain in members of ECF42, and contacts between the distal part of the SnoaL-like containing C-terminal extension of ECF41 and the linker between  $\sigma_2$  and  $\sigma_4$  domain. The evaluation of the functional role of these predictions, carried out by Dr. Qiang Liu, Franziska Dürr and Prof. Thorsten Mascher from the Technische Universität Dresden, revealed that mutations in the predicted residues of the C-terminal extension of ECF42 abolished ECF activity, whereas mutation in the predicted residues of ECF41 significantly increased ECF activity (Wu *et al.*, 2019). This agrees with previous reports for ECF41, where the deletion of the distal part of the C-terminal extension in Ecf41 from *Bacillus licheniformis* and *Rhodobacter sphaeroides* increased ECF activity, although their binding affinity to the core RNAP was reduced (Wecke *et al.*, 2012). Instead, in ECF42 the full TPR-containing C-terminal extension is essential for ECF activity (Liu, Pinto and Mascher, 2018). Further details of this study were published in Molecular Microbiology (Wu *et al.*, 2019).

Even though ECF41 and ECF56 have the same type of C-terminal extension, the protein sequence of their core  $\sigma$  regions is different, making them form two distinct groups in both the new and the original ECF classifications (Staroń *et al.*, 2009; Huang *et al.*, 2015a). These differences are reflected in the DCA results, which revealed a similar fold of the core ECF domains and C-terminal extension, but different contacts are predicted between both regions. In the case of ECF56, high scoring predictions are scattered across several regions of the C-terminal extension and the core domains, suggesting a more intimate contact between the C-terminal extension and the core area of members of ECF56. Likewise ECF41, most of the predictions for ECF56 are in the linker between  $\sigma_2$  and  $\sigma_4$  domain (Wu *et al.*, 2019). In contrast with ECF41, the only significant contact predicted in the distal part of members of ECF56 is with the  $\sigma_2$  domain, in a region involved in the contact with the core

RNAP (L. Li *et al.*, 2019). Mutation of these residues would confirm the interactions and reveal their role.

It has been suggested that the SnoaL-like domain of the C-terminal extension of members of ECF41 binds some ligand and could function as receptor in the modulation of  $\sigma^J$  activity in *M. tuberculosis* (Goutam, Gupta and Gopal, 2017). A similar role of the SnoaL-like C-terminal extension of ECF56 cannot be discarded. This could partially explain the presence of two members of ECF56 in *S. tsukubaensis*, since they could have different ligand affinities.

STSU\_11555, the protein of unknown function encoded downstream of the STSU\_11560, may be an anti- $\sigma$  factor, as preliminary experimental results suggested (Oliveira *et al.*, unpublished). The lack of conservation of this protein across members of ECF56 and its predicted tertiary structure composed of  $\beta$ -sheets argue for a new type of anti- $\sigma$  factor that is not conserved across members of ECF56. A possible reason is that STSU\_11555 could have arisen as anti- $\sigma$  factor recently in evolution. More experimental data is required to test the precise mechanism of ECF inhibition carried out by this protein. Members of ECF41 are encoded in close proximity to a flavin-containing amine oxidoreductase or a carboxymuconolactone decarboxylase (Staroń *et al.*, 2009). Given the catalytic role of these enzymes, it could be possible that STSU\_11555 also functions as an enzyme specific of the STSU\_11560 pathway. Favoring this idea, a BLAST search identified some distant similarity of STSU\_11555 to a hydrolase.

In conclusion, C-terminal extensions perform different roles in the regulation of ECF  $\sigma$  factors. The results of this section are based on DCA, which has proven to be a useful tool for the prediction of important binding interfaces that regulate ECF activity. C-terminal extensions can repress ECF activity, as in the case of ECF41 (Wu *et al.*, 2019), or be essential for activity, as in the case of ECF42 (Wu *et al.*, 2019). Even though they harbor the same domain, ECF56 C-terminal extension binds differently to the core ECF domains than in ECF41. Here I provide a list of residues that can be tested to decipher the regulation of members of ECF56 (Table 6.1). More analyses are required to understand the role of the C-terminal extension of members of ECF56 and the mechanism of regulation exerted by STSU\_11555 over STSU\_11560 activity in *S. tsukubaensis*.

## 7. Discussion and conclusion

In this work, I harness the comprehensive classification of ECF  $\sigma$  factors to analyze the regulation of ECFs in the different phylogenetic groups. For each group I provide a putative function, regulation and target promoter motif, when possible. Although most of this information is included in Table S3.1, the analysis of the conserved elements found in the genetic neighborhood of ECFs from each group and the search for experimentally addressed ECFs is not included in this thesis due to space constraints. This information will be available through the ECF hub, the web resource that will facilitate exploration and analysis of the ECF classification. This expansion of the ECF classification allowed to study the type of interactions that govern ECF  $\sigma$  factor regulation by anti- $\sigma$  factors and C-terminal extensions. Furthermore, the phosphosites targeted by STKs were predicted for several ECF groups.

### 7.1. Evolution of ECF $\sigma$ factors

The function of regulatory proteins is defined by the interaction with other elements. In order to preserve these interactions and preserve protein function, proteins co-evolve with their interaction partners (Goh *et al.*, 2000) (Section 4.1). This is because, when two proteins interact, changes in one need to be compensated in the second to sustain the interaction, modelling their protein sequences (de Juan, Pazos and Valencia, 2013). ECFs essentially interact with three partners, namely their activity regulator, their target promoter and the RNA polymerase. Therefore, these interactions condition the type of residues present in each position of ECF proteins. Indeed, ECF groups, whose members contain a similar protein sequence, generally reflect these three elements – they usually have a conserved type of regulator and target promoter motif, and they tend to be present in organisms from the same phylum, hence with a more similar RNAP. However, there are important exceptions of this homogeneity. For instance, members of group ECF39, which share a similar protein sequence and cluster together in both the founding and the current ECF classifications, are regulated either post-translationally by anti- $\sigma$  factors or at a transcriptional level by a 2CS (Section 3.4) (Pinto and Mascher, 2016). Moreover, work in the four members of ECF26 present in *S. meliloti* revealed that they tend to recognize slightly different promoter motifs (Lang *et al.*, 2018). Lastly, some ECF groups are present in a large number of taxonomic phyla, and hence are able to bind to different RNAP complexes with taxon-dependent differences (Lane and Darst, 2010b). For instance, members of group ECF41 and ECF42 are present in over 10 bacterial phyla each (Table S3.1) (Wu *et al.*, 2019).

Considering the ECF tree (Fig. 3.5), it seems that certain areas are overrepresented in ECFs from a certain bacterial phylum. For instances, the clade defined between ECF245 (light brown) and ECF30 (magenta) seems to be rich in proteins from Firmicutes (Fig. 3.5). Similarly, the clade that spans from ECF216 (light green) to ECF39 (orange) seems to be rich in ECFs from Actinobacteria (Fig. 3.5). This suggests that 1) ECFs tend to be vertically transferred, or transferred horizontally between

organisms from the same taxonomic phylum, and 2) the main factor that determines the distribution of ECF groups in the ECF tree is taxonomic origin. However, the construction of genetic circuits based on heterologous ECFs from taxonomically distant organisms (Rhodius *et al.*, 2013; Pinto *et al.*, 2018, 2019) argues against the contact with the RNAP being the major modeler of ECF protein sequences. It is possible that other taxon-specific traits aside from RNAP contact could determine ECF sequence.

Other areas of the ECF tree seem to be exceptionally taxonomically diverse. This is the case of the clade that spans from ECF111 (salmon) to ECF41 (olive green) (Fig. 3.5). In this case, other interaction partner could be more important than RNAP in shaping ECF sequence. For instance, four groups of this clade (ECF41, ECF46, ECF295 and ECF294) contain SnoaL-like C-terminal extensions of their sequence. In this case, horizontal gene transfer (HGT) could be responsible of the transfer of seemingly similar ECFs to organisms of different phyla. Higher protein similarity to proteins from another taxon respect to homologous proteins from the same taxon is a common criterion to define HGT (Koonin, Makarova and Aravind, 2001; M. Nguyen *et al.*, 2015). Nevertheless, taxonomically homogenous groups such as ECF118, ECF121 and ECF123 (Table S3.1) have also been reported to be originated by HGT (Pinto and da Fonseca, 2020), suggesting that HGT could also happen for ECFs from the same taxon and that HGT could be an important factor explaining the widespread distribution of ECFs. Gene duplication, required for the generation of new specialized versions of a protein, appears more often in genes originally inherited by HGT (Hooper and Berg, 2003). Therefore, ECF evolution could be a result of HGT, gene duplication and diversification, which could have allowed for the generation of ECF paralogs to fulfill a new function or to subspecialize in a certain role previously carried out by the original ECF.

In favor of this, several works have found an overlap between the sigmulons of the different  $\sigma$  factors in *B. subtilis* (Mascher, Hachmann and Helmann, 2007; Nicolas *et al.*, 2012) and *P. aeruginosa*, where ~30% of the coding sequences are regulated by more than one ECF (Schulz *et al.*, 2015). A partial overlap of ECF sigmulons is also observed in *S. meliloti* (Lang *et al.*, 2018). Furthermore,  $\sigma$  factors cross-regulate each other expression in *M. tuberculosis*, creating a highly-connected network, where clusters of  $\sigma$  factors related to specialized stress responses display a greater connectivity (Chauhan *et al.*, 2016). These works indicate that partially overlapping sigmulons are a common phenomenon in bacteria.

## 7.2. ECF $\sigma$ factor multiplicity

Bacterial genomes can have a really large number of  $\sigma$  factors. Considering ECFs, the record is held by *Labilithrix luteola*, with 173 ECFs in its ~12Mbp. One problem that comes with the high numbers of  $\sigma$  factors contained in the same genome is the competition for binding to core RNAP, which can cause toxicity (Malik, Zalenskaya and Goldfarb, 1987). This issue seems to be partially solved during favorable growth conditions, since most of the ECFs are likely inhibited either by their anti- $\sigma$  factor

(Section 4), by the lack of phosphorylation in  $\sigma_{2.2}$ , as observed for members of group ECF43 (Section 5), or by their C-terminal extension, as in group ECF41 (Section 6). Therefore, less ECFs are competing for binding to the RNAP. However, TPR-containing C-terminal extensions present in members of group ECF42 seem to be required for ECF activity (Wu *et al.*, 2019) (Section 6.1), raising the question of whether members of ECF42 are constitutively active or whether they are the subject of another unknown regulator that would keep them inactive when they are not needed. The latter seems to be the case of a member of ECF56, STSU\_11560 from *S. tsukubaensis*, which has been suggested to be regulated by both an anti- $\sigma$  factor and a C-terminal extension (Oliveira *et al.*, unpublished) (Section 6). Competition among  $\sigma$  factors is more prominent under adverse growth conditions. During stringent response, this competition generates the bases for the passive up-regulation of the expression of coding sequences regulated by alternative  $\sigma$  factors (Mauri and Klumpp, 2014). When amino acids become scarce, the stringent response arrests the transcription of ribosomal components, leading to the drastic reduction of cell growth rate. This makes free RNAP core complexes available for alternative  $\sigma$  factors (Bremer, Dennis and Ehrenberg, 2003; Mauri and Klumpp, 2014). Therefore, bacteria have different ways of dealing with, or harnessing, ECF multiplicity in different stages of the bacterial growth cycle.

What remains to be clarified is the type of advantage for bacterial fitness that large numbers of ECFs give. On one hand, bacteria that can deal with more environmental conditions have a clear advantage. However, additional genetic material needs more resources to be maintained. It is important to consider that genetic material that does not have a clear benefit for bacterial fitness tends to be removed (Kuo and Ochman, 2010). Although there are several models to explain how gene duplications evolve and become fixed in the population, they usually agree in that the resulting duplicated genes would need to mutate to specialize in a certain function in order to be preserved in the genome (reviewed in (Innan and Kondrashov, 2010)). Therefore, ECFs involved in the same response that are present in the same genome must hold some adaptive importance. ECFs tend to have partially overlapping sigmulons; however, most of the genes they regulate are unique, at least in *S. meliloti* and *P. aeruginosa* (Schulz *et al.*, 2015; Lang *et al.*, 2018). These sigmulon differences across ECFs in the same organism could be combined with different *ecf* expression rates across growth and development stages, and with their regulation by different factors, resulting in highly specialized ECFs even in cases where they seem to respond to the same type of stress. It would be interesting to analyze the expression patterns of different ECFs across different growth conditions, for instance using the transcriptomic data from Nicolas and colleagues (Nicolas *et al.*, 2012).

Although often ECF groups appear only once per genome, some ECF groups are particularly enriched in certain phyla (Fig. 3.7). For instance, there is an average of 5.3 copies of ECF57 per planctomycetal genome, and an average of 3.4 copies of ECF240 per Bacteroidetes genome (Fig. 3.7). In these cases, it is not clear whether the function of members of the same group is redundant in the same organism, or they rather hold specialized functions. In favor of the latter, members of ECF240,

which inherits most of its characteristics from the original FecI-like group ECF10, are involved in carbohydrate scavenging in Bacteroidetes (Martens, Koropatkin, *et al.*, 2009; Martens, Roth, *et al.*, 2009). The redundancy of members of ECF240 in the same genome is required for the activation of different Sus-like systems, that would lead to the degradation of different carbohydrates and the adaptation to different conditions (Bjursell, Martens and Gordon, 2006; Martens, Koropatkin, *et al.*, 2009). A similar case occurs in the proteobacterial group ECF243, which merges original FecI-like groups ECF05-09 and is in charge of iron uptake (Braun, Mahren and Ogierman, 2003; Staroń *et al.*, 2009). I found an average of 1.13 members of ECF243 per proteobacterial genome. However, under closer inspection, only 33% of the proteobacterial genomes contain members of ECF243, indicating that, when present, members of ECF243s are duplicated and appear 3.4 times per organism on average. Interestingly, only 8.9% of the organisms contain members of ECF243 from the same subgroup, suggesting that different subgroups fulfill different physiological functions. One possibility is that members of different subgroups detect signals from different FecR-like anti- $\sigma$  factors, which in turn, detect the presence of iron-siderophore complexes from different FecA-like transporters (see (Braun, Mahren and Ogierman, 2003) for a review). Future analyses would answer whether the different members of the same ECF group in the same genome have acquired different functions and whether this specificity is a general feature of ECF  $\sigma$  factors.

A complementary idea to explain the presence of multiple members of a group in the same organism is that new ECFs generated by gene duplication evolved and became fixed in the population to compensate for mutations that appear in the promoters of certain target genes of the original ECF. This mechanism has been proposed to explain the presence of a high number of  $\sigma$  factors (usually  $\sim 6$ ) for the transcription of  $\sim 100$  plastid genes, the so-called “spoiled kid hypothesis” (Lefebvre-Legendre *et al.*, 2014; Chi *et al.*, 2015). In this model, the high mutation rate of chloroplast promoters would be compensated by the high multiplicity of plant  $\sigma$  factors (Maier *et al.*, 2008; Chi *et al.*, 2015). Indeed, mutations in ECF regulators are suppressed by mutations in ECFs, as shown for ECFs and class I anti- $\sigma$  factors (Section 4.1). However, it is not clear that the duplication of ECFs could compensate for mutations in the promoters of elements of the sigmulon of the original ECF. This could be tested analyzing the number, group and protein sequence of ECFs found in organisms with a high rate of evolution, which would be more sensitive to changes in promoter regions. To date, the evolution rate of only 16 pathogenic bacteria is available (Duchêne *et al.*, 2016). The two order of magnitude difference in evolution rates and the lack of correlation between evolution rate and bacterial taxon (Duchêne *et al.*, 2016) makes it difficult to extrapolate evolution rates to other bacterial species.

### 7.3. New modes of regulation of ECFs

The comprehensive description of ECF groups revealed their most common types of regulators when their coding sequences share the same genetic neighborhood. Most of these regulators had already been associated to ECF groups in the original classification (Pinto and Mascher, 2016), including anti-

$\sigma$  factors, C-terminal extensions, STKs, 2CSs and transcriptional regulators (Mascher, 2013). As a novelty, this work stresses the importance of short C-terminal extensions, this is, extensions of the ECF coding sequence of less than 50aa. These elements contain conserved sequence motifs and are associated to ECF groups that lack any other type of regulator, suggesting their functional role. Indeed, the cysteine-rich, short C-terminal extension of CorE2, from *M. xanthus* (group ECF238), is involved in  $\text{Cd}^{2+}$  and  $\text{Zn}^{2+}$  recognition (Marcos-Torres *et al.*, 2016; Pérez, Muñoz-Dorado and Moraleda-Muñoz, 2018). SigZ from *B. subtilis* is also part of ECF238. SigZ is not regulated by any anti- $\sigma$  factor and the studies about its function are very limited, since its deletion is not linked to any important phenotype (Luo *et al.*, 2010). The association of SigZ with group ECF238 suggests that the two cysteine residues present in its C-terminus could have a functional role.

During the ECF reclassification I found 16 ECF groups that are not associated to a clear regulator (Fig. 3. 10). Even though it is possible to speculate about a functional role of the conserved elements encoded in their genetic neighborhoods, either in regulation or as part of the response triggered by ECF activity, the value of these data is limited without any experimental characterization of members of these groups. Interestingly, two of these groups have one described member each. These ECFs are regulated by two novel mechanisms: protein stabilization, observed for SigP from *P. gingivalis* (ECF228) (Kadowaki *et al.*, 2016), and proteolysis, observed for  $\sigma^{\text{AntA}}$  from *S. albus* (ECF282) (Seipke, Patrick and Hutchings, 2014). It was not possible to determine whether these modes of regulation are preserved in other members of the same groups; however, new experimental works could test this hypothesis. Given the singularity of these types of regulation, overlooked in the original ECF classification, it is likely that new modes of regulation would appear in the remaining 14 ECF groups with no clear regulator, or regulating ECF groups associated to other regulator as a new level of control.

One important insight of this work is that ECF groups controlled by several regulatory layers are more common than originally thought. For instance, members of ECF121 are dually regulated by anti- $\sigma$  factors and N-terminal extensions, some members of ECF12 are regulated by both anti- $\sigma$  factors and alternative promoters that generate an unstable, longer version of the ECF (Kim *et al.*, 2009), and members of ECF19, and possibly ECF18, are not only regulated by RskA-like anti- $\sigma$  factors, but also by a pair of conserved cysteine residues known to form a disulfide bridge that senses oxidative stress in SigK from *M. tuberculosis* (ECF19s1) (Shukla *et al.*, 2014). While these regulatory layers have only been deciphered for a few well-studied ECFs, they point towards the presence of several regulatory mechanisms in additional ECF groups. For instance, several ECF groups feature conserved cysteine residues potentially able to form disulfide bridges (Table S3.1), and members of ECF267 contain both a FecR-like anti- $\sigma$  factor and a conserved protein kinase in their genomic neighborhood. Given their multi-layered regulation, abundance and diversity, it could be possible that ECF  $\sigma$  factors have higher signal integration capabilities than previously anticipated.

#### 7.4. Prediction of the type of regulator that targets ECFs

One of the main outcomes of this work is the possibility of extracting hypotheses on the function, regulation and target promoter motif of uncharacterized ECFs from protein sequence data alone, as long as the ECF classifies into one of the groups defined in this work. However, when looking at the presence of certain types of regulators across the ECF phylogenetic tree, closely related ECF groups can be the target of a different type of regulator (Table S3.1, Fig. 3.5). Given that phylogenetically close ECF groups share certain protein sequence features, it is not clear how similar proteins have different types of regulators. This raises the question of whether it is possible to extract the protein sequence features that determine the binding of the ECF to the different types of regulators. If this is the case, ECF protein sequence would be enough to determine its most likely type of regulator without a previous classification of the ECF. Moreover, this would shed light into the ECF residues that are important for the regulation exerted by a certain type of protein. Among the challenges that this study may have are the multiple regulators that some types of ECFs have, as discussed in Section 7.3.

#### 7.5. Anti- $\sigma$ factor binding across the $\sigma^{70}$ family

The most common regulator of ECF  $\sigma$  factors are class I anti- $\sigma$  factors, which have a common secondary structure with four alpha helices (Campbell *et al.*, 2007; Sineva, Savkina and Ades, 2017). Even though their binding mechanism to ECFs varies in the four different structures of ECF/ASDI complexes (Fig. 4.1), in this thesis I defined the common residues that determine the binding of these two families of proteins (Section 4). In this model, ASDI helix 4 establishes a large contact interface with  $\sigma$  regions 2.1 and 2.2. This primary binding interface is preserved in all the structures of ECF/ASDI complexes (Campbell *et al.*, 2003, 2007; Shukla *et al.*, 2014; Devkota *et al.*, 2017). A second binding interface exists between a single residue of ASDI helix 1 and two residues of ECF's  $\sigma_4$  domain (Fig. 4.4). Whereas the residues in the primary binding interface seem to be less conserved, indicating the possibility that this region defines specificity within ECF/ASDI groups, the secondary binding interface is generally conserved within ECF/ASDI groups and is composed by either charged contacts or hydrophobic interactions. Importantly, these two binding interfaces appear in the two types of ASDIs defined by Paget, i. e. ASDIs that insert between  $\sigma_2$  and  $\sigma_4$ , such as RseA from *E. coli*, RskA from *M. tuberculosis* and ChrR from *R. sphaeroides*, and ASDIs that wrap around these domains, such as RsiW from *B. subtilis* (Paget, 2015). The importance of these residues in defining the specificity between ECFs and ASDIs needs to be assessed experimentally. The first step to test that these ASDI residues are important for binding to the ECF is mutating them to an amino acid with different physicochemical properties and test the ability of the mutated ASDI to inhibit ECF activity. Then, loss of ECF activity should be recovered by mutations in the appropriate residue of the partner ECF.

A question that arises with the analysis of the ECF/ASDI binding mechanism is whether this could be extended to other types of anti- $\sigma$  factors. A similar dual binding mode can be observed in the crystal structure of the ECF CnrH in complex with the class II anti- $\sigma$  factor CnrY, from *Cupriavidus metallidurans* (Maillard *et al.*, 2014). The two  $\alpha$  helices of CnrY wrap around CnrH in a conformation where CnrY's first  $\alpha$  helix mimics the function of ASDI's first helix and binds to  $\sigma_4$  domain, and CnrY's second and last  $\alpha$  helix binds to  $\sigma_2$  domain in a similar manner as ASDI's fourth helix. The only crystal structure of a class III anti- $\sigma$  factor, BldN, in complex with the ECF  $\sigma$  factor RsbN from *Streptomyces venezuelae* (Schumacher *et al.*, 2018) also shows this dual binding mode. In this case, the first and second  $\alpha$  helices of BldN bind to the  $\sigma_4$  domain, whereas its third and last  $\alpha$  helix binds to the ECF regions 2.1 and 2.2, similarly to ASDI's fourth helix, but in this case of a different RsbN molecule (Schumacher *et al.*, 2018). The similar binding between the three types of ECF anti- $\sigma$  factors is striking and contrasts with their low level of sequence similarity, which is limited to ~11% for RseA-BldN and ~3% for RseA-CnrY (using global pairwise alignments calculated by Needleman-Wunsch algorithm (Madeira *et al.*, 2019)). This may explain why, even though the same regions of the anti- $\sigma$  factor interact with a similar area of the ECF in the three types of ECF anti- $\sigma$  factors, the specific residues that carry out the interaction may differ.

It is unclear why bacteria need at least three types of ASDs. On one hand, different ASDs may provide extra specificity to ECF inhibition, which could help to reduce the apparent tendency to cross-talk of anti- $\sigma$  factors (Jamithireddy, Runthala and Gopal, 2019). On the other hand, the three types of ASDs could have emerged from different proteins and optimized their ECF inhibition by blocking the same ECF regions through convergent evolution. Future analyses that include all the ASDs known to date could help in understanding their evolution.

The anti- $\sigma$  factor FliM of the class 3 ( $\sigma_3$ -containing)  $\sigma$  factor FliA, in *E. coli*, also targets  $\sigma_2$  and  $\sigma_4$  regions with two different areas of the protein (Sorenson, Ray and Darst, 2004). However, FliM inhibition is inverted with respect to ECF anti- $\sigma$  factors. FliM is composed of four  $\alpha$  helices, of which the first and second bind to the  $\sigma_2$  domain surface, similarly to the fourth helix of ASDIs, and third and fourth helices bind to  $\sigma_4$  (Sorenson, Ray and Darst, 2004), similarly to the first helix of ASDIs. Interestingly, FliM does not bind to  $\sigma_3$  domain, supporting that the blockage of  $\sigma_2$  and  $\sigma_4$  domains is the core of  $\sigma^{70}$  inhibition. In contrast, Rsd, the anti- $\sigma$  factor of the housekeeping  $\sigma$  factor RpoD in *E. coli*, seems to target almost exclusively  $\sigma_4$  domain (Jishage, Dasgupta and Ishihama, 2001; Patikoglou *et al.*, 2007). However, this structure is based on a truncated RpoD only containing the  $\sigma_4$  domain, thus not solving whether Rsd also targets  $\sigma_2$  and whether blockage of  $\sigma_4$  is enough to provide  $\sigma$  factor inhibition.

#### 7.6. Vibrionales and Alteromonadales ECF43s could compose a new ECF group

The region between  $\sigma_{2.1}$  and  $\sigma_{2.2}$   $\alpha$  helices is longer in members of ECF43 than in ECFs that are not regulated by STKs (Section 5.1, Fig. 5.2, ring #4). The function of this extended region is unknown.

On one hand this extended region could enable the binding and recognition of the STK. In favor of this hypothesis, members of original groups ECF59 and ECF60 also have this extended region (Fig. 5.2, ring #4). Another possibility is that this region has a functional role in transcription, either binding to the DNA or to the RNAP core subunits. Supporting this, the ECF  $\sigma$  factor SigH, from *M. tuberculosis*, contacts the promoter discriminator at base G(-4) with P51, located at the end of  $\sigma_{2.1}$   $\alpha$  helix (L. Li *et al.*, 2019), suggesting a similar function for the extended region that spans between  $\sigma_{2.1}$  and  $\sigma_{2.2}$  in members of ECF43. Given that this extended region is not present in any crystalized ECF, thorough structural analyses of EcfP in contact with the RNAP will be required to determine if this region does interact with the discriminator or with any other area of the promoter or the RNAP core subunits.

Members ECF43 in Vibrionales and Alteromonadales contain an even longer extended region between  $\sigma_{2.1}$  and  $\sigma_{2.2}$   $\alpha$  helices. While this region occupies ~15aa in canonical ECF43s, it spans over ~23aa in Vibrionales and Alteromonadales ECF43s (Fig. 5.2, ring #4). Vibrionales and Alteromonadales ECF43s are part of the same phylogenetic clade, which is separated by a large evolutionary distance from the remaining ECF43s (Fig. 5.2, ring #4). Altogether, this indicates that Vibrionales and Alteromonadales ECF43s are a specialized version of ECF43 and form a new group within ECF43. Work by Gao and colleges showed that Vibrionales and Alteromonadales are located in the same clade within Gammaproteobacteria and share a common ancestor (Gao, Mohan and Gupta, 2009), suggesting that these divergent ECF43s appeared in the common ancestor of these two orders. However, other orders that are part of this clade - Aeromonadales, Pasteurellales and Enterobacteriales - lack any member of ECF43, suggesting that after the transition from canonical ECF43s to Vibrio-like ECF43s, members of ECF43s disappeared in these orders. This suggests that Vibrionales and Alteromonadales ECF43s appeared relatively recently compared to other members of ECF43.

I focused on *Vibrio* spp. in order to shed light into the evolution of Vibrionales and Alteromonadales ECF43s, to which EcfP from *V. parahaemolyticus* belongs. Proteins with similarity to its STK, PknT, are generally encoded near T6SSs in *Vibrio* spp. that lack ECF43s (Fig. 5.14). Since the amount of *Vibrio* species that contain T6SSs but lack ECF43s (14 species) is larger than the amount of species with both (7 species) or only with ECF43 (3 species) (Section 5.5), it seems that STKs associated to T6SSs appeared earlier in evolution than STKs associated to ECF43s in *Vibrio* spp. This agrees with a potential recent origin of Vibrionales and Alteromonadales ECF43s. A possible model for the evolution of STKs associated to *Vibrio* spp. ECF43s would be that first only T6SS-associated STKs were present in the genome (14 *Vibrio* spp. organisms), then ECF43s appear with their STKs (7 *Vibrio* spp.) and later some *Vibrio* spp. lost their T6SS-associated STK conserving their ECF43-STK system (3 *Vibrio* spp.). However, the data of this thesis is not enough to prove this model. Moreover, this model does not explain the evolutionary origin of Vibrionales and Alteromonadales ECF43s. One possibility is that ECF43s and STKs evolved separately. The STKs contained in the T6SS clusters

could be the evolutionary source of ECF43-associated STKs, given their sequence similarity in *Vibrio* spp. (Fig. 5.13). These kinases might have duplicated and changed their specificity towards ECF  $\sigma$  factors. However, members of ECF43, and specifically Vibrionales and Alteromonadales ECF43s, possess unique sequence features that are not present in any other ECF group. Therefore, a simultaneous acquisition of a member of ECF43 by *Vibrio* spp. is required to couple the duplicated STK to the phosphorylation of an ECF  $\sigma$  factor. Therefore, it is difficult to imagine that the ECF and the STK have a different evolutionary origin in *Vibrio* spp., which suggests that ECF43s and STKs are inherited together as a signaling module, and that the similarity between ECF43 and T6SS STKs evolved later to adjust to other unknown cellular process or component in *Vibrio* spp. There are two main possible options: 1) Vibrionales and Alteromonadales ancestors acquired both ECF43 and its associated STK, which then evolved into Vibrionales and Alteromonadales ECF43 variant, or 2) ECF43s and STKs were both horizontally transferred from other organisms after the split of Vibrionales and Alteromonadales. The clear differences between Vibrionales and Alteromonadales ECF43s and the remaining ECF43s argue in favor of the former. The comparison of the phylogenetic tree of all the members of ECF43 and the tree of their associated STKs would reveal whether these proteins are co-evolving, which would indicate that they are transferred together to new organisms.

#### 7.7. Advantages of alternative modes of regulation over anti- $\sigma$ factors

During this thesis I have discussed STKs and C-terminal extensions as alternative ECF regulators. The advantages of a phosphorylation-mediated signal transduction mechanism over anti- $\sigma$  factor sequestration are not clear. On one hand, phosphorylation is reversible (Fischer and Krebs, 1955) and could be faster than the proteolysis of an anti- $\sigma$  factor. Additionally, the promiscuity of STKs could lead to the activation of several targets (Cousin *et al.*, 2013). The ECF could also be, in turn, phosphorylated by other STKs that act in response to different stimuli. In the case of EcfP, the deletion of PknT completely abolished its phosphorylation (Fig.5.4C); however, deletion mutants of EcfP and PknT show slightly different gene expression profiles, indicating that both systems may take part in other pathways (Chandrashekar Iyer *et al.*, accepted).

The advantage of C-terminal extensions over anti- $\sigma$  factor inhibition could be related to a more tight control of ECF activity since both the regulatory domain and the core ECF are in the same protein. This arrangement, where regulatory and output domains are part of the same protein, is common in proteins from signal transduction mechanisms. For instance, the response regulators of 2CSs are fused to DNA binding domains, and 1CSs contain the sensing and the output domain in the same protein. Indeed, C-terminal extension-containing ECFs have been compared to 1CSs, since they seem to sense intracellular signals and the sensing and signal output are presumably contained in the same protein (Pinto, Liu and Mascher, 2019). Some reports have hypothesized that SnoaL-like C-terminal extensions contained in members of group ECF41 could have a similar catalytic activity as limonene-1,2-epoxide hydrolase or polyketide cyclases, given their structural similarity (Goutam, Gupta and

Gopal, 2017). A certain catalytic activity has been proven for  $\sigma^J$  from *M. tuberculosis*, a member of ECF41 (Goutam, Gupta and Gopal, 2017). Although a similar function could be possible for this extension in members of ECF56, the function of the TPR-containing C-terminal extension of members of ECF42 remains unknown. The full TPR-containing extension was required for activity when members of ECF42 were heterologously expressed (Liu, Pinto and Mascher, 2018; Wu *et al.*, 2019). Since these ECFs were not expressed in their native organisms, it is unlikely that their host had their inducer, arguing in favor of members of ECF42 being constitutively active. This raises the question on whether members of ECF42 are inhibited instead of activated as part of their regulation. If this is the case, this would be the first report of an ECF that is inactivated by a signal.

## 7.8. Limitations of this study

### 7.8.1. ECF retrieval pipeline

The study of homologous proteins relies on all of them having a common evolutionary origin and the same function. Since testing all the ECFs retrieved during this study for  $\sigma$  factor activity would require interrogating over 170,000 proteins for *in vivo* activity in their native organism, a task that is intractable nowadays, I performed their extraction through a conservative homology search. This homology search involved several quality filters required to accept a protein as putative ECF. These filters included: 1) the evaluation of the score against the general ECF HMM (Section 3.1 and 8.2), 2) both  $\sigma_2$  and  $\sigma_4$  domains present in the protein sequence, 3) lack of  $\sigma_3$  domain, and 4) lack of ambiguous amino acid characters such as X. In this way, the retrieved proteins are highly similar to prototype ECF  $\sigma$  factors and I could fairly assume that they still preserve ECF activity.

One of the advantages of this strategy is that it has a high specificity. Aside from group 3  $\sigma^{70}$ s, which were discarded by the retrieval filters, ECFs have sequence similarity to proteins that function as anti-anti- $\sigma$  factors. Anti-anti- $\sigma$  factors are typically associated to members of group ECF15 and are similar enough to the members of this group to bind to their class II anti- $\sigma$  factors, but different enough to not have ECF activity (Francez-Charlot *et al.*, 2015). Anti-anti- $\sigma$  factors contain divergent  $\sigma_2$  and/or  $\sigma_4$  domains, but in some cases they can hit Pfam models for these domains. The main difference respect to real ECFs is that they are usually fused to the response regulator of a 2CS in their N-terminus and that their linker between  $\sigma_2$  and  $\sigma_4$  domain is often shorter. While shorter linkers are difficult to identify since this area is variable in sequence and length in ECFs, the identification of response regulators fused to ECF proteins in the ECF library (>170,000 proteins) was simple and yielded only 2 ECFs with this domain. However, the area where the response regulator was located differed from the N-terminal position of known anti-anti- $\sigma$  factors. If these putative ECFs are anti-anti- $\sigma$  factor, this still supposes only <0.1% of the ECFs identified in this work.

As a consequence of this stringent selection of putative ECFs for the library expansion, the degree of sequence diversity in the extracted ECFs is limited. In particular, I noticed that two main types of ECF  $\sigma$  factors could not be captured, namely, ECF  $\sigma$  factors from phages and ECFs whose conserved

$\sigma_2$  and  $\sigma_4$  domains are divergent.  $\sigma$  factors of phage origin have been described in literature; nevertheless, they are usually divergent from canonical  $\sigma^{70}$ s (Nechaev and Severinov, 2003) since they incorporate alternative domains replacing  $\sigma^{70}$  core domains in some cases. For instance, in *Bacillus* phage vB\_BceM-HSE3, the ECF Gp17 contains a double zinc ribbon domain (Pfam: PF12773) in the position where the  $\sigma_2$  domain usually is, while a generic  $\sigma_4$  Pfam domain is not found (Peng and Yuan, 2018). Similarly,  $\sigma$  factors Gp01 and Gp103 contain only  $\sigma_2$  domain or no Pfam domain, respectively (Peng and Yuan, 2018). Another reason for the lack of phage proteins in the present work is that viral genomes are usually not annotated in NCBI (Brister *et al.*, 2015) and did not enter the ECF search in most of the cases. Other types of ECF-like  $\sigma$  factors not included in the current version are ECFs whose  $\sigma_4$  (e.g. SigI from *Bacillus subtilis*) or  $\sigma_2$  domain (such as EcfP from *V. parahaemolyticus* or ComX from *Streptococcus pneumoniae*) do not hit their Pfam models. A special example of this are  $\sigma^1$ -like ECFs, which contain a  $\sigma_{1-C}$  domain instead of a canonical  $\sigma_4$  domain (Ortiz de Ora *et al.*, 2018). These ECFs are involved in the synthesis of cellulosome components in cellulolytic clostridia (Ortiz de Ora *et al.*, 2018). Attempts to classify these proteins against the current ECF classification were unsuccessful. The group with the highest probability of containing  $\sigma^1$ -like ECFs is ECF201 (probability = 1.12e-19), the outermost group of the ECF classification, indicating that  $\sigma^1$ -like ECFs are distant from canonical ECFs and might have evolved in parallel to them from group 3  $\sigma^{70}$ s.

Staroń and colleagues addressed the issue of the limited diversity of the extracted ECFs by defining a new HMM with the so-called “singletons”, this is, ECFs that are not classified, and hence are divergent (Staroń *et al.*, 2009). Inspired by this strategy, I used the 293 ECFs defined by previous classification efforts that are not retrieved in this expansion (hereafter called “library of excluded ECFs”) to build an HMM (“singleton” HMM) that described these divergent ECFs. As a performance control, I checked whether the library of excluded ECFs produced matches against the singleton HMM. Surprisingly, only 103 proteins (~35%) were able to hit this singleton HMM. This is likely the consequence of the large diversity present within the library of excluded ECFs, which cannot be explained by a single model. One explanation is that, when looking at a heterogeneous set of proteins, the amino acid changes present in different variants are covered up by the most common consensus residue. Therefore, the idea of recovering more ECFs using singleton models built from sets of divergent proteins was abandoned.

An alternative strategy to partially recover more divergent ECF variants is exemplified by the retrieval of more ECFs associated to STKs (Section 5). I explored two options for this. The first was based on finding proteins that hit an HMM built from conserved sets of proteins, this is, HMMs built from ECF groups. This resulted in the expansion of five out of the seven groups associated to STKs (Section 5.4). The remaining two groups were not expanded since some of their proteins lacked STKs at <5Kbp from their coding sequence. An alternative strategy is based on using the HMM built from proteins with sequence similarity to a single target of interest that are not initially included in the ECF

library. I used this strategy to find members of ECF43 associated to STKs using EcfP from *V. parahaemolyticus* as an input, since this protein was not captured in the initial ECF expansion due to its divergent  $\sigma_2$  domain (Section 5.1). This allowed for the expansion of ECF43 from 69 to 931 unique protein sequences. However, the first strategy, based on using an HMM built from the current members of ECF43, outperforms the usage of EcfP as a bait since it was able to extract 995 unique protein sequences from group ECF43. This strategy should be the preferred one for future expansions of the ECF database.

A clear limitation of these strategies is that they are based on already known ECF variants. This could be partially solved by iterative searches of new ECFs, where the results of one search are further used to perform the next, until no more proteins are found. This strategy is already implemented in popular protein search engines, such as PSI-BLAST (Altschul *et al.*, 2009) and HHblits (Remmert *et al.*, 2012). Future expansions of the ECF database could try to use iterative searches to increase the diversity of the proteins retrieved.

#### 7.8.2. ECF classification

The ECF classification presented in this work has two layers, which are useful when choosing the degree of protein sequence similarity needed for downstream analyses. The first clustering level is based on ECF subgroups, which are conserved clusters of proteins with a maximum k-tuple distance  $<0.6$  (Section 8.3). ECF subgroups are further clustered into ECF groups, in the same sequence diversity level as ECF groups from the original classification (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015a). ECF groups are more diverse at a protein sequence level, but they still contain a conserved genetic neighborhood with the same type of regulator (Section 3.2). ECF subgroups are useful when studying the closest homologs to the ECF of interest, whereas ECF groups are useful for the application of co-variation-based tools for the prediction of interactions between ECF and their regulator.

The two-layered structure of the ECF classification presented in this work was provided by a two-step clustering process. Even though there is a large range of clustering methods that can be used for protein classification (reviewed in (Holder and Lewis, 2003)), they are not usually able to deal with the large number of protein sequences ( $>170,000$ ) present in this work in a feasible time frame and with enough accuracy. To solve this issue, I used MMSeqs2 as a pre-clustering step. MMSeqs2 is specifically designed for working with large protein datasets (Steinegger and Söding, 2017). In exchange of its speed, MMSeqs2 outputs some heterogeneous clusters. For this reason, I corrected MMSeqs2 results using bisecting K-means, which gives as final output ECF subgroups. One of the issues of this approach is that similar (but not identical) ECF proteins may fall into different subgroups. The scatter observed for ECF subgroups is solved in ECF groups, since subgroups composed of similar protein sequences cluster together in the same clade of the phylogenetic tree. This problem is reflected in the lower accuracy of HMMs built from ECF subgroups respect to ECF

groups. ECF subgroups are able to correctly classify ECFs in ~94% of the cases, whereas this value is ~99% for ECF groups.

One of the problems of hierarchical classification is that the resulting phylogenetic tree needs to be partitioned into groups, but the optimal number and size of these groups is usually unknown and user-defined. On one hand, the number of groups should be large enough to not merge proteins with different characteristics into the same group and to be meaningful, but at the same time a large number of tiny groups that share the same type of proteins is not ideal. There are different types of methods to determine the most appropriate number of clusters in a classification (reviewed in (Chiang and Mirkin, 2010)). The most common methods focus on the proximity of the elements included in a cluster (compactness) respect to their distance from elements form other clusters (separation) (Halkidi, Batistakis and Vazirgiannis, 2001; Chiang and Mirkin, 2010). This is the case of the Silhouette coefficient (Rousseeuw, 1987). In the present work, the selection of the number of ECF groups is solved using the conserved genetic neighborhood of the subgroups, since the elements that they contain give indications of the function of the ECFs. In this way, the ECF classification is composed of the appropriate number of ECF groups, coherent with phylogenetics and genetic neighborhood conservation as a proxy of protein function.

As a result of the clustering strategy, the number of unclassified ECFs per organism is larger in bacterial phyla underrepresented in biological databases (Fig. 3.7). Furthermore, they tend to occupy peripheral subgroups when present in taxonomically diverse groups. A possible reason is that proteins from organisms underrepresented in databases are likely too diverse and scarce to be clustered with the currently available dataset. The analysis of 79 Planctomycetes revealed 30 extra ECF phylogenetic groups specific of this phyla (Wiegand *et al.*, 2019), indicating that focusing on specific underrepresented taxa could allow the further expansion of the ECF classification. Future work would try to determine whether the 30 extra ECF groups found by Wiegand and colleges agree with any of the planctomycetal ECF groups newly described in this work.

Phylogenetic studies that focus on proteins with multiple domains can be challenging since the full-length proteins may not align well when the domain content is variable. A major problem is the distinction between homologous proteins with similarity to the full-length target protein, and other proteins that may share some promiscuous domains with the target protein i.e., domains that fulfill general functions, such as conferring affinity to another protein, and are present in several families of proteins. This problem is partially solved through scoring metrics that allow for the identification of proteins with the same domain content as the target protein (Song, Sedgewick and Durand, 2007). Moreover, standard phylogenetics do not explain the domain shuffling events that are part of the evolution of some multidomain protein families (Stolzer *et al.*, 2015). ECF  $\sigma$  factors contain two conserved domains,  $\sigma_2$  and  $\sigma_4$ . As a result of the extraction pipeline (Section 3.1), ECFs classified in this work contain the exact same number and type of domains, and some of the issues of the classification of multidomain proteins are relieved. However, throughout the classification process I

did not consider the possibility of an alternative evolution of the ECF  $\sigma$  factor subfamily, where the two domains are shuffled independently. To account for this,  $\sigma_2$  and  $\sigma_4$  domains should be classified independently and the resulting two classifications should be compared. As a result of this process, domain swapping in ECF groups could be revealed and it would be possible to see ECF groups with a similar  $\sigma_2$  domain and a different  $\sigma_4$  domain and *vice versa*. These types of analyses were out of the scope of this work, but it could be interesting to perform them in the future.

### 7.8.3. Clustering validation

Clustering is an unsupervised machine learning technique that needs to be validated before considering its results as explanatory of the underlying data. Validation may be done using external or internal criteria (Halkidi, Batistakis and Vazirgiannis, 2001). Internal criteria evaluate that the structure of the classification fits the input data (Halkidi, Batistakis and Vazirgiannis, 2001). Methods of internal validation include measurements of compactness and separation, as already discussed for the Silhouette coefficient (Halkidi, Batistakis and Vazirgiannis, 2001). External criteria compare the clustering with other classifications considered to be the closest to reality (Halkidi, Batistakis and Vazirgiannis, 2001). I validated the ECF groups using the bootstrap values of group-defining branches and I compared the new ECF groups against the original classification as internal and external criteria, respectively (Section 3).

Bootstrapping is a common method for defining the confidence of a certain clade of a phylogenetic tree (Felsenstein, 1985). Standard nonparametric bootstrapping is based on the construction of phylogenetic trees from random samples of columns of the original alignment, with replacement (Felsenstein, 1985). The bootstrap values assigned to a certain branch represent the percentage of the bootstrap trees that contain the same clade. In this way, large bootstrap values of group-defining branches indicate that these clades appear in a broad range of the bootstraps. It may seem that bootstrap is a good manner of testing repeatability and accuracy of certain branches of phylogenetic trees. However, work by Hillis and Bull showed that under the most common conditions of phylogenetic analyses, bootstrap values are a highly imprecise way of measuring repeatability and they are conservative estimates of the probability of correctly inferring a certain clade (Hillis and Bull, 1993). Despite these drawbacks, bootstrap is the most common internal validation method for phylogenetic trees and several parametric implementations have managed to make it fast for large maximum-likelihood phylogenetic trees and, in some cases, to correct the conservative behavior of nonparametric bootstrapping (Hoang *et al.*, 2018). Parametric bootstrapping is based on resampling the estimated log-likelihoods calculated for each site of the original alignment, instead of resampling directly the alignment (Kishino, Miyata and Hasegawa, 1990). Alternative methods of internal validation involve the comparison of the total branch length with trees derived from random data (Lapointe, 1998); however, this method does not allow for the assessment of the partitions of the tree. Alternatively, one could use indexes, such as Silhouette coefficient, to assess the compactness and

separation of clades. These metrics would be high in clades where leaves have a large average distance to the rest of the tree, but are connected by short distances among themselves. Groups such as ECF42 would perform well according to the Silhouette coefficient, but other groups that have a large average internal divergence but share the same regulatory mechanism, such as the FecI-like group ECF243, would not perform well. Given the lack of a clear way of evaluating internally ECF group definition, nonparametric bootstrap seems to be the most widely accepted way of assessing the stability of the branches that root ECF groups. Therefore, I applied nonparametric bootstrapping to the ECF tree (Fig. 3.4E), and I assumed that large bootstrap values of group-rooting branches indicate a stable group definition.

As an external validation criterion, the overlap between estimated original groups and new groups was used. This comparison could have been done in a quantitative manner, using one of the numerous scores available to measure the similarity between two clustering methods, such as the Rand statistic or the Jaccard coefficient. The difficulty of these quantitative measures is that most of the classified ECFs are not part of the original classification, and estimates of their original ECF group are required. This makes difficult to decide whether a new ECF group is similar to an original group. Moreover, unclassified proteins should be considered as the sole members of their own group, which biases the results of these coefficients since most of the groups correspond to ungrouped ECFs. Nevertheless, taking the external validation in a quantitative manner would permit to evaluate different tree architectures according to the agreement with the original classification. There are also issues that go together with this strategy. For instance, the new classification would be forced to inherit problems of the original classification, such as the definition of heterogeneous groups such as ECF01 and ECF20. To avoid biasing the new classification to the original ECF groups, I compared the new groups with predictions of the original ones once new ECF groups had been defined according to the genetic neighborhood composition.

#### **7.8.4. Challenges of the application of DCA**

Tools based on protein co-variation, such as DCA, are a powerful way to predict the most important pairs of residues that connect two interacting families of proteins, or the ternary structure of the members of a single family of proteins. These methods require enough residue variation within each family to derive the positions that co-variate, but at the same time they require that the interaction between any pair of proteins is carried out using the same residues (Martin Weigt *et al.*, 2009). The great enrichment of groups in phylogenetically diverse proteins allowed for the application of DCA in individual groups such as ECF41, ECF42 and ECF56. This led to the discovery of functional differences between the C-terminal extensions from ECF41, ECF42 (Wu *et al.*, 2019) and ECF56 (Section 6), confirming previous reports (Wecke *et al.*, 2012; Goutam, Gupta and Gopal, 2017; Liu, Pinto and Mascher, 2018). However, it was not possible to extract valuable results from DCA in single ECF groups when analyzing the interaction between ASDIs and ECFs. This is probably due to

the lack of enough proteins in ASDI groups, since the largest group, AS12, contains only 691 pairs of ECFs/ASDIs. In order to predict contacts that agree with the available ECF/ASDI crystal structures, DCA had to be applied to all pairs of ECFs/ASDIs that share genomic proximity (Section 4). This strategy defined contacts that are likely important for the overall ASDI family, since they are conserved in most of the ECF/ASDI co-crystal structures. However, ASDI is a very diverse family of proteins, with an average identity of 10.55% using the alignment of subgroup consensus sequences. Therefore, it is likely that different members of different ASDI groups contact their ECFs with group-specific contacts, on top of the general contacts predicted in this work. Prediction of these group-specific contacts does not seem possible with the current data. Nevertheless, future updates of the ECF classification could result in an increase of the number of ECF/ASDI pairs associated to each group, which could allow for the prediction of the contacts that are specific of a single ECF/ASDI groups, similarly as in the case of C-terminal extensions.

Another study that took advantage of the enrichment in ECF  $\sigma$  factors addressed ECF phosphorylation as modulator of ECF binding affinity to RNAP core enzyme in members of ECF43 (Section 5). Attempts to apply DCA to predict important residues for the interaction between STKs and ECFs in members of ECF43 did not give any feasible prediction. The number of ECF43/STK pairs, 835 non-redundant entries, could be enough to provide good results when the contact is stable (non-transient) and conserved across the families. Aside from the low number of non-redundant protein sequences, another reason why DCA failed could be related to the large conformational changes of the STK “activation loop” during STKs transition between inactive and active states (Huse and Kuriyan, 2002). In inactive STKs, the activation loop blocks the binding of substrate or ATP, whereas in active STKs the phosphorylated activation loop moves away from the catalytic center, allowing for substrate binding (Huse and Kuriyan, 2002). Since several STK residues contact the activation loop in different stages of STK activation, it is possible that these contacts have a larger contribution to the amino acidic composition of this loop than the binding to the substrate peptide. This would prevent DCA results from being accurate, since the observed covariation is due to intra-molecular contacts. This dynamic binding of STKs to ECFs contrasts with the seemingly more stable binding of C-terminal extensions and anti- $\sigma$  factors to ECFs, which can be predicted by DCA. One of the main differences is that STKs are catalyzing the covalent binding of a chemical group to the ECF, whereas the function of C-terminal extensions and ASDIs seems to be related to steric effects. Even though some reports have hypothesized that ECF41 C-terminal extensions could have catalytic activity (Goutam, Gupta and Gopal, 2017), this did not hamper the ability of DCA to make valuable predictions. In this line, it has been suggested that the SnoaL-like C-terminal extension of members of ECF41 could behave as a sensor, rather than as an enzyme, to modify the activity of the ECF according to the presence of its ligand (Goutam, Gupta and Gopal, 2017).

### 7.9. Final remarks

The first aim of this thesis was to expand the ECF classification in number of proteins and diversity. I used all the annotated genomes in NCBI to fulfil this aim, resulting in a 50-fold increase in the number of unique ECF proteins and 22 completely new ECF groups, where ECFs could not be assigned to any original group. This shows that not only the size of the ECF library increased, but also the diversity of the proteins included within. Results of the analysis of the new ECF groups confirmed the findings of the original classification and added new putative modes of regulation to the set of possibilities in ECF  $\sigma$  factors, including short C-terminal extensions, pairs of cysteines with the capacity of creating disulfide bridges under oxidative conditions, unstable ECFs stabilized by other proteins and ECFs regulated by proteolysis. These results suggest that ECFs have a multi-layered regulation with a higher signal integration capacity than previously thought.

The second aim of this thesis was to analyze the mechanisms that govern the interaction between ECFs and anti- $\sigma$  factors. Focusing on class I anti- $\sigma$  factors as the most common anti- $\sigma$  factor type, I predicted two main binding interfaces that are present in all the ECF/ASDI complexes co-crystallized to date. These two interfaces could regulate the specific binding of ASDIs to their cognate ECF, although experimental confirmation of this part is missing. The most important binding interface involved ASDI helix 4 and ECF  $\sigma_2$  domain. Residues contained in these areas are diverse within ECF/ASDI phylogenetic groups, suggesting their involvement in the fine specificity of ASDIs for their cognate ECF. The second binding interface involved a single residue from ASDI helix 1 and at least two residues from ECF  $\sigma_4$  domain. These residues seem to be conserved within groups and could define specificity for members of the same ECF/ASDI group.

The last aim of this thesis was to establish the functional role of STKs and C-terminal extensions in ECF activity in different ECF groups. This study revealed that the functional role of C-terminal extensions was different in groups ECF41 and ECF42. Furthermore, some predicted contacts between the SnoaL-like C-terminal extension of ECF56 and its ECF core domains suggested that members of ECF56 might have a different regulation than their close relatives from group ECF41. The role of STKs was studied in ECF43, finding that ECF phosphorylation was required for the binding to the RNAP core complex. Other putative phosphorylation sites were predicted for other six groups with a conserved STK in their genetic neighborhood.

In conclusion, the comprehensive classification of ECF  $\sigma$  factors into phylogenetic groups provides the scientific community with a comprehensive guide on their regulation, target promoter and function. Scientists studying new types of ECFs could benefit from this work as long as their ECF classifies against an ECF group.

## 8. Material and methods

### 8.1. General bioinformatic tools

Throughout this work, shell wrapper scripts containing calls to custom Python, R and MATLAB scripts were used. These wrapper scripts often contained calls to published programs, referenced throughout this work.

Generally, multiple-sequence alignments (MSAs) were generated by Clustal Omega 1.2.3. with options `--iter=2` and `--max-guidetree-iterations=1` (Sievers *et al.*, 2011). In some cases, these alignments are manually curated. However, UPP (N. P. D. Nguyen *et al.*, 2015) with default options was used when indicated. Hidden Markov Models (HMMs) were built using *hmmbuild* function and used for scanning libraries using *hmmsearch* function, both from HMMER suite 3.1b2 (Finn, Clements and Eddy, 2011) and both with default parameters. The envelope region was used when the area of a protein that hit a certain HMM needed to be extracted. Protein structures were visualized using UCSF Chimera version 1.10.2 (Pettersen *et al.*, 2004). Sequence alignments were visualized using CLC Main Workbench 8.

Phylogenetic trees were built mostly with IQ-Tree with automatic model selection and default parameters if nothing else is stated (L.-T. Nguyen *et al.*, 2015), although in some cases RAxML version 8.2.12 (Stamatakis, 2014) was preferred. RAxML ran with `-f a`, 100 rapid bootstraps, automatic protein substitution model selection and two threads (Stamatakis, 2014). Phylogenetic trees were visualized in iTOL (Letunic and Bork, 2016).

Protein structure from WP\_006346870.1 (locus STSU\_11560) from *S. tsukubaensis* was modelled using online I-TASSER with default options (Yang *et al.*, 2014). For this, the SnoaL-like C-terminal extension (amino acid 240 to 369) and the ECF core region (amino acids 1 to 239) were run independently. EcfP structure was modelled with Swiss-model (Waterhouse *et al.*, 2018) using SigH as template (PDB: 5ZX2 (L. Li *et al.*, 2019)).

Motif searches were done with MEME suite (Bailey *et al.*, 2009). Specifically, TomTom with default parameters was used for searching in all prokaryotic DNA motif databases for matches to a given DNA motif (Gupta *et al.*, 2007).

### 8.2. Extraction of new ECFs from NCBI

The amino acid sequence between the start of  $\sigma_2$  and the end of  $\sigma_4$  domains (core ECF region) was extracted from the MSA of the 3,755 ECFs from the original library (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015b) (position 912 and 2059 in the MSA) and was used to build the general HMM for ECF sequence retrieval. Then, ECFs from the original library and proteins with a  $\sigma_3$  domain (Pfam: PF04539) retrieved from Pfam (release 31.0 (El-Gebali *et al.*, 2019)) were scored against this model. The resulting bit scores were used to construct a Receiver Operator Characteristic (ROC) curve (using the function “roc\_curve” from the “scikit-learn” Python package (Pedregosa *et al.*, 2012)) in order to select the threshold score able to capture the greatest number of ECFs (highest

sensitivity), while minimizing the number of  $\sigma_3$ -containing proteins (highest specificity). This resulted in an optimum bit score threshold of 60.8, a sensitivity of 0.92, a specificity of 0.98 and an area under the curve (AUC) of 0.98, accounting for a robust performance of the classifier. Then, proteins with a score higher than the optimum threshold were selected as putative ECF  $\sigma$  factors for the next steps. For this search, I considered all protein sequences from genomes with annotation (.gff), protein (.faa) and genome (.fna) file available in NCBI (as of February 2017), including RefSeq and GenBank entries, accounting for a total of 156,241 genomes and 554,108,437 proteins. Subsequent quality controls were applied to ensure that all the putative ECFs contain  $\sigma_2$  (Pfam: PF04542) and  $\sigma_4$  (Pfam: PF04545 and PF08281) domains and lack the  $\sigma_3$  domain using Pfam HMM profiles (hmmsearch with default settings). Only the non-redundant proteins without ambiguous amino acids were considered further. Sequence redundancy was assessed with Cd-hit (Li and Godzik, 2006) at 100% sequence identity. In summary, from the 554,108,437 annotated proteins, 714,848 had a positive score against the ECF model. Of those, 177,910 had  $\sigma_2$  and  $\sigma_4$  domains, lacked  $\sigma_3$ , lacked non-amino acidic symbols and were non-redundant, constituting the extended ECF library used in the next steps.

Within this dataset I found that 1,217 proteins generated low-scoring hits against the model of the  $\sigma_3$  domain (Pfam: PF04539) when hmmscan was applied for  $\sigma_3$ 's HMM alone. This is likely due to the size of the ECF database, where E-values are less significant than for smaller datasets (see HMMER documentation (Finn, Clements and Eddy, 2011)). Given that the Pfam HMM of  $\sigma_3$  wrongly hits the  $\sigma_2$  or  $\sigma_4$  domain in some cases, only the proteins where the highest scoring  $\sigma_2$ ,  $\sigma_3$  and  $\sigma_4$  domains are not overlapping are considered as  $\sigma_3$ -containing proteins. After their identification, the 1,217  $\sigma_3$ -containing proteins were used as outliers for the clustering, since they are the closest  $\sigma$  factors to ECFs.

The average number of ECFs per genome was computed considering only the 12,539 ECFs from the 1,234 complete genomes tagged as 'representative' or 'reference' in NCBI, giving priority to RefSeq genomes over GenBank if both exist for the same organism, unless stated otherwise.

### 8.3. ECF clustering

Non-redundant ECF sequences were stripped to  $\sigma_2$  and  $\sigma_4$  regions using hmmscan (HMMER suite 3.1b2 (Finn, Clements and Eddy, 2011)) and Pfam models for  $\sigma_2$  and  $\sigma_4$ . I selected the 'envelope' region of the hit with the lowest E-value. For the first step of clustering, I applied MMseqs2 (Steinegger and Söding, 2017) with default parameters. However, the phylogenetic distance between pairs of sequences within clusters, as calculated from the k-tuple distance (Wilbur and Lipman, 1983) obtained from Clustal Omega 1.2.3. with options '--full, --full-iter and --distmat-out' (Sievers *et al.*, 2011), was large in some cases, indicating imperfect clustering at this stage. To solve this issue, MMseqs2 clusters were split using a bisecting k-Means algorithm until the maximum pairwise distance between sequences of the clusters was  $\leq 0.6$ . This threshold was the largest maximum k-tuple distance whose associated clusters contained proteins similar enough to produce homogeneous MSAs.

The resulting 2,380 clusters with more than 10 sequences are defined as ECF subgroups. A total of 137,452 ECFs (77.26%) were classified into subgroups. Then, the consensus sequences of subgroups (computed with Biopython (Cock *et al.*, 2009)) were used for the construction of a phylogenetic tree. As outliers of the tree I included the consensus of a MSA calculated from all proteins included in Pfam (release 31.0 (El-Gebali *et al.*, 2019)) that contain all of the three domains  $\sigma_2$ ,  $\sigma_3$  and  $\sigma_4$ . I also included the 1,217 closest proteins to ECF  $\sigma$  factors with  $\sigma_3$  domain. The subsequent phylogenetic tree was manually split into monophyletic clades using a divisive strategy, where two clades were kept together in the same ECF group unless the genetic context or the putative anti- $\sigma$  factor (when present) differed. ECF groups featuring only a single subgroup were only labelled if they have significant similarity to an original ECF group, while all other single subgroups were maintained as singleton subgroups (for nomenclature see next paragraph). This strategy resulted in 157 ECF groups that contained 135,259 ECF  $\sigma$  factors, corresponding to 76.03% of the new ECF library.

Names of original ECF groups are maintained for groups with the same characteristics. When several original groups are represented in an ECF group or the ECF group has no significant similarity to any original group, the name of this ECF group follows the pattern ECF2XX, standing for ECF classification 2.0, where XX is a running number assigned according to the position in the phylogenetic tree. For instance, ECF201 is closer to the base of the tree than ECF260. Subgroups are referred with the name of the ECF group they are part of, followed by 's', standing for subgroup, followed by a running number that increases for decreasing subgroup size. For instance, subgroups ECF02s1 (ECF02 subgroup 1) and ECF02s2 (ECF02 subgroup 2) are both part of group ECF02, and s1 contains more non-redundant proteins than s2. Subgroups that are not part of any ECF group are named 'ECFs' followed by a running number according to their position in the phylogenetic tree.

Groups and subgroups were evaluated according to the performance of their HMMs, built from the concatenated  $\sigma_2$  and  $\sigma_4$  domains of the constituting protein sequences. These HMMs were used to score proteins from all groups or subgroups. Bit scores below the reporting threshold were considered equal to 0. The average score of members of each group against each HMM was normalized by dividing by the score of the group against its own HMM. The average normalized scores are plotted in a heatmap (Fig. 3.4G). I validated the subgroups by generating 100 randomly permuted sets of proteins with the same size distribution as the subgroups (Fig. 3.4F). The mean average k-tuple distance in the permuted data was  $0.79 \pm 0.01$ , whereas this value was  $0.29 \pm 0.11$  for ECF subgroups. The difference between the distributions of average pairwise k-tuple distances of ECF subgroups and permuted clusters was statistically significant (two-tailed Student's t-test p-value  $< 1e-16$ ). Furthermore, I evaluated the support of the branches of the phylogenetic tree by running 100 bootstrap replicates, as implemented in IQ-TREE (L.-T. Nguyen *et al.*, 2015). As a further plausibility test, I verified the agreement between original and new classification (Fig. 3.6).

#### 8.4. ECF group analysis

For the analysis of ECF group characteristics I only included proteins from ‘representative’ and ‘reference’ genomes as defined by NCBI (<https://www.ncbi.nlm.nih.gov>), thereby reducing the bias towards frequently sequenced organisms. Only RefSeq assemblies were considered when both RefSeq and GenBank assemblies are available. To define coherent ECF groups and to elucidate the putative function of members of each group, I analyzed the protein domain composition of the proteins encoded at a distance of  $\pm 10$  coding sequences from the ECFs coding sequence. First, I queried these proteins against the HMMs of Pfam 31.0 (El-Gebali *et al.*, 2019). For every protein, I only considered the non-overlapping Pfam domains with the lowest E-value, leading to a set of specific domains (which I defined as the ‘domain architecture’) for each protein in the genomic neighborhood of the ECF. For each ECF subgroup, I analyzed the conservation of the domain architecture in specific positions up- and downstream of the ECF. A domain architecture is defined as conserved if it appears in more than 75% of the genomic contexts of an ECF subgroup. To avoid biases due to low number of ECFs, I only analyzed the genetic context of subgroups with more than 10 ECFs.

For anti- $\sigma$  factor identification I used 1) Pfam domains of known anti- $\sigma$  factors, 2) detectable sequence similarity to anti- $\sigma$  factors of the founding classification (Staroń *et al.*, 2009) and 3) presence of transmembrane helices, as described in the following. Most of the anti- $\sigma$  factors cannot be predicted due to the lack of Pfam domains that describe them. Therefore, I used the anti- $\sigma$  factors retrieved by (Staroń *et al.*, 2009) as the database to query candidate anti- $\sigma$  factors using BLAST (Altschul *et al.*, 1990) with an E-value  $< 0.01$ . Moreover, since anti- $\sigma$  factors are usually transmembrane proteins, I predicted the presence of transmembrane helices using the consensus prediction of TopCons (Tsirigos *et al.*, 2015). In cases where the presence of the transmembrane helix was not clear, I aligned the sequences of the putative anti- $\sigma$  factors to determine the presence of conserved hydrophobic regions likely to be transmembrane domains. Since anti- $\sigma$  factors are usually located in positions  $\pm 2$  from the ECF coding sequence, those were the main positions I focused the search on.

#### 8.5. Classification of new ECFs against ECF clusters

Parameters  $k_i$  and  $c_i$  from Eq. 3.1 were fitted to the bit scores obtained by members and non-members of the ECF cluster under evaluation using the function ‘curve\_fit’ (with Levenberg-Marquardt algorithm) from the Python package ‘scipy.optimize’ in SciPy library (Jones *et al.*, 2001). Only ECFs stripped to their  $\sigma_2$  and  $\sigma_4$  domain were used for this. For more details, see Section 3.5.

#### 8.6. Prediction of ECF target promoter motifs

Since ECFs often auto-regulate their own transcription, their putative target promoters can be predicted from the regions upstream of the operon in which the ECF is encoded. To this end I searched for conserved bipartite nucleotide motifs in the 200 base pairs (bp) upstream of the ECF

operon, by executing BioProspector (Liu, Brutlag and Liu, 2001) with the parameters ‘-W 7 -w 5 -G 18 -g 15 -n 150 -d 1’, as previously described (Rhodius *et al.*, 2013). Operons were defined as the set of coding sequences transcribed in the same direction as the ECF and with an intergenic distance shorter than 50bp. Only the highest scoring motif of each input sequences was further considered. The region containing the -35 and -10 motifs, in addition to  $\pm 10$ bp up and downstream is represented in a sequence logo produced by WebLogo 3 (Crooks *et al.*, 2004).

### 8.7. Class I anti- $\sigma$ factor extraction

Anti- $\sigma$  factors have little sequence similarity, hampering their identification. Some indications that help in their retrieval are 1) that their coding sequence is usually in close proximity to the one of the ECF, 2) that they are usually transmembrane, even though this has important exceptions (Campbell *et al.*, 2007; Staroń *et al.*, 2009), and 3) that they are conserved within ECF phylogenetic groups (Staroń *et al.*, 2009).

Putative ASDIs were extracted from the genetic neighborhood ( $\pm 10$  coding sequences) of a library of 46,293 ECF  $\sigma$  factors from “representative” and “reference” organisms, as defined by NCBI (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>), using only RefSeq entries when both RefSeq and GenBank records are available for the same genome. In this way I minimized the taxonomic bias present in the library. To identify ASDI domain-containing proteins, I first used two HMMs, one built from the zinc-binding and another from the non-zinc-binding anti- $\sigma$  factors from the work of Staroń and colleagues (Staroń *et al.*, 2009). I selected the optimal bit score threshold for the retrieval of new ASDIs for each HMM by optimizing a Receiver Operator Characteristic (ROC) curve using the function *roc\_curve* from *sklearn.metrics* (Pedregosa *et al.*, 2011). Proteins used for the construction of each model were used as positive controls, and the remaining non-ASDI anti- $\sigma$  factors from Staroń *et al.* (Staroń *et al.*, 2009) as negative controls. The resulting bit score thresholds, 0.4 for non-zinc binding and 14.2 for zinc-binding models, were applied for the extraction of ASDIs from the set of putative anti- $\sigma$  factors. This resulted in 7,490 ASDIs, which were subsequently used for the construction of an extended HMM of the ASDI family. The thresholding bit score that best separates real ASDIs from other proteins was optimized using a ROC curve as described above, resulting in a bit score threshold of 0.2. I used this extended HMM to look for further members of the ASDI family in the genetic neighborhood of ECFs ( $\pm 10$  coding sequences) from the ECF classification in Section 3. In order to lessen the bias towards frequently sequenced organisms, I only included proteins from representative or reference genomes as labelled by NCBI (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>), using only RefSeq entries when both RefSeq and GenBank records are available for the same genome. This yielded 11,939 putative ASDI-containing proteins. I further curated these data removing proteins with anti- $\sigma$  domains shorter than 50 amino acids, since these could be anti- $\sigma$  factors of class II (Sineva, Savkina and Ades, 2017). The area of the anti- $\sigma$  domain was defined as the envelope region of the highest scoring hit of the extended

HMM, discarding areas that are part of the transmembrane helices or extracellular. This resulted in 10,930 ASDIs, with an average length of  $101 \pm 33$  (standard deviation) amino acids.

### 8.8. ASDI classification

I clustered ASDIs according to amino acid sequence similarity. Given the large number of proteins the retrieval returned (over 10,000), I first grouped them into clusters or closely related sequences, the so-called subgroups. These were built with a divisive strategy, where proteins were subjected to a bisecting K-means clustering approach until the maximum k-tuple distance between any protein of the cluster is smaller than 0.6, as measured by Clustal Omega 1.2.3. with `--distmat-out --full` and `--full-iter` flags (Wilbur and Lipman, 1983; Sievers *et al.*, 2011). Bisecting K-means was implemented using *KMeans* function from *sklearn.cluster* module (Pedregosa *et al.*, 2011). The 3,790 proteins that did not enter any subgroup were left ungrouped. Thanks to this grouping it was easier to see subgroups that may contain outliers that passed the HMM threshold, but do not likely display anti- $\sigma$  factor activity. In order to distinguish and discard these outliers from our clustering, I assessed the presence of Pfam domains (Pfam 31.0 (El-Gebali *et al.*, 2019)) in the anti- $\sigma$  factors from each subgroup. I discarded 132 subgroups (606 proteins) where their Pfam domains indicated an unlikely anti- $\sigma$  factor function (data not shown). The resulting 1,475 subgroups contained 6,534 proteins (~60% of the starting ASDIs), with a median group size of 3 proteins and a standard deviation of 6.17 proteins. Given the low size of proteins in each subgroup, I further clustered the manually curated alignment of the consensus sequences of each subgroup, into a maximum-likelihood phylogenetic tree using IQ-TREE version 1.5.5 (L.-T. Nguyen *et al.*, 2015) with 1000 ultrafast bootstraps. As an outgroup of this tree, I included the class II anti- $\sigma$  factor CnrY, from *Cupriavidus metallidurans*. The resulting tree was split into monophyletic ASDI groups according to the ECF group of their cognate partner. With this strategy, I defined 23 ASDI groups, of which 12 contain more than 100 proteins.

The presence of a zinc-binding domain was assumed in ASDIs with a H<sub>x</sub><sub>3</sub>C<sub>x</sub><sub>2</sub>C sequence signature that expands over helix 2 and helix 3. Presence of transmembrane helices was assessed using the consensus prediction from online TopCons (Tsirigos *et al.*, 2015). The presence of other Pfam domains in full-length class I anti- $\sigma$  factors was evaluated using hmmscan function from HMMER suite 3.1b2 (Finn, Clements and Eddy, 2011) with the library of HMMs from Pfam 31.0 (El-Gebali *et al.*, 2019). Pfam domains present in a certain position of the MSA of the full-length anti- $\sigma$  factors in more than 50% of the members of a subgroup were plotted in the ASDI tree.

### 8.9. Evaluation of ECF-ASDI co-evolution

In order to evaluate the co-evolution of ECFs and ASDIs, I calculated the Pearson correlation coefficient (PCC) of the distances between cognate pairs of proteins, as introduced by Goh *et al.* (Goh *et al.*, 2000). The significance of this PCC was evaluated similarly as in (Dintner *et al.*, 2011). For this purpose, the PCCs between ASDIs, ECFs and of two extra families of proteins that do not co-evolve

and/or interact with ECFs or ASDIs were evaluated as negative controls. These negative controls were homologs of *E. coli*'s housekeeping  $\sigma$  factor  $\sigma^{70}$  (RefSeq: NP\_417539.1) and of *Bacillus subtilis*' anti- $\sigma$  factor RsbW (RefSeq: WP\_061902497), since proteins for these types have never been described to interact with ASDIs nor ECFs, respectively. I extracted proteins from these types using online HMMER (Finn, Clements and Eddy, 2011) with parameters `-E 1 --domE 1 --incE 0.01 --incdomE 0.03 --mx BLOSUM62 --pextend 0.4 --popen 0.02 --seqdb uniprotrefprot`, and mapped the hit IDs from UniProt to GenBank using the UniProt's ID conversion tool (Huang *et al.*, 2011). A total of 409 genomes contained the four protein families, this is ECFs, ASDIs, RsbW and RpoD. For each organism, I selected one of the ECF-anti- $\sigma$  factor pairs and one homolog of RsbW and RpoD. These proteins had a taxonomic diverse origin, with 39% of the proteins from Firmicutes, 28% from Actinobacteria, 11% from Cyanobacteria and the rest from other eight bacterial phyla. I calculated the pairwise distance for each protein family using Clustal Omega with `-full` and `--distmat-out` flags (Sievers *et al.*, 2011). The PCC was calculated from the flattened distance matrices using *pearsonr* function from Python's *scipy.stats* resource (Jones *et al.*, 2001).

#### 8.10. Specificity Determining Positions (SDPs)

SDPs were calculated with S3det (Rausell *et al.*, 2010) on the 12 ASDI groups with more than 100 proteins, and on their cognate ECFs. Aligned ASDIs (or ECFs) were extracted from the MSA used for DCA to preserve the same positional mapping. S3det was executed on every pair of ASDI (or ECF) groups, resulting in a set of ranked SDP predictions for every pair of groups. I scored the SDPs associated to every group as the sum of the inverse of their ranks across the different S3det runs with contribution of the group. The highest scoring SDP for every group was considered positive, resulting in five SDPs. For the extraction of the amino acid residue interactions between ECF and ASDI from co-crystal structures, I used Voronoi tessellation as implemented in Voronota version 1.19 (Olechnovič and Venclovas, 2014).

#### 8.11. Extraction of EcfP-like proteins

To identify proteins similar to EcfP with a STK encoded in the vicinity of their coding sequence, I built (HMMs) from the multiple-sequence alignment (constructed using Clustal Omega 1.2.3 (Sievers and Higgins, 2014)) of the ECFs and STKs extracted from a PSI-BLAST search (Altschul *et al.*, 2009) (E-value < 10) of EcfP and PknT. I used these HMMs to look for proteins with similarity to EcfP and PknT in the annotated genomes in NCBI (version from February 2017). Then, I filtered for pairs of ECF-STK whose coding sequences were separated by less than 5Kbp. The 224 unique matches where several ECFs or several STKs are found in the same neighborhood were discarded since the interaction pair is not clear. I also discarded 12 variants where the ECF and the STK coding sequences were fused. In order to reduce the number of proteins, I removed some pairs that, when combined (ECF concatenated with STK), yielded an identity greater than 98%. This means that if two

ECF/STK pairs share more than 98% of their sequence, I removed one pair. The search yielded 1617 ECF/STK pairs with a combined amino acid sequence identity <98%. From those, 14 putative STKs did not match the PFAM model for proteins kinases (PFAM: PF00069) and were discarded, ending with 1,603 pairs of ECF-STK.

I aligned the resulting ECFs with Clustal Omega 1.2.3 (Sievers and Higgins, 2014). A maximum likelihood phylogenetic tree was built from the manually curated alignment using IQ-Tree with automatic model selection and default parameters (L.-T. Nguyen *et al.*, 2015). I included as outliers of this phylogenetic tree SigM from *B. subtilis*, RpoE and FecI from *E. coli*, Ecf41 from *R. sphaeroides* and SigE from *M. tuberculosis*. The latter is regulated through the phosphorylation of its anti- $\sigma$ -factor, RseA, which is then cleaved by the protease ClpC1P2 (Barik *et al.*, 2010). All the remaining outliers are not known to be regulated by STKs. The domain architecture of STKs was computed using the models from PFAM database release 31.0 (El-Gebali *et al.*, 2019).

For the assignment of the retrieved ECFs to ECF groups, HMMs from the available ECF classifications (Staroń *et al.*, 2009; Jogler *et al.*, 2012; Huang *et al.*, 2015a) were used to score every extracted ECF against every ECF group. Proteins were assigned to the group for which the bit score was highest.

The presence or absence of an extended region between helices  $\sigma_{2.1}$  and  $\sigma_{2.2}$  was assessed from the alignment of the extracted ECFs, including outliers. I considered that a protein has an extended region when the length between  $\sigma_{2.1}$  and  $\sigma_{2.2}$  helices is larger than three times the standard deviation of the length in the outliers, that is, 11.12 amino acids.

#### **8.12. Extension of STK-associated groups**

I scanned all the genomes in NCBI (release from February 2017) with the HMMs of the seven new groups (ECF43, ECF59, ECF61, ECF62, ECF217, ECF267 and ECF283) that are co-conserved with protein kinases (Fig. 5.1), using the Pfam model for the protein kinase domain (Pfam: Pkinase) from Pfam release 31.0. Resulting hits were evaluated according to 1) their assignment to an STK-associated ECF group and 2) the presence of a protein kinase-containing protein encoded near the candidate's coding sequence. For classification of proteins to STK-associated ECF groups, the HMMs of these seven groups were used to score all the proteins from RefSeq and GenBank libraries in NCBI. The decision of whether a protein is a true member of a STK-associated ECF group inherits from the pipeline designed for protein classification (Sections 3.5 and 8.5), but differs from it in that full-length ECFs and HMMs are used, and only ECF groups associated to STKs are tested. Only proteins that score higher than noise or trusted thresholds of a group are further considered as candidate members. Then, the probability that these proteins belong to the ECF group is calculated. This is done applying the bit scores to the sigmoid formula (Equation 3.1), with the parameters obtained from the least squares fit of the sigmoid curve to the bit scores obtained by members (probability=1) and non-members (probability=0) using full-length ECFs, as explained in Section 8.5.

Proteins are assigned to the group that achieves the highest probability, as long as this probability is higher than the ROC-optimized threshold of 0.34% (see Section 3.5 for details). As an extra filter, only proteins that are encoded in less than 5Kbp from a coding sequence with a protein kinase domain (Pfam: Pkinase) are further considered (HMMER E-value < 10). As a result of this pipeline, 4,719 ECFs were extracted, of which 1,707 had unique protein sequence.

The analysis of the positions with conserved serine and/or threonine residues was performed based on an MSA of the ECFs extracted in the previous step, together with two random, control ECFs from each of the remaining 150 ECF groups in the ECF classification presented in Section 3. RpoE from *E. coli* (RefSeq: WP\_001295364.1) was chosen as reference to name the amino acid coordinates of the results. The MSA was performed using UPP with default parameters (N. P. D. Nguyen *et al.*, 2015). The resulting alignment was scanned for positions that typically contain negatively-charged residues (aspartate (D) or glutamate (E)) in the control sequences. The top 15 columns of the alignment according to D+E frequency (frequency of D+E from 98.29% to 36.3% in the controls) were chosen to calculate the amount of serine or threonine (S+T) in the STK-associated ECFs. Conservation of S+T residues in the STK-associated groups in positions that typically feature negative charges indicates that they might be phosphorylated by their associated STK, compensating for the lack of negative charge. As a positive control, the position equivalent to T63 in *V. parahaemolyticus*, turns positive in ECF43 in this analysis.

### 8.13. Evolution of ECF43

Members of ECF43 were extracted from the expansion of members STK-associated groups (Section 8.12). For that, the whole proteome of the ten representative and reference organisms that contain members of ECF43 in their genome was scanned searching for protein kinase domains (Pfam: Pkinase version 27 May 2019) using hmmscan (HMMER suite 3.1b2 (Finn, Clements and Eddy, 2011)) with default options. Only RefSeq entries were considered when both RefSeq and GenBank genomes were available. The 49 proteins with hits to the protein kinase domain with E-value <0.01 and domain c-E-value <0.01 are considered positive. Then, the envelope region of Pkinase HMM was used for the construction of an MSA. This alignment was used to build a maximum likelihood phylogenetic tree. The kinase domains of the STKs PknA and PknB, from *M. tuberculosis*, were used as outliers of the tree. The tree was built with RAxML version 8.2.12 with -f a option, 100 rapid bootstraps, automatic protein substitution model selection and two threads (Stamatakis, 2014). The proteins encoded at a distance of 10 CDSs up- and down-stream the STK coding sequence were analyzed according to their domain content, using Pfam domains from release 31.0 (El-Gebali *et al.*, 2019) that have a hit with an E-value <0.01 and domain c-value <0.01.

PknT-like proteins were extracted from representative and reference genomes from *Vibrio* genus using blastp function with an E-value <0.01 from a locally installed version of BLAST 2.7.1+ (Camacho *et al.*, 2009) and full-length PknT as query. The resulting sequences were aligned using

Clustal Omega 1.2.3 with options `--iter=2 --max-guidetree-iterations=1` (Sievers and Higgins, 2014). PknT protein sequence was included as a control. A maximum-likelihood phylogenetic tree of these proteins was built with RAxML version 8.2.12 with `-f a` option, 100 rapid bootstraps, automatic protein substitution model selection and two threads (Stamatakis, 2014). The proteins encoded at a distance of 10 CDSs up- and down-stream the PknT-like proteins were analyzed according to their domain content, using Pfam domains from release 31.0 (El-Gebali *et al.*, 2019) that have a hit with an E-value  $<0.01$  and domain c-value  $<0.01$ .

#### **8.14. Direct coupling analysis (DCA)**

The DCA pipeline consisted of a wrapper shell script that call several others scripts, namely Python scripts for the pre-processing of the sequences of the families of proteins under study, the MATLAB script for Gaussian DCA (Baldassi *et al.*, 2014) and the different Python scripts that allow for plotting DCA results into user-defined crystal structures. The pre-processing included the selection of the area or domain of the protein/s known to be interacting. This is important in the case of class I anti- $\sigma$  factors, where only the ASDI domain was analyzed by DCA. Then, an MSA of the families of proteins under study was produced. Clustal Omega 1.2.3 (Sievers and Higgins, 2014) was used for conserved families of proteins, such, members of ECF56 and ECF41, whereas UPP 4.3.8 (N. P. D. Nguyen *et al.*, 2015) was used for families of proteins with a lower degree of conservation, such as ASDIs. Alignment columns with  $>70\%$  gaps were removed to reduce the amount of data DCA had to process. Pairs of interacting proteins were fused into the same FASTA entry when interaction between two different families of proteins is under analysis, such as ECF-ASDI interaction. The resulting alignment was used as an input of Gaussian DCA (Baldassi *et al.*, 2014). The results were mapped to the original full-length alignment and a Chimera (Pettersen *et al.*, 2004) script was created to plot the top DCA predictions in the crystal structures of members of the family under analysis, when available. A table indicating the top predictions and its corresponding amino acids in target sequences was produced. DCA direct information was plotted as a contact map using custom Python scripts.

#### **8.15. Promoter search for STSU\_11560**

I searched for the predicted target promoter motif of ECF56s3 in the genome from *S. tsukubaensis* using online Virtual Footprint version 3.0 (Münch *et al.*, 2005). The annotated genome of *S. tsukubaensis* was obtained from Rute Oliveira and Dr. Marta V. Mendes (Oliveira *et al.*, unpublished). This genome is equivalent to GCF\_000297155.2 genome record from NCBI. The predicted target promoter motif of ECF56 was split into the two regions corresponding to -35 (consensus GATGAG) and -10 (consensus CGTCA) elements in order to submit a bipartite pattern to Virtual Footprint. Virtual Footprint was executed with 0.8 sensitivity threshold, 0.9 core sensitivity, size 5 and spacer between 14 and 20 base pairs. This resulted in 2,261 hits. Of those, 183 were at a distance  $\leq 300$ bp

from an ORF and in the same transcription direction. Proteins encoded downstream, in the same direction and in the same operon as the 183 coding sequences identified in the previous step – defined as ORFs separated by  $\leq 50$ bp – were extracted and scanned for Pfam domains using the models from Pfam release 31.0 (El-Gebali *et al.*, 2019) and hmmscan function from HMMER suite 3.1b2 (Finn, Clements and Eddy, 2011). Only hits with an E-value and c-Value  $< 0.01$  were further considered. The association between Pfam domains and GO terms was extracted from InterPro2GO mapping as per January 2019 (Mitchell *et al.*, 2015).

*De novo* prediction of target promoter motifs was done using the six regions where ChIP-seq reads align with a log fold-change  $> 0.6$  in the putative anti- $\sigma$  factor (STSU\_11555) deletion mutant respect to the ECF (STSU\_11560) deletion mutant (Oliveira *et al.* 2020, in preparation). The region of STSU\_06338 that was used for promoter prediction needed to be extended from the above-mentioned area in order to have enough base pairs to accommodate an ECF  $\sigma$  factor binding site. Bipartite overrepresented sequences were predicted with BioProspector (Liu, Brutlag and Liu, 2001) with options -W 7 -w 4 -G 20 -g 15 -n 100 -d 1. BioProspector predictions were plotted in *S. tsukubaensis* genome to compare with the position of the ChIP-seq peaks.

#### **8.16. Search for a common regulon for members of ECF56s3**

Only the 79 genomes that contain a member of ECF56s3 tagged as “representatives” or “reference” by NCBI were considered (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>). Only RefSeq genomes were considered when both a RefSeq and a GenBank assemblies of the same genome exist. Genomes from these organisms were scanned, searching for the predicted target promoter motif of ECF56s3 using Virtual Footprint with default options (Münch *et al.*, 2005). Only predicted promoters located at  $\leq 300$ bp of an ORF in the right direction were further considered. The first ORF after an ECF56s3-like predicted promoter was annotated against actinobacterial non-supervised orthologous groups (NOGs) and EggNOG mapper with default options (Huerta-Cepas *et al.*, 2017). Finally, the number of organisms where a certain NOG appeared in the set of proteins downstream of an ECF56s3-like predicted promoters was calculated.

#### **8.17. Search for the putative anti- $\sigma$ factor STSU\_11555**

STSU\_11555-like proteins were found using the online version of HHblits (Zimmermann *et al.*, 2018) with default parameter. The local alignment of the 100 sequences with the lowest E-value were used for the construction of an HMM. This HMM was used for searching over the library of proteins in the genetic neighborhood of ECF  $\sigma$  factors. Only proteins with an E-value and a c-Value  $< 0.01$  were further considered. Genetic neighborhoods were defined as the coding sequences located in  $\pm 10$  from the ECF coding sequence. Only ECFs from organisms tagged as “representatives” or “reference” by NCBI were considered (<https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/>). Only RefSeq

genomes were considered when both a RefSeq and a GenBank assembly of the same genome exist. Results were plotted on the ECF phylogenetic tree.

#### 8.18. Bacterial strains and growth conditions

*E. coli* strains, plasmids and vectors used in this study are listed in Table 8.1, Table 8.2 and Table 8.3. *E. coli* was generally cultured in lysogeny broth (LB) (Bertani, 1951). For plate reader experiments, I used MOPS minimal medium (Neidhardt, Bloch and Smith, 1974), M9 minimal medium (Cold Spring Harbor Laboratory, 2010), both supplemented with 0.5% (v/v) glycerol, or MSgg minimal medium (Branda *et al.*, 2004). Cells were cultured at 37°C and 250rpm shaking, when liquid. Strains carrying CRIM helper plasmids were cultured at 30°C (Haldimann and Wanner, 2001). To maintain plasmids, chloramphenicol 25µg/ml, kanamycin 50µg/ml, ampicillin 100µg/mL and/or spectinomycin 100µg/ml were used. For the selection of integrants, chloramphenicol at 6µg/ml was used. For blue-white screening of plasmids with insert, LB plates containing isopropyl β-D-1-thiogalactopyranoside (IPTG) 0.1mM and 5-Bromo-4-chloro-3-indolyl-β-D-galactoside (X-Gal) 40µg/ml were used.

Arabinose was stored at room temperature in 20% (w/v) dilution in ultrapure water (purified by Milli-Q®) and filter-sterilized. ATc was stored at -20°C in a 100µg/mL diluted in 50% ethanol from a stock solution of 2mg/mL. Dilutions from this stock in ultrapure water were freshly prepared before experiments. The final concentration of arabinose in experiments ranged from 0.2% to 0.00001%, whereas the final concentration of ATc ranged from 0 to 100ng/mL, depending on the experiment. Valinomycin was diluted in DMSO and stored at -20°C at a stock concentration of 4mM. Polymyxin B nonapeptide (PMBN) was stored at -20°C at a concentration of 5mg/mL in ultrapure water after filter sterilization. Valinomycin at a working concentration of 3µM (~3.33µg/mL) was used in combination with 5µg/mL of PMBN, as indicated in (Alatossava, Vaara and Baschong, 1984).

#### 8.19. Molecular biology techniques

Polymerase chain reaction (PCR) (Mullis, 1990) was performed to check constructs using Taq DNA Polymerase (New England Biolabs) with oligonucleotides provided by Sigma-Aldrich. Type IIs restriction enzymes (BpiI and BsaI) and T4 DNA ligase were purchased from Thermo Scientific. Transformation of different chemically competent *E. coli* strains was performed according to the Inoue method (Sambrook and Russell, 2006) or using the transformation and storage solution (TSS) methodology (Chung, Niemela and Miller, 1989). Plasmid DNA was extracted from a single colony grown in LB medium using E.Z.N.A. Plasmid Mini Kit I (Omega Bio-Tek).

##### 8.19.1. Construction of genetic circuits

Circuit construction was done using modular cloning (MoClo) (Weber *et al.*, 2011) following the protocol on (Pinto *et al.*, 2018). Like Golden Gate method, MoClo is based on the assembly of genetic parts into bigger constructs using type IIS restriction enzymes, which create overhanging ends outside

of their recognition sequence. Overhanging ends of genetic parts that need to be assembled next to each other are complementary (Weber *et al.*, 2011). Level 0 parts contain either an ECF coding sequence, an anti- $\sigma$  factor coding sequence or a  $P_{ecf}$  promoter (the promoter controlled by an ECF), extracted from (Rhodius *et al.*, 2013), as well as a ribosome binding site (RBS), a terminator or an insulator (Pinto *et al.*, 2018) (Table 8.1). Level 0 parts were assembled into transcription units in MoClo level 1, which were further assembled into level M constructs (Table 8.1). Level M assembly was done into pSVM-mc vector (Pinto *et al.*, 2018) (Table 8.2) for the construction of medium copy number reporter plasmids. Standard MoClo vectors were used for the insertion of the genetic parts in the different levels (Engler *et al.*, 2014). However, pSV004 vector from the CRIMoClo system was used for the assembly of level M parts for genomic integration (Vecchione and Fritz, 2019) (Table 8.2). MoClo constructs were tested by PCR using appropriate primers (Table 8.4).

#### 8.19.2. Genomic integration using CRIMoClo

Genomic integration at the HK022 *attB* was done using CRIMoClo as in (Vecchione and Fritz, 2019). This plasmid contains the  $\gamma$  replication origin of R6K, which can only be replicated by the trans-acting  $\Pi$  protein encoded by *pir* in DH5 $\alpha$   $\lambda$ pir strain (Haldimann and Wanner, 2001). Inserts for integration built during this work were composed of the ECF coding sequence under the control of the arabinose-inducible promoter ( $P_{BAD-ecf}$ ) and the coding sequence of its partner anti- $\sigma$  factor under the control of an ATc-inducible promoter ( $P_{tet-as}$ ) (Table 8.3).  $P_{BAD-ecf}$  and  $P_{tet-as}$  were separated by a  $\sim$ 300bp insulator with several terminators (Vecchione and Fritz, 2019). This circuit was ensembled in pSV004, a level M MoClo vector compatible with CRIM integration at HK022 *attB* (Vecchione and Fritz, 2019), and cloned into DH5 $\alpha$   $\lambda$ pir. This plasmid was integrated into the genome of GFC0203, a SV01 strain (*pir*<sup>-</sup>) with the helper plasmid pAH69 (Vecchione and Fritz, 2019). pAH69 contains the coding sequence of the integrase for integration in the HK022 *attB* under the control of a promoter only functional at 42°C (Haldimann and Wanner, 2001). Due to the utilization of *oriR101* replication origin, this helper plasmid would be cured at 42°C (Haldimann and Wanner, 2001). Therefore, transformation of the pSV004-based plasmids into this strain would result in the integration of their insert, yielding chloramphenicol-resistant strains (Table 8.3, GFC0411-GFC0413). Single integration events were verified by colony PCR using primers P1, P2, P3 and P4, where P1 and P4 are specific of HK022 *attB* site, as indicated in (Haldimann and Wanner, 2001). These strains were latter transformed with a pSVM-mc-based reporter plasmid that contained the target promoter of the ECF controlling the expression of *gfp* ( $P_{ecf-gfp}$ ) (Table 8.3, GFC0414-GFC0418). In the case of ECF02\_2817, which corresponds to RpoE in *E. coli*, the reporter plasmid with  $P_{ecf02\_2817-gfp}$  was transformed into SV01 reporter strain (Table 8.3, GFC0419).

**Table 8.1. Plasmids used for this study**

<b>Level 0 MoClo plasmids</b>						
<b>Name</b>	<b>Genetic part</b>	<b>Description</b>	<b>Backbone</b>	<b>Selection resistance</b>	<b>Chassis</b>	<b>Source</b>
pDM217	<i>Pecf02_2817</i>	ECF02_2817 promoter	pICH41233	Spectinomycin 50µg/mL	DH5α	Dr. Doreen Meier based on (Rhodius <i>et al.</i> , 2013)
pSV0-9_003	RBS	Strong RBS	pICH41246	Spectinomycin 50µg/mL	DH5α	Dr. Stefano Vecchione based on (Vellanoweth and Rabinowitz, 1992)
pSV0-15_001	<i>gfp</i>	GFP coding sequence	pICH41308	Spectinomycin 50µg/mL	DH5α	Dr. Stefano Vecchione based on (Bisicchia, Botella and Devine, 2010)
pSV0-11_001	L3S2P21	Terminator	pICH41276	Spectinomycin 50µg/mL	DH5α	Dr. Stefano Vecchione based on (Chen <i>et al.</i> , 2013)
<b>Level 1 MoClo plasmids</b>						
<b>Name</b>	<b>Genotype (insert)</b>	<b>Description</b>	<b>Backbone</b>	<b>Resistance</b>	<b>Host</b>	<b>Source</b>
pSV1-1L_0022	<i>Pecf15_up436</i>	ECF15_436 promoter	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-3L_0004	<i>Pecf16_3622</i>	ECF16_3622 promoter	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-1L_0024	<i>Pecf22_up1147</i>	ECF22_1147 promoter	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-3L_0007	<i>Pecf19_up1315</i>	ECF28 promoter	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-1L_0027	<i>Pecf38_up1322</i>	ECF38_1322 promoter	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-1L_0031	<i>P<sub>BAD</sub>-ecf22_4450</i>	ECF22_4450 coding sequence under the control of P <sub>BAD</sub>	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-2L_0007	<i>Du300+L3S2P21</i>	Insulator (300bp) + Terminator	pICH47742	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pICH54033	Dummy 3	Dummy needed for MoClo	pBIN19	Ampicillin 100µg/mL	DH5α	(Engler <i>et al.</i> , 2014)
pSV1-4L_0004	<i>Du300+L3S2P21</i>	Insulator (300bp) + Terminator	pICH47761	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-5L_0026	<i>P<sub>tet</sub>-as22_4450</i>	AS22_4450 coding sequences under the control of P <sub>tet</sub>	pICH47772	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pICH50914	L5E	End linker 5	pUC19	Ampicillin 100µg/mL	DH5α	(Engler <i>et al.</i> , 2014)
pSV1-1L_0016	<i>P<sub>BAD</sub>-ecf28_1088</i>	ECF28_1088 coding sequence under the control of P <sub>BAD</sub> promoter	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-5L_0029	<i>P<sub>tet</sub>-as28_1088</i>	AS28_1088 coding sequence under the control of P <sub>tet</sub>	pICH47772	Ampicillin 100µg/mL	DH5α	Dr. Stefano Vecchione
pSV1-1L_0034	<i>P<sub>BAD</sub>-ecf38_1322</i>	ECF38_1322 coding sequence under the control of P <sub>BAD</sub> promoter	pICH47732	Ampicillin 100µg/mL	DH5α	(Pinto <i>et al.</i> , 2018)
pSV1-	<i>P<sub>tet</sub>-as38_1322</i>	AS38_1322 coding	pICH4777	Ampicillin	DH5α	Dr. Stefano

5L_0033		sequence under the control of $P_{tet}$	2	100µg/mL		Vecchione
pDC1-1L_0115	pDM217+pSV0-9_003+pSV0-15_001+pSV0-11_001	<i>Pecf02_2817-gfp</i>	pICH4773 2	Ampicillin 100µg/mL	DH5α	This work

**Level M MoClo plasmids**

Name	Genotype (insert)	Description	Backbone	Resistance	Host	Source
pDCM-MC_244	pDC1-1L_0001	<i>Pecf02_2817-gfp</i>	pSVM-mc	Kanamycin 50µg/mL	DH5α	This work
pDCM-MC_245	pSV1-1L_0022	<i>Pecf15_up436-gfp</i>	pSVM-mc	Kanamycin 50µg/mL	DH5α	This work
pDCM-MC_246	pSV1-3L_0004	<i>Pecf16_3622-gfp</i>	pSVM-mc	Kanamycin 50µg/mL	DH5α	This work
pDCM-MC_247	pSV1-1L_0024	<i>Pecf22_up1147-gfp</i>	pSVM-mc	Kanamycin 50µg/mL	DH5α	This work
pDCM-MC_248	pSV1-3L_0007	<i>Pecf19_up1315(Pecf28)-gfp</i>	pSVM-mc	Kanamycin 50µg/mL	DH5α	This work
pDCM-MC_249	pSV1-1L_0027	<i>Pecf38_up1322-gfp</i>	pSVM-mc	Kanamycin 50µg/mL	DH5α	This work

**CRIMoClo plasmids**

Name	Insert from	Description	Backbone	Resistance	Host	Source
pDC001	pSV1-1L_0031, pSV1-2L_0007, pICH54033, pSV1-4L_0004, pSV1-5L_0026+pICH50914	$P_{BAD-ecf22_4450}$ , $P_{tet-as22_4450}$	pSV004	Chloramphenicol 125µg/mL	DH5α λpir	This work
pDC002	pSV1-1L_0016, pSV1-2L_0007, pICH54033, pSV1-4L_0004, pSV1-5L_0029+pICH50914	$P_{BAD-ecf28_1088}$ , $P_{tet-as28_1088}$	pSV004	Chloramphenicol 125µg/mL	DH5α λpir	This work
pDC003	pSV1-1L_0034, pSV1-2L_0007, pICH54033, pSV1-4L_0004, pSV1-5L_0033, pICH50914	$P_{BAD-ecf38_1322}$ , $P_{tet-as38_1322}$	pSV004	Chloramphenicol 125µg/mL	DH5α λpir	This work

**Other plasmids**

Name	Insert from	Description	Backbone	Resistance	Host	Source
pAH69	Pr+int (HK022), cI, oriR101, repA101	Helper plasmid with integrase for integration in HK022 <i>attB</i> site, temperature-conditional replication and integrase expression	pINT-ts	Ampicillin 100µg/mL	DH5α	(Haldimann and Wanner, 2001)

**Table 8.2. Vectors used for this study**

Name	Description	Host	Resistance	Source
pICH41233	Level 0-1 (Promoter)	DH5α	Spectinomycin 50µg/mL	(Engler <i>et al.</i> , 2014)
pICH41246	Level 0-9 (RBS)	DH5α	Spectinomycin 50µg/mL	(Engler <i>et al.</i> , 2014)
pICH41308	Level 0-15 (CDS)	DH5α	Spectinomycin 50µg/mL	(Engler <i>et al.</i> , 2014)
pICH41276	Level 0-11 (terminator)	DH5α	Spectinomycin 50µg/mL	(Engler <i>et al.</i> , 2014)
pICH47732	Level 1-1L	DH5α	Ampicillin 100µg/mL	(Engler <i>et al.</i> , 2014)
pSVM-mc	Medium-copy number (p15A ori) for cloning in level M	DH5α	Kanamycin 50µg/mL	(Pinto <i>et al.</i> , 2018)
pSV004	HK0022 <i>attP</i> site for usage as level M	DH5α	Chloramphenicol	(Vecchione and

MoClo for chromosomal integration		$\lambda$ pir	25 $\mu$ g/mL	Fritz 2019)	
<b>Table 8.3. Strains used during this study.</b>					
Name	Source strain	Genotype	Description	Resistance	Source
DH5 $\alpha$	DH1	F- $\Phi$ 80 <i>lacZ</i> $\Delta$ M15 $\Delta$ ( <i>lacZYA-argF</i> ) U169 <i>recA1 endA1 hsdR17</i> (rK-, mK+) <i>phoA</i> <i>supE44 thi-1 gyrA96</i> <i>relA1</i> $\lambda$ -	Cloning strain	-	(Grant <i>et al.</i> , 1990)
SV01	MK01	<i>rph-1</i> , $\lambda$ -, $\Delta$ ( <i>araD-araB</i> )567, $\Delta$ <i>lacZ</i> 4787(:: <i>rrnB-3</i> ), $\Delta$ ( <i>araH-araF</i> )570(:: <i>FRT</i> ), $\Delta$ <i>AraEp-532</i> :: <i>FRT</i> ), $\phi$ <i>Pcp8araE535</i> , $\Delta$ ( <i>rhaD-rhaB</i> )568, <i>hsdR514</i> , $\Delta$ <i>lacI</i> :: <i>Lox</i> )	Reporter strain. Unable to catabolize arabinose.	-	(Pinto <i>et al.</i> , 2018)
DH5 $\alpha$ $\lambda$ pir	DH5 $\alpha$	$\lambda$ +	Cloning of CRIMoClo plasmids	Nalidixic acid	(Schindler <i>et al.</i> , 2016)
GFC0203	SV01	pAH69	Reporter strain for integration in HK022 <i>attB</i> site	Ampicillin 100 $\mu$ g/mL	(Vecchione and Fritz 2019)
GFC0013	SV01	$P_{tet}$ - <i>gfp</i>	Positive control	Chloramphenicol 6 $\mu$ g/mL, Kanamycin 50 $\mu$ g/mL	Dr. Stefano Vecchione
GFC0014	SV01	$P_{BAD}$ - <i>gfp</i>	Positive control	Chloramphenicol 6 $\mu$ g/mL, Kanamycin 50 $\mu$ g/mL	Dr. Stefano Vecchione
GFC0348	GFC0203	-	$P_{BAD}$ - <i>ecf15_436</i> , $P_{tet}$ - <i>as15_436</i> integrated in HK022 <i>attB</i>	Chloramphenicol 6 $\mu$ g/mL	Dr. Stefano Vecchione
GFC0349	GFC0203	-	$P_{BAD}$ - <i>ecf16_3622</i> , $P_{tet}$ - <i>as16_3622</i> integrated in HK022 <i>attB</i>	Chloramphenicol 6 $\mu$ g/mL	Dr. Stefano Vecchione
GFC0411	GFC0203	pDC001 integrated in HK022 <i>attB</i>	$P_{BAD}$ - <i>ecf22_4450</i> , $P_{tet}$ - <i>as22_4450</i> integrated at HK0022 <i>attB</i>	Chloramphenicol 6 $\mu$ g/mL	This work
GFC0412	GFC0203	pDC002 integrated in HK022 <i>attB</i>	$P_{BAD}$ - <i>ecf28_1088</i> , $P_{tet}$ - <i>as28_1088</i> integrated at HK0022 <i>attB</i>	Chloramphenicol 6 $\mu$ g/mL	This work
GFC0413	GFC0203	pDC003 integrated in HK022 <i>attB</i>	$P_{BAD}$ - <i>ecf38_1322</i> , $P_{tet}$ - <i>as38_1322</i> integrated at HK0022 <i>attB</i>	Chloramphenicol 6 $\mu$ g/mL	This work
GFC0414	GFC0348	pDCM-MC_245	$P_{BAD}$ - <i>ecf15_436</i> , $P_{tet}$ - <i>as15_436</i> integrated in HK022 <i>attB</i> , <i>Pecf15_up436-gfp</i>	Chloramphenicol 6 $\mu$ g/mL, Kanamycin 50 $\mu$ g/mL	This work
GFC0415	GFC0349	pDCM-MC_246	$P_{BAD}$ - <i>ecf16_3622</i> , $P_{tet}$ - <i>as16_3622</i> integrated in HK022 <i>attB</i> , <i>Pecf16_3622-gfp</i>	Chloramphenicol 6 $\mu$ g/mL, Kanamycin 50 $\mu$ g/mL	This work
GFC0416	GFC0411	pDCM-MC_247	$P_{BAD}$ - <i>ecf22_4450</i> , $P_{tet}$ - <i>as22_4450</i> integrated at HK0022 <i>attB</i> , <i>Pecf22_up1147-gfp</i>	Chloramphenicol 6 $\mu$ g/mL, Kanamycin 50 $\mu$ g/mL	This work
GFC0417	GFC0412	pDCM-MC_248	$P_{BAD}$ - <i>ecf28_1088</i> , $P_{tet}$ - <i>as28_1088</i> integrated at HK0022 <i>attB</i> , <i>Pecf19_up1315(Pecf28)-gfp</i>	Chloramphenicol 6 $\mu$ g/mL, Kanamycin 50 $\mu$ g/mL	This work
GFC0418	GFC0413	pDCM-MC_249	$P_{BAD}$ - <i>ecf38_1322</i> , $P_{tet}$ - <i>as38_1322</i> integrated at HK0022 <i>attB</i> , <i>Pecf38_up1322-gfp</i>	Chloramphenicol 6 $\mu$ g/mL, Kanamycin 50 $\mu$ g/mL	This work

GFC0419	GFC0203	pDCM-MC_244	<i>Pecf02_2817-gfp</i>	Kanamycin 50µg/mL	This work
---------	---------	-------------	------------------------	----------------------	-----------

Table 8.4. Primers used during this study.

Primers			
Name	Description	Sequence (5' to 3')	Purpose
GF0516	HK022 P1	GGAATCAATGCCTGAGTG	
GF0517	HK022 P4	GGCATCAACAGCACATTC	Check single chromosomal
GF0520	P2	ACTTAACGGCTGACATGG	integration
GF0521	P3	ACGAGTATCGAGATGGCA	
GF0002	$P_{BAD}$ rev	AAATCTCGAGGCCCAAAAAACGGGTATG	Check MoClo level 1 and CRIMoClo
GF0537	MoClo M-P mcs+ori fw	CACATTGCGGACGTTTTTAATG	plasmid inserts
GF0070	Level M fw	GCTGGTGGCAGGATATATTG	Check insertion in pSVM-mc
GF0071	Level M rev	GATAAACCTTTTCACGCCCT	

### 8.19.3. Tracking fluorescence emitted by GFP and optical density

The protocol for fluorescence and growth measurement in *E. coli* is similar to the one used in (Pinto *et al.*, 2018). A single colony from the appropriate strain was cultured in LB medium with the appropriate selection antibiotics at 37°C and 250rpm overnight. 1µL of culture was transferred into 3mL of minimal medium (MOPS, M9 or MSgg minimal media). Cultures were supplemented with the appropriate selection antibiotics, ATc, arabinose, and cultured at 37°C and 250rpm. When cultures reached OD<sub>600</sub> 0.4-0.8, indicating exponential phase, they were diluted to OD<sub>600</sub> 0.05 in 37°C minimal medium with the appropriate inducers. OD<sub>600</sub> and GFP fluorescence emission of 100µL cultures in transparent 96-well plates were measured every 5min in a 10-hour time course experiment, by using a TECAN<sup>®</sup> Infinite 200 PRO plate reader. SV01 was used as wild-type strain for blanking fluorescence. Strains that contain a reporter plasmid with either  $P_{BAD}$ -*ecf* (GFC0014) or  $P_{ter}$ -*ecf* (GFC0013) were used as positive controls of arabinose and ATc activity. Minimal medium was used for blanking the OD<sub>600</sub>. When needed, ATc was removed collecting the cells by centrifugation during 2min at 2500g and room temperature, and resuspending them in media without ATc. This washing step was repeated twice. Cultures were induced after 2 hours from the start of the plate reader experiment (at the beginning of cycle 24) with 5µL of inducer solution, when needed.

## References

- Ahmed, S. and Booth, I. R. (1983) 'The use of valinomycin, nigericin and trichlorocarbanilide in control of the protonmotive force in *Escherichia coli* cells', *Biochemical Journal*, 212(1), pp. 105–112. doi: 10.1042/bj2120105.
- Ait-Bara, S. *et al.* (2017) 'Competitive folding of RNA structures at a termination-antitermination site', *RNA*. Cold Spring Harbor Laboratory Press, 23(5), pp. 721–734. doi: 10.1261/rna.060178.116.
- Alatossava, T., Vaara, M. and Baschong, W. (1984) 'Polymyxin B nonapeptide sensitizes *Escherichia coli* to valinomycin and A23187 ionophores', *FEMS Microbiology Letters*, 22(3), pp. 249–251. doi: 10.1111/j.1574-6968.1984.tb00736.x.
- Alba, B. M. *et al.* (2002) 'DegS and YaeL participate sequentially in the cleavage of RseA to activate the  $\sigma$ E-dependent extracytoplasmic stress response', *Genes and Development*, 16(16), pp. 2156–2168. doi: 10.1101/gad.1008902.
- Altschul, S. F. *et al.* (1990) 'Basic local alignment search tool', *Journal of Molecular Biology*. doi: 10.1016/S0022-2836(05)80360-2.
- Altschul, S. F. *et al.* (2009) 'PSI-BLAST pseudocounts and the minimum description length principle', *Nucleic Acids Research*, 37(3), pp. 815–824. doi: 10.1093/nar/gkn981.
- Anthony, J. R., Newman, J. D. and Donohue, T. J. (2004) 'Interactions between the *Rhodospirillum rubrum* ECF sigma factor,  $\sigma$  E, and its anti-sigma factor, ChrR', *Journal of Molecular Biology*, 341(2), pp. 345–360. doi: 10.1016/j.jmb.2004.06.018.
- Anthony, J. R., Warczak, K. L. and Donohue, T. J. (2005) 'A transcriptional response to singlet oxygen, a toxic byproduct of photosynthesis', *Proceedings of the National Academy of Sciences of the United States of America*, 102(18), pp. 6502–6507. doi: 10.1073/pnas.0502225102.
- Armache, K. J. *et al.* (2005) 'Structures of complete RNA polymerase II and its subcomplex, Rpb4/7', *Journal of Biological Chemistry*, 280(8), pp. 7131–7134. doi: 10.1074/jbc.M413038200.
- Bailey, T. L. *et al.* (2009) 'MEME SUITE: tools for motif discovery and searching.', *Nucleic acids research*, 37(Web Server issue), pp. W202–8. doi: 10.1093/nar/gkp335.
- Baldassi, C. *et al.* (2014) 'Fast and accurate multivariate Gaussian modeling of protein families: Predicting residue contacts and protein-interaction partners', *PLoS ONE*. doi: 10.1371/journal.pone.0092721.
- Balogh, D. *et al.* (2017) 'Insights into ClpXP proteolysis: heterooligomerization and partial deactivation enhance chaperone affinity and substrate turnover in *Listeria monocytogenes*', *Chemical Science*. doi: 10.1039/c6sc03438a.
- Barchinger, S. E. *et al.* (2016) 'Regulation of gene expression in *Shewanella oneidensis* MR-1 during electron acceptor limitation and bacterial nanowire formation', *Applied and Environmental Microbiology*. American Society for Microbiology, 82(17), pp. 5428–5443. doi: 10.1128/AEM.01615-16.
- Barik, S. *et al.* (2010) 'RseA, the SigE specific anti-sigma factor of *Mycobacterium tuberculosis*, is inactivated by phosphorylation-dependent ClpC1P2 proteolysis', *Molecular Microbiology*, 75(3), pp. 592–606. doi: 10.1111/j.1365-2958.2009.07008.x.
- Basehoar, A. D., Zanton, S. J. and Pugh, B. F. (2004) 'Identification and distinct regulation of yeast TATA box-containing genes', *Cell*, 116(5), pp. 699–709. doi: 10.1016/S0092-8674(04)00205-3.
- Bastiaansen, K. C. *et al.* (2014) 'The Prc and RseP proteases control bacterial cell-surface signalling activity', *Environmental Microbiology*, 16(8), pp. 2433–2443. doi: 10.1111/1462-2920.12371.
- Bastiaansen, K. C. *et al.* (2015) 'Self-cleavage of the *Pseudomonas aeruginosa* cell-surface signaling anti-sigma factor FoxR occurs through an N-O acyl rearrangement', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 290(19), pp. 12237–12246. doi:

10.1074/jbc.M115.643098.

Bastiat, B. *et al.* (2012) ‘Sinorhizobium meliloti sigma factors RpoE1 and RpoE4 are activated in stationary phase in response to sulfite.’, *PLoS one*, 7(11), p. e50768. doi: 10.1371/journal.pone.0050768.

Bayer-Santos, E. *et al.* (2018) ‘Xanthomonas citri T6SS mediates resistance to Dictyostelium predation and is regulated by an ECF  $\sigma$  factor and cognate Ser/Thr kinase.’, *Environmental microbiology*, 20(4), pp. 1562–1575. doi: 10.1111/1462-2920.14085.

Bertani, G. (1951) ‘Studies on lysogenesis. I. The mode of phage liberation by lysogenic Escherichia coli.’, *Journal of bacteriology*, 62(3), pp. 293–300.

Bibb, M. J. and Buttner, M. J. (2003) ‘The streptomyces coelicolor developmental transcription factor  $\sigma$ BldN is synthesized as a proprotein’, *Journal of Bacteriology*. doi: 10.1128/JB.185.7.2338-2345.2003.

Bisicchia, P., Botella, E. and Devine, K. M. (2010) ‘Suite of novel vectors for ectopic insertion of GFP, CFP and IYFP transcriptional fusions in single copy at the amyE and bglS loci in Bacillus subtilis’, *Plasmid*, 64(3), pp. 143–149. doi: 10.1016/j.plasmid.2010.06.002.

Bjursell, M. K., Martens, E. C. and Gordon, J. I. (2006) ‘Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, Bacteroides thetaiotaomicron, to the suckling period’, *Journal of Biological Chemistry*, 281(47), pp. 36269–36279. doi: 10.1074/jbc.M606509200.

Blom, N., Gammeltoft, S. and Brunak, S. (1999) ‘Sequence and structure-based prediction of eukaryotic protein phosphorylation sites’, *Journal of Molecular Biology*. Academic Press, 294(5), pp. 1351–1362. doi: 10.1006/jmbi.1999.3310.

Branda, S. S. *et al.* (2004) ‘Genes involved in formation of structured multicellular communities by Bacillus subtilis’, *Journal of Bacteriology*, 186(12), pp. 3970–3979. doi: 10.1128/JB.186.12.3970-3979.2004.

Braun, V. (1997) ‘Surface signaling: Novel transcription initiation mechanism starting from the cell surface’, *Archives of Microbiology*, pp. 325–331. doi: 10.1007/s002030050451.

Braun, V. and Mahren, S. (2005) ‘Transmembrane transcriptional control (surface signalling) of the Escherichia coli Fec type’, *FEMS Microbiology Reviews*, pp. 673–684. doi: 10.1016/j.femsre.2004.10.001.

Braun, V., Mahren, S. and Ogierman, M. (2003) ‘Regulation of the FecI-type ECF sigma factor by transmembrane signalling.’, *Current opinion in microbiology*, 6(2), pp. 173–80. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/12732308> (Accessed: 19 July 2019).

Bremer, H., Dennis, P. and Ehrenberg, M. (2003) ‘Free RNA polymerase and modeling global transcription in Escherichia coli’, *Biochimie*. Elsevier, 85(6), pp. 597–609. doi: 10.1016/S0300-9084(03)00105-6.

Brister, J. R. *et al.* (2015) ‘NCBI viral Genomes resource’, *Nucleic Acids Research*. doi: 10.1093/nar/gku1207.

Broberg, C. A., Calder, T. J. and Orth, K. (2011) ‘Vibrio parahaemolyticus cell biology and pathogenicity determinants’, *Microbes and Infection*, pp. 992–1001. doi: 10.1016/j.micinf.2011.06.013.

Brown, D. P., Krishnamurthy, N. and Sjölander, K. (2007) ‘Automated protein subfamily identification and classification.’, *PLoS computational biology*, 3(8), p. e160. doi: 10.1371/journal.pcbi.0030160.

Brueckner, F., Ortiz, J. and Cramer, P. (2009) ‘A movie of the RNA polymerase nucleotide addition cycle’, *Current Opinion in Structural Biology*, pp. 294–299. doi: 10.1016/j.sbi.2009.04.005.

Burgess, R. R. *et al.* (1969) ‘Factor stimulating transcription by RNA polymerase’, *Nature*,

221(5175), pp. 43–46. doi: 10.1038/221043a0.

Burr, T. *et al.* (2000) ‘DNA sequence elements located immediately upstream of the -10 hexamer in *Escherichia coli* promoters: A systematic study’, *Nucleic Acids Research*. Oxford University Press, 28(9), pp. 1864–1870. doi: 10.1093/nar/28.9.1864.

Čabart, P. *et al.* (2011) ‘Transcription factor TFIIF is not required for initiation by RNA polymerase II, but it is essential to stabilize transcription factor TFIIB in early elongation complexes’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(38), pp. 15786–15791. doi: 10.1073/pnas.1104591108.

Camacho, C. *et al.* (2009) ‘BLAST+: Architecture and applications’, *BMC Bioinformatics*, 10. doi: 10.1186/1471-2105-10-421.

Campbell, E. A., Masuda, S., *et al.* (2002) ‘Crystal structure of the *Bacillus stearothermophilus* anti- $\sigma$  factor SpoIIAB with the sporulation  $\sigma$  factor  $\sigma^F$ ’, *Cell*. Cell Press, 108(6), pp. 795–807. doi: 10.1016/S0092-8674(02)00662-1.

Campbell, E. A., Muzzin, O., *et al.* (2002) ‘Structure of the bacterial RNA polymerase promoter specificity  $\sigma$  subunit’, *Molecular Cell*. doi: 10.1016/S1097-2765(02)00470-7.

Campbell, E. A. *et al.* (2003) ‘Crystal structure of *Escherichia coli*  $\sigma^E$  with the cytoplasmic domain of its anti- $\sigma$  RseA’, *Molecular Cell*. doi: 10.1016/S1097-2765(03)00148-5.

Campbell, E. A. *et al.* (2007) ‘A conserved structural module regulates transcriptional responses to diverse stress signals in bacteria.’, *Molecular cell*, 27(5), pp. 793–805. doi: 10.1016/j.molcel.2007.07.009.

Carafa, Y. d. A., Brody, E. and Thermes, C. (1990) ‘Prediction of rho-independent *Escherichia coli* transcription terminators. A statistical analysis of their RNA stem-loop structures’, *Journal of Molecular Biology*, 216(4), pp. 835–858. doi: 10.1016/S0022-2836(99)80005-9.

Carter, R. and Drouin, G. (2009) ‘Structural differentiation of the three eukaryotic RNA polymerases’, *Genomics*, 94(6), pp. 388–396. doi: 10.1016/j.ygeno.2009.08.011.

Casas-Pastor, D. *et al.* (2019) ‘Expansion and re-classification of the extracytoplasmic function (ECF)  $\sigma$  factor family’, *bioRxiv*. Cold Spring Harbor Laboratory, p. 2019.12.11.873521. doi: 10.1101/2019.12.11.873521.

Cashel, M., Hsu, L. M. and Hernandez, V. J. (2003) ‘Changes in conserved region 3 of *Escherichia coli*  $\sigma^{70}$  reduce abortive transcription and enhance promoter escape’, *Journal of Biological Chemistry*, 278(8), pp. 5539–5547. doi: 10.1074/jbc.M211430200.

Casino, P., Rubio, V. and Marina, A. (2010) ‘The mechanism of signal transduction by two-component systems’, *Current Opinion in Structural Biology*, pp. 763–771. doi: 10.1016/j.sbi.2010.09.010.

Castro, A. N. *et al.* (2018) ‘Signal peptidase is necessary and sufficient for site 1 cleavage of RsiV in *Bacillus subtilis* in response to lysozyme’, *Journal of Bacteriology*. American Society for Microbiology, 200(11). doi: 10.1128/JB.00663-17.

Cermakian, N. *et al.* (1997) ‘On the evolution of the single-subunit RNA polymerases’, *Journal of Molecular Evolution*, 45(6), pp. 671–681. doi: 10.1007/PL00006271.

Cervený, L. *et al.* (2013) ‘Tetratricopeptide repeat motifs in the world of bacterial pathogens: Role in virulence mechanisms’, *Infection and Immunity*, 81(3), pp. 629–635. doi: 10.1128/IAI.01035-12.

Chaba, R. *et al.* (2011) ‘Signal integration by DegS and RseB governs the  $\sigma^E$ -mediated envelope stress response in *Escherichia coli*’, *Proceedings of the National Academy of Sciences of the United States of America*, 108(5), pp. 2106–2111. doi: 10.1073/pnas.1019277108.

Chabert, V. *et al.* (2019) ‘Model peptide for anti-sigma factor domain HHCC zinc fingers: High reactivity toward 1O2 leads to domain unfolding’, *Chemical Science*. Royal Society of Chemistry, 10(12), pp. 3608–3615. doi: 10.1039/c9sc00341j.

- Chamberlin, M., Mcgrath, J. and Waskell, L. (1970) 'New RNA polymerase from escherichia coli infected with bacteriophage T7', *Nature*, 228(5268), pp. 227–231. doi: 10.1038/228227a0.
- Chauhan, R. *et al.* (2016) 'Reconstruction and topological characterization of the sigma factor regulatory network of Mycobacterium tuberculosis', *Nature Communications*. Nature Publishing Group, 7. doi: 10.1038/ncomms11062.
- Chen, Y. J. *et al.* (2013) 'Characterization of 582 natural and synthetic terminators and quantification of their design constraints', *Nature Methods*, 10(7), pp. 659–664. doi: 10.1038/nmeth.2515.
- Chevalier, S. *et al.* (2019) 'Extracytoplasmic function sigma factors in Pseudomonas aeruginosa.', *Biochimica et biophysica acta. Gene regulatory mechanisms*, 1862(7), pp. 706–721. doi: 10.1016/j.bbagr.2018.04.008.
- Chi, W. *et al.* (2015) 'Plastid sigma factors: Their individual functions and regulation in transcription', *Biochimica et Biophysica Acta - Bioenergetics*. Elsevier B.V., pp. 770–778. doi: 10.1016/j.bbabi.2015.01.001.
- Chiang, M. M. T. and Mirkin, B. (2010) 'Intelligent choice of the number of clusters in k-means clustering: An experimental study with different cluster spreads', *Journal of Classification*, 27(1), pp. 3–40. doi: 10.1007/s00357-010-9049-5.
- Chihara, S. *et al.* (1973) 'Enzymatic Degradation of Colistin Isolation and Identification of  $\alpha$ -N-Acyl  $\alpha,\gamma$ -Diaminobutyric Acid and Colistin Nonapeptide', *Agricultural and Biological Chemistry*, 37(11), pp. 2455–2463. doi: 10.1080/00021369.1973.10861030.
- Chung, C. T., Niemela, S. L. and Miller, R. H. (1989) 'One-step preparation of competent Escherichia coli: transformation and storage of bacterial cells in the same solution', *Proceedings of the National Academy of Sciences of the United States of America*, 86(7), pp. 2172–2175. doi: 10.1073/pnas.86.7.2172.
- Cock, P. J. A. *et al.* (2009) 'Biopython: Freely available Python tools for computational molecular biology and bioinformatics', *Bioinformatics*. doi: 10.1093/bioinformatics/btp163.
- Cold Spring Harbor Laboratory (2010) *M9 minimal medium (standard)*, *Cold Spring Harbor Protocols*. Cold Spring Harbor Laboratory. doi: 10.1101/pdb.rec12295.
- Conaway, R. C. *et al.* (1991) 'Mechanism of promoter selection by RNA polymerase II: Mammalian transcription factors  $\alpha$  and  $\beta\gamma$  promote entry of polymerase into the preinitiation complex', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 88(14), pp. 6205–6209. doi: 10.1073/pnas.88.14.6205.
- Cooper, S. J. *et al.* (2006) 'Comprehensive analysis of transcriptional promoter structure and function in 1% of the human genome', *Genome Research*, 16(1), pp. 1–10. doi: 10.1101/gr.4222606.
- Cousin, C. *et al.* (2013) 'Protein-serine/threonine/tyrosine kinases in bacterial signaling and regulation', *FEMS Microbiology Letters*, 346(1), pp. 11–19. doi: 10.1111/1574-6968.12189.
- Cramer, P. *et al.* (2000) 'Architecture of RNA polymerase II and implications for the transcription mechanism.', *Science (New York, N.Y.)*, 288(5466), pp. 640–9. doi: 10.1126/science.288.5466.640.
- Cramer, P. (2002) 'Multisubunit RNA polymerases', *Current Opinion in Structural Biology*. Elsevier Ltd, pp. 89–97. doi: 10.1016/S0959-440X(02)00294-4.
- Cramer, P. *et al.* (2008) 'Structure of Eukaryotic RNA Polymerases', *Annual Review of Biophysics*. Annual Reviews, 37(1), pp. 337–352. doi: 10.1146/annurev.biophys.37.032807.130008.
- Crick, F. H. (1958) 'On protein synthesis', *Symposia of the Society for Experimental Biology*, 12, pp. 138–163.
- Crooks, G. E. *et al.* (2004) 'WebLogo: A sequence logo generator', *Genome Research*. doi: 10.1101/gr.849004.
- D'Andrea, L. D. and Regan, L. (2003) 'TPR proteins: the versatile helix.', *Trends in biochemical*

*sciences*, 28(12), pp. 655–62. doi: 10.1016/j.tibs.2003.10.007.

Deighan, P. and Hochschild, A. (2006) ‘Conformational toggle triggers a modulator of RNA polymerase activity’, *Trends in Biochemical Sciences*, pp. 424–426. doi: 10.1016/j.tibs.2006.06.004.

Devkota, S. R. *et al.* (2017) ‘Structural insights into the regulation of *Bacillus subtilis* SigW activity by anti-sigma RsiW’, *PLoS ONE*. doi: 10.1371/journal.pone.0174284.

Dintner, S. *et al.* (2011) ‘Coevolution of ABC transporters and two-component regulatory systems as resistance modules against antimicrobial peptides in Firmicutes bacteria’, *Journal of Bacteriology*. doi: 10.1128/JB.05175-11.

Dombroski, A. J., Walter, W. A. and Gross, C. A. (1993) ‘Amino-terminal amino acids modulate  $\sigma$ -factor DNA-binding activity’, *Genes and Development*, 7(12 A), pp. 2446–2455. doi: 10.1101/gad.7.12a.2446.

Duchêne, S. *et al.* (2016) ‘Genome-scale rates of evolutionary change in bacteria’, *Microbial genomics*, 2(11), p. e000094. doi: 10.1099/mgen.0.000094.

Dufour, A. and Haldenwang, W. G. (1994) ‘Interactions between a *Bacillus subtilis* anti-sigma Factor (*RsbW*) and Its Antagonist (*RsbV*)’, *Journal of Bacteriology*.

Duncan, M. C. *et al.* (2018) ‘*Vibrio cholerae* motility exerts drag force to impede attack by the bacterial predator *Bdellovibrio bacteriovorus*’, *Nature communications*. NLM (Medline), 9(1), p. 4757. doi: 10.1038/s41467-018-07245-3.

Dvir, A., Conaway, R. C. and Conaway, J. W. (1997) ‘A role for TFIIF in controlling the activity of early RNA polymerase II elongation complexes’, *Proceedings of the National Academy of Sciences of the United States of America*, 94(17), pp. 9006–9010. doi: 10.1073/pnas.94.17.9006.

El-Gebali, S. *et al.* (2019) ‘The Pfam protein families database in 2019’, *Nucleic Acids Research*. Oxford University Press, 47(D1), pp. D427–D432. doi: 10.1093/nar/gky995.

Engler, C. *et al.* (2014) ‘A Golden Gate modular cloning toolbox for plants’, *ACS Synthetic Biology*. American Chemical Society, 3(11), pp. 839–843. doi: 10.1021/sb4001504.

Estrem, S. T. *et al.* (1998) ‘Identification of an UP element consensus sequence for bacterial promoters’, *Proceedings of the National Academy of Sciences of the United States of America*, 95(17), pp. 9761–9766. doi: 10.1073/pnas.95.17.9761.

Fang, C. *et al.* (2019) ‘Structures and mechanism of transcription initiation by bacterial ECF factors’, *Nucleic Acids Research*. Oxford University Press (OUP), 47(13), pp. 7094–7104. doi: 10.1093/nar/gkz470.

Feklistov, A. and Darst, S. A. (2011) ‘Structural basis for promoter -10 element recognition by the bacterial RNA polymerase  $\sigma$  subunit’, *Cell*, 147(6), pp. 1257–1269. doi: 10.1016/j.cell.2011.10.041.

Felle, H. *et al.* (1980) ‘Quantitative measurements of membrane potential in *Escherichia coli*’, *Biochemistry*, 19(15), pp. 3585–3590. doi: 10.1021/bi00556a026.

Felsenstein, J. (1985) ‘Confidence Limits on Phylogenies: An Approach Using the Bootstrap’, *Evolution*. JSTOR, 39(4), p. 783. doi: 10.2307/2408678.

Finn, R. D., Clements, J. and Eddy, S. R. (2011) ‘HMMER web server: Interactive sequence similarity searching’, *Nucleic Acids Research*. doi: 10.1093/nar/gkr367.

Fischer, E. H. and Krebs, E. G. (1955) ‘Conversion of phosphorylase b to phosphorylase a in muscle extracts.’, *The Journal of biological chemistry*, 216(1), pp. 121–132.

Fong, N. *et al.* (2015) ‘Effects of Transcription Elongation Rate and Xrn2 Exonuclease Activity on RNA Polymerase II Termination Suggest Widespread Kinetic Competition’, *Molecular Cell*. Cell Press, 60(2), pp. 256–267. doi: 10.1016/j.molcel.2015.09.026.

Foreman, R., Fiebig, A. and Crosson, S. (2012) ‘The *lovK-lovR* two-component system is a regulator of the general stress pathway in *Caulobacter crescentus*’, *Journal of Bacteriology*, 194(12), pp. 3038–

3049. doi: 10.1128/JB.00182-12.

Forrest, D. *et al.* (2017) 'Single-peptide DNA-dependent RNA polymerase homologous to multi-subunit RNA polymerase', *Nature Communications*. Nature Publishing Group, 8. doi: 10.1038/ncomms15774.

Francez-Charlot, A. *et al.* (2009) 'Sigma factor mimicry involved in regulation of general stress response', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0810291106.

Francez-Charlot, A. *et al.* (2015) 'The general stress response in Alphaproteobacteria', *Trends in Microbiology*. Elsevier Ltd, pp. 164–171. doi: 10.1016/j.tim.2014.12.006.

Galperin, M. Y. (2004) 'Bacterial signal transduction network in a genomic perspective+', *Environmental Microbiology*, 6(6), pp. 552–567. doi: 10.1111/j.1462-2920.2004.00633.x.

Gao, B., Mohan, R. and Gupta, R. S. (2009) 'Phylogenomics and protein signatures elucidating the evolutionary relationships among the Gammaproteobacteria', *International Journal of Systematic and Evolutionary Microbiology*, 59(2), pp. 234–247. doi: 10.1099/ijs.0.002741-0.

Gaudion, A. *et al.* (2013) 'Characterisation of the mycobacterium tuberculosis alternative sigma factor SigG: Its operon and regulon', *Tuberculosis*, 93(5), pp. 482–491. doi: 10.1016/j.tube.2013.05.005.

Gehring, A. M., Walker, J. E. and Santangelo, T. J. (2016) 'Transcription regulation in archaea', *Journal of Bacteriology*. American Society for Microbiology, 198(14), pp. 1906–1917. doi: 10.1128/JB.00255-16.

Geiger, J. H. *et al.* (1996) 'Crystal structure of the yeast TFIIA/TBP/DNA complex', *Science*. American Association for the Advancement of Science, 272(5263), pp. 830–836. doi: 10.1126/science.272.5263.830.

Gershenson, N. I. and Ioshikhes, I. P. (2005) 'Synergy of human Pol II core promoter elements revealed by statistical sequence analysis', *Bioinformatics*, 21(8), pp. 1295–1300. doi: 10.1093/bioinformatics/bti172.

Goh, C. S. *et al.* (2000) 'Co-evolution of proteins with their interaction partners', *Journal of Molecular Biology*. doi: 10.1006/jmbi.2000.3732.

Goldman, S. R., Ebright, R. H. and Nickels, B. E. (2009) 'Direct detection of abortive RNA transcripts in vivo', *Science*, 324(5929), pp. 927–928. doi: 10.1126/science.1169237.

Gómez-Santos, N. *et al.* (2011) 'Core from *Myxococcus xanthus* is a Copper-Dependent RNA polymerase sigma factor', *PLoS Genetics*, 7(6). doi: 10.1371/journal.pgen.1002106.

Goutam, K., Gupta, A. K. and Gopal, B. (2017) 'The fused Snoal<sub>L</sub>\_2 domain in the Mycobacterium tuberculosis sigma factor  $\sigma$ J modulates promoter recognition.', *Nucleic acids research*, 45(16), pp. 9760–9772. doi: 10.1093/nar/gkx609.

Grant, S. G. N. *et al.* (1990) 'Differential plasmid rescue from transgenic mouse DNAs into *Escherichia coli* methylation-restriction mutants', *Proceedings of the National Academy of Sciences of the United States of America*, 87(12), pp. 4645–4649. doi: 10.1073/pnas.87.12.4645.

Grass, G., Fricke, B. and Nies, D. H. (2005) 'Control of expression of a periplasmic nickel efflux pump by periplasmic nickel concentrations', in *BioMetals*. doi: 10.1007/s10534-005-3718-6.

Greenwell, R., Nam, T. W. and Donohue, T. J. (2011) 'Features of *Rhodobacter sphaeroides* ChrR required for stimuli to promote the dissociation of  $\sigma$ e/ChrR complexes', *Journal of Molecular Biology*, 407(4), pp. 477–491. doi: 10.1016/j.jmb.2011.01.055.

Grosse, C., Friedrich, S. and Nies, D. H. (2007) 'Contribution of extracytoplasmic function sigma factors to transition metal homeostasis in *Cupriavidus metallidurans* strain CH34.', *Journal of molecular microbiology and biotechnology*, 12(3–4), pp. 227–40. doi: 10.1159/000099644.

Gruber, T. M. and Gross, C. A. (2003) 'Multiple sigma subunits and the partitioning of bacterial

- transcription space.’, *Annual review of microbiology*, 57, pp. 441–66. doi: 10.1146/annurev.micro.57.030502.090913.
- Gupta, S. *et al.* (2007) ‘Quantifying similarity between motifs.’, *Genome biology*, 8(2), p. R24. doi: 10.1186/gb-2007-8-2-r24.
- Hakkila, K. *et al.* (2019) ‘Group 2 Sigma Factors are Central Regulators of Oxidative Stress Acclimation in Cyanobacteria’, *Plant and Cell Physiology*, 60(2), pp. 436–447. doi: 10.1093/pcp/pcy221.
- Haldimann, A. and Wanner, B. L. (2001) ‘Conditional-replication, integration, excision, and retrieval plasmid-host systems for gene structure-function studies of bacteria’, *Journal of Bacteriology*, 183(21), pp. 6384–6393. doi: 10.1128/JB.183.21.6384-6393.2001.
- Halkidi, M., Batistakis, Y. and Vazirgiannis, M. (2001) ‘On clustering validation techniques’, *Journal of Intelligent Information Systems*, 17(2–3), pp. 107–145. doi: 10.1023/A:1012801612483.
- Han, K. *et al.* (2013) ‘Extraordinary expansion of a *Sorangium cellulosum* genome from an alkaline milieu.’, *Scientific reports*, 3, p. 2101. doi: 10.1038/srep02101.
- Hanks, S. K. and Hunter, T. (1995) ‘Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification.’, *The FASEB Journal*, 9(8), pp. 576–596. doi: 10.1096/fasebj.9.8.7768349.
- Hansen, U. M. and McClure, W. R. (1980) ‘Role of the sigma subunit of *Escherichia coli* RNA polymerase in initiation. II. Release of sigma from ternary complexes.’, *The Journal of biological chemistry*, 255(20), pp. 9564–70. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/7000759> (Accessed: 14 November 2019).
- Harper, M. *et al.* (2010) ‘Natural selection in the chicken host identifies 3-deoxy-D-manno-octulosonic acid kinase residues essential for phosphorylation of *Pasteurella multocida* lipopolysaccharide’, *Infection and Immunity*, 78(9), pp. 3669–3677. doi: 10.1128/IAI.00457-10.
- Heinrich, J., Hein, K. and Wiegert, T. (2009) ‘Two proteolytic modules are involved in regulated intramembrane proteolysis of *Bacillus subtilis* RsiW’, *Molecular Microbiology*, 74(6), pp. 1412–1426. doi: 10.1111/j.1365-2958.2009.06940.x.
- Heinrich, J. and Wiegert, T. (2009) ‘Regulated intramembrane proteolysis in the control of extracytoplasmic function sigma factors’, *Research in Microbiology*, 160(9), pp. 696–703. doi: 10.1016/j.resmic.2009.08.019.
- Helmann, J. D. (2002) ‘The extracytoplasmic function (ECF) sigma factors’, *Advances in Microbial Physiology*. doi: 10.1016/S0065-2911(02)46002-X.
- Herrou, J. *et al.* (2012) ‘Structural basis of a protein partner switch that regulates the general stress response of  $\alpha$ -proteobacteria.’, *Proceedings of the National Academy of Sciences of the United States of America*, 109(21), pp. E1415–23. doi: 10.1073/pnas.1116887109.
- Heruth, D. P. *et al.* (1994) ‘Characterization of genetic determinants for R body synthesis and assembly in *Caedibacter taeniospiralis* 47 and 116’, *Journal of Bacteriology*. doi: 10.1128/jb.176.12.3559-3567.1994.
- Hillis, D. M. and Bull, J. J. (1993) ‘An Empirical Test of Bootstrapping as a Method for Assessing Confidence in Phylogenetic Analysis’, *Systematic Biology*, 42(2), pp. 182–192. doi: 10.1093/sysbio/42.2.182.
- Hirata, A., Klein, B. J. and Murakami, K. S. (2008) ‘The X-ray crystal structure of RNA polymerase from Archaea’, *Nature*. Nature Publishing Group, 451(7180), pp. 851–854. doi: 10.1038/nature06530.
- Hirtreiter, A., Grohmann, D. and Werner, F. (2010) ‘Molecular mechanisms of RNA polymerase—the F/E (RPB4/7) complex is required for high processivity in vitro’, *Nucleic Acids Research*, 38(2), pp. 585–596. doi: 10.1093/nar/gkp928.
- Ho, T. D. *et al.* (2011) ‘The *Bacillus subtilis* extracytoplasmic function  $\sigma$  factor  $\sigma_v$  is induced by

- lysozyme and provides resistance to lysozyme', *Journal of Bacteriology*, 193(22), pp. 6215–6222. doi: 10.1128/JB.05467-11.
- Ho, T. D. and Ellermeier, C. D. (2012) 'Extra cytoplasmic function  $\sigma$  factor activation', *Current Opinion in Microbiology*, pp. 182–188. doi: 10.1016/j.mib.2012.01.001.
- Hoang, D. T. *et al.* (2018) 'UFBoot2: Improving the Ultrafast Bootstrap Approximation', *Molecular Biology and Evolution*, 35(2), pp. 518–522. doi: 10.1093/molbev/msx281.
- Hoch, J. A. (2000) 'Two-component and phosphorelay signal transduction', *Current Opinion in Microbiology*. Current Biology Ltd, pp. 165–170. doi: 10.1016/S1369-5274(00)00070-9.
- Holder, M. and Lewis, P. O. (2003) 'Phylogeny estimation: Traditional and Bayesian approaches', *Nature Reviews Genetics*, pp. 275–284. doi: 10.1038/nrg1044.
- Holstege, F. C. P., Fiedler, U. and Timmers, H. T. M. (1997) 'Three transitions in the RNA polymerase II transcription complex during initiation', *EMBO Journal*, 16(24), pp. 7468–7480. doi: 10.1093/emboj/16.24.7468.
- Hooper, S. D. and Berg, O. G. (2003) 'Duplication is more common among laterally transferred genes than among indigenous genes.', *Genome biology*, 4(8). doi: 10.1186/gb-2003-4-8-r48.
- Van Hove, B., Staudenmaier, H. and Braun, V. (1990) 'Novel two-component transmembrane transcription control: Regulation of iron dicitrate transport in *Escherichia coli* K-12', *Journal of Bacteriology*, 172(12), pp. 6749–6758. doi: 10.1128/jb.172.12.6749-6758.1990.
- Huang, H. *et al.* (2011) 'A comprehensive protein-centric ID mapping service for molecular data integration.', *Bioinformatics (Oxford, England)*, 27(8), pp. 1190–1. doi: 10.1093/bioinformatics/btr101.
- Huang, H. D. *et al.* (2005) 'KinasePhos: A web tool for identifying protein kinase-specific phosphorylation sites', *Nucleic Acids Research*, 33(SUPPL. 2). doi: 10.1093/nar/gki471.
- Huang, X. *et al.* (2015a) 'Environmental sensing in *Actinobacteria*: a comprehensive survey on the signaling capacity of this phylum', *Journal of Bacteriology*, 197(15), pp. 2517–2535. doi: 10.1128/JB.00176-15.
- Huang, X. *et al.* (2015b) 'Environmental Sensing in *Actinobacteria*: a Comprehensive Survey on the Signaling Capacity of This Phylum', *Journal of Bacteriology*. American Society for Microbiology Journals, 197(15), pp. 2517–2535. doi: 10.1128/JB.00176-15.
- Huerta-Cepas, J. *et al.* (2016) 'EGGNOG 4.5: A hierarchical orthology framework with improved functional annotations for eukaryotic, prokaryotic and viral sequences', *Nucleic Acids Research*. Oxford University Press, 44(D1), pp. D286–D293. doi: 10.1093/nar/gkv1248.
- Huerta-Cepas, J. *et al.* (2017) 'Fast genome-wide functional annotation through orthology assignment by eggNOG-mapper', *Molecular Biology and Evolution*. Oxford University Press, 34(8), pp. 2115–2122. doi: 10.1093/molbev/msx148.
- Hurwitz, J. *et al.* (1961) 'The enzymatic incorporation of ribonucleotides into RNA and the role of DNA.', *Cold Spring Harbor symposia on quantitative biology*, 26, pp. 91–100. doi: 10.1101/SQB.1961.026.01.014.
- Huse, M. and Kuriyan, J. (2002) 'The conformational plasticity of protein kinases', *Cell*. Cell Press, pp. 275–282. doi: 10.1016/S0092-8674(02)00741-9.
- Ibrahim, I. M. *et al.* (2016) 'Probing the nucleotide-binding activity of a redox sensor: two-component regulatory control in chloroplasts', *Photosynthesis Research*. Springer Netherlands, 130(1–3), pp. 93–101. doi: 10.1007/s11120-016-0229-y.
- Innan, H. and Kondrashov, F. (2010) 'The evolution of gene duplications: Classifying and distinguishing between models', *Nature Reviews Genetics*, pp. 97–108. doi: 10.1038/nrg2689.
- Iyer, L. M., Koonin, E. V. and Aravind, L. (2004) 'Evolution of bacterial RNA polymerase:

- Implications for large-scale bacterial phylogeny, domain accretion, and horizontal gene transfer', *Gene*, 335(1–2), pp. 73–88. doi: 10.1016/j.gene.2004.03.017.
- Jamithireddy, A. K., Runthala, A. and Gopal, B. (2019) 'Evaluation of specificity determinants in Mycobacterium tuberculosis  $\sigma$ /anti- $\sigma$  factor interactions', *Biochemical and Biophysical Research Communications*. doi: 10.1016/j.bbrc.2019.10.198.
- Janczarek, M. *et al.* (2018) 'Hanks-Type Serine/Threonine Protein Kinases and Phosphatases in Bacteria: Roles in Signaling and Adaptation to Various Environments.', *International journal of molecular sciences*, 19(10). doi: 10.3390/ijms19102872.
- Jishage, M. *et al.* (2002) 'Regulation of  $\sigma$  factor competition by the alarmone ppGpp', *Genes and Development*, 16(10), pp. 1260–1270. doi: 10.1101/gad.227902.
- Jishage, M., Dasgupta, D. and Ishihama, A. (2001) 'Mapping of the Rsd contact site on the sigma 70 subunit of Escherichia coli RNA polymerase', *Journal of Bacteriology*, 183(9), pp. 2952–2956. doi: 10.1128/JB.183.9.2952-2956.2001.
- Jogler, C. *et al.* (2012) 'Identification of proteins likely to be involved in morphogenesis, cell division, and signal transduction in Planctomycetes by comparative genomics.', *Journal of bacteriology*. American Society for Microbiology Journals, 194(23), pp. 6419–30. doi: 10.1128/JB.01325-12.
- Johnson, D. S. *et al.* (2007) 'Genome-wide mapping of in vivo protein-DNA interactions', *Science*, 316(5830), pp. 1497–1502. doi: 10.1126/science.1141319.
- Jokerst, R. S. *et al.* (1989) 'Analysis of the gene encoding the largest subunit of RNA polymerase II in Drosophila.', *Molecular & general genetics : MGG*, 215(2), pp. 266–75. doi: 10.1007/bf00339727.
- Jones, E. *et al.* (2001) 'SciPy: Open source scientific tools for Python'. Available at: <http://www.scipy.org/>.
- Jonkers, I. and Lis, J. T. (2015) 'Getting up to speed with transcription elongation by RNA polymerase II', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 167–177. doi: 10.1038/nrm3953.
- de Juan, D., Pazos, F. and Valencia, A. (2013) 'Emerging methods in protein co-evolution.', *Nature reviews. Genetics*, 14(4), pp. 249–61. doi: 10.1038/nrg3414.
- Kadowaki, T. *et al.* (2016) 'A two-component system regulates gene expression of the type IX secretion component proteins via an ECF sigma factor', *Scientific Reports*. doi: 10.1038/srep23288.
- Kazmierczak, M. J., Wiedmann, M. and Boor, K. J. (2005) 'Alternative Sigma Factors and Their Roles in Bacterial Virulence', *Microbiology and Molecular Biology Reviews*. doi: 10.1128/mmb.69.4.527-543.2005.
- Kim, M. S. *et al.* (2009) 'Positive and negative feedback regulatory loops of thiol-oxidative stress response mediated by an unstable isoform of  $\sigma$ R in actinomycetes', *Molecular Microbiology*. doi: 10.1111/j.1365-2958.2009.06824.x.
- Kim, Y. *et al.* (1993) 'Crystal structure of a yeast TBP/TATA-box complex', *Nature*, 365(6446), pp. 512–520. doi: 10.1038/365512a0.
- Kingston, R. E. and Chamberlin, M. J. (1981) 'Pausing and attenuation of in vitro transcription in the rrnB operon of *E. coli*', *Cell*, 27(3 PART 2), pp. 523–531. doi: 10.1016/0092-8674(81)90394-9.
- Kishino, H., Miyata, T. and Hasegawa, M. (1990) 'Maximum likelihood inference of protein phylogeny and the origin of chloroplasts', *Journal of Molecular Evolution*. Springer-Verlag, 31(2), pp. 151–160. doi: 10.1007/BF02109483.
- Koonin, E. V., Makarova, K. S. and Aravind, L. (2001) 'Horizontal Gene Transfer in Prokaryotes: Quantification and Classification', *Annual Review of Microbiology*, 55(1), pp. 709–742. doi: 10.1146/annurev.micro.55.1.709.

- Kostrewa, D. *et al.* (2009) 'RNA polymerase II-TFIIB structure and mechanism of transcription initiation', *Nature*, 462(7271), pp. 323–330. doi: 10.1038/nature08548.
- Kumar, A. *et al.* (1993) 'The minus 35-recognition region of Escherichia coli sigma 70 is inessential for initiation of transcription at an "extended minus 10" promoter', *Journal of Molecular Biology*. Academic Press, 232(2), pp. 406–418. doi: 10.1006/jmbi.1993.1400.
- Kuo, C.-H. and Ochman, H. (2010) 'The Extinction Dynamics of Bacterial Pseudogenes', *PLoS Genetics*. Edited by J. Zhang, 6(8), p. e1001050. doi: 10.1371/journal.pgen.1001050.
- Lambert, L. J. *et al.* (2004) 'T4 AsiA blocks DNA recognition by remodeling sigma70 region 4.', *The EMBO journal*, 23(15), pp. 2952–62. doi: 10.1038/sj.emboj.7600312.
- Lan, L. *et al.* (2006) 'Genome-wide gene expression analysis of Pseudomonas syringae pv. tomato DC3000 reveals overlapping and distinct pathways regulated by hrpL and hrpRS.', *Molecular plant-microbe interactions : MPMI*, 19(9), pp. 976–87. doi: 10.1094/MPMI-19-0976.
- Lane, W. J. and Darst, S. A. (2006) 'The structural basis for promoter -35 element recognition by the group IV sigma factors.', *PLoS biology*, 4(9), p. e269. doi: 10.1371/journal.pbio.0040269.
- Lane, W. J. and Darst, S. A. (2010a) 'Molecular Evolution of Multisubunit RNA Polymerases: Sequence Analysis', *Journal of Molecular Biology*, 395(4), pp. 671–685. doi: 10.1016/j.jmb.2009.10.062.
- Lane, W. J. and Darst, S. A. (2010b) 'Molecular Evolution of Multisubunit RNA Polymerases: Structural Analysis', *Journal of Molecular Biology*. doi: 10.1016/j.jmb.2009.10.063.
- Lang, C. *et al.* (2018) 'Most Sinorhizobium meliloti Extracytoplasmic Function Sigma Factors Control Accessory Functions', *mSphere*. American Society for Microbiology, 3(5). doi: 10.1128/mspheredirect.00454-18.
- Lapointe, F.-J. (1998) 'How to validate phylogenetic trees? A stepwise procedure', in, pp. 71–88. doi: 10.1007/978-4-431-65950-1\_6.
- Lawson, M. R. *et al.* (2018) 'Mechanism for the Regulated Control of Bacterial Transcription Termination by a Universal Adaptor Protein', *Molecular Cell*. Cell Press, 71(6), pp. 911–922.e4. doi: 10.1016/j.molcel.2018.07.014.
- Lee, J. and Borukhov, S. (2016) 'Bacterial RNA polymerase-DNA interaction-The driving force of gene expression and the target for drug action', *Frontiers in Molecular Biosciences*. Frontiers Media S.A. doi: 10.3389/fmolb.2016.00073.
- Lefebvre-Legendre, L. *et al.* (2014) 'On the Complexity of Chloroplast RNA Metabolism: psaA Trans-splicing Can be Bypassed in Chlamydomonas', *Molecular Biology and Evolution*, 31(10), pp. 2697–2707. doi: 10.1093/molbev/msu215.
- Leibman, M. and Hochschild, A. (2007) 'A  $\sigma$ -core interaction of the RNA polymerase holoenzyme that enhances promoter escape', *EMBO Journal*, 26(6), pp. 1579–1590. doi: 10.1038/sj.emboj.7601612.
- Letchumanan, V., Chan, K. G. and Lee, L. H. (2014) 'Vibrio parahaemolyticus: A review on the pathogenesis, prevalence, and advance molecular identification techniques', *Frontiers in Microbiology*. Frontiers Media S.A. doi: 10.3389/fmicb.2014.00705.
- Letunic, I. and Bork, P. (2016) 'Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees', *Nucleic acids research*. doi: 10.1093/nar/gkw290.
- Lewerke, L. T. *et al.* (2018) 'Bacterial sensing: A putative amphipathic helix in RsiV is the switch for activating  $\sigma^V$  in response to lysozyme.', *PLoS genetics*, 14(7), p. e1007527. doi: 10.1371/journal.pgen.1007527.
- Li, L. *et al.* (2019) 'Structural basis for transcription initiation by bacterial ECF  $\sigma$  factors', *Nature Communications*. doi: 10.1038/s41467-019-09096-y.

- Li, P. *et al.* (2017) 'Acute hepatopancreatic necrosis disease-causing *Vibrio parahaemolyticus* strains maintain an antibacterial type VI secretion system with versatile effector repertoires', *Applied and Environmental Microbiology*. American Society for Microbiology, 83(13). doi: 10.1128/AEM.00737-17.
- Li, S. *et al.* (2019) 'Structural basis for the recognition of MucA by MucB and AlgU in *Pseudomonas aeruginosa*', *The FEBS Journal*, p. febs.14995. doi: 10.1111/febs.14995.
- Li, W. *et al.* (2002) 'Identification and structure of the anti-sigma factor-binding domain of the disulphide-stress regulated sigma factor  $\sigma$ R from *Streptomyces coelicolor*', *Journal of Molecular Biology*. doi: 10.1016/S0022-2836(02)00948-8.
- Li, W. *et al.* (2003) 'The role of zinc in the disulphide stress-regulated anti-sigma factor RsrA from *Streptomyces coelicolor*', *Journal of Molecular Biology*. Academic Press, 333(2), pp. 461–472. doi: 10.1016/j.jmb.2003.08.038.
- Li, W. and Godzik, A. (2006) 'Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences.', *Bioinformatics (Oxford, England)*, 22(13), pp. 1658–9. doi: 10.1093/bioinformatics/btl158.
- Li, X. *et al.* (2009) 'Cleavage of RseA by RseP requires a carboxyl-terminal hydrophobic amino acid following DegS cleavage', *Proceedings of the National Academy of Sciences of the United States of America*, 106(35), pp. 14837–14842. doi: 10.1073/pnas.0903289106.
- Lima, S. *et al.* (2013) 'Dual molecular signals mediate the bacterial response to outer-membrane stress', *Science*. American Association for the Advancement of Science, 340(6134), pp. 837–841. doi: 10.1126/science.1235358.
- Liu, B. *et al.* (2017) 'Structural basis of bacterial transcription activation.', *Science (New York, N.Y.)*, 358(6365), pp. 947–951. doi: 10.1126/science.aao1923.
- Liu, Q., Pinto, D. and Mascher, T. (2018) 'Characterization of the Widely Distributed Novel ECF42 Group of Extracytoplasmic Function  $\sigma$  Factors in *Streptomyces venezuelae*', *Journal of Bacteriology*. doi: 10.1128/jb.00437-18.
- Liu, X. *et al.* (2010) 'Structure of an RNA polymerase II-TFIIB complex and the transcription initiation mechanism.', *Science (New York, N.Y.)*, 327(5962), pp. 206–9. doi: 10.1126/science.1182015.
- Liu, X., Brutlag, D. L. and Liu, J. S. (2001) 'BioProspector: discovering conserved DNA motifs in upstream regulatory regions of co-expressed genes.', *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*.
- Llamas, M. A. *et al.* (2009) 'A novel extracytoplasmic function (ECF) sigma factor regulates virulence in *Pseudomonas aeruginosa*', *PLoS Pathogens*. doi: 10.1371/journal.ppat.1000572.
- Lonetto, M. A. *et al.* (1994) 'Analysis of the *Streptomyces coelicolor* sigE gene reveals the existence of a subfamily of eubacterial RNA polymerase sigma factors involved in the regulation of extracytoplasmic functions.', *Proceedings of the National Academy of Sciences*, 91(16), pp. 7573–7577. doi: 10.1073/pnas.91.16.7573.
- Lonetto, M., Gribskov, M. and Gross, C. A. (1992) 'The sigma 70 family: sequence conservation and evolutionary relationships.', *Journal of bacteriology*, 174(12), pp. 3843–9. doi: 10.1128/jb.174.12.3843-3849.1992.
- Lourenço, R. F., Kohler, C. and Gomes, S. L. (2011) 'A two-component system, an anti-sigma factor and two paralogous ECF sigma factors are involved in the control of general stress response in *Caulobacter crescentus*', *Molecular Microbiology*, 80(6), pp. 1598–1612. doi: 10.1111/j.1365-2958.2011.07668.x.
- Luo, S. *et al.* (2014) 'An extracytoplasmic function sigma factor,  $\sigma$ (25), differentially regulates avermectin and oligomycin biosynthesis in *Streptomyces avermitilis*.', *Applied microbiology and biotechnology*, 98(16), pp. 7097–112. doi: 10.1007/s00253-014-5759-7.

- Luo, Y. *et al.* (2010) 'Transcriptomic and phenotypic characterization of a *Bacillus subtilis* strain without extracytoplasmic function  $\sigma$  factors.', *Journal of bacteriology*, 192(21), pp. 5736–45. doi: 10.1128/JB.00826-10.
- Luse, D. S. (2013) 'Promoter clearance by RNA polymerase II', *Biochimica et Biophysica Acta - Gene Regulatory Mechanisms*, pp. 63–68. doi: 10.1016/j.bbagr.2012.08.010.
- Madeira, F. *et al.* (2019) 'The EMBL-EBI search and sequence analysis tools APIs in 2019.', *Nucleic acids research*, 47(W1), pp. W636–W641. doi: 10.1093/nar/gkz268.
- Mahren, S. and Braun, V. (2003) 'The *FecI* extracytoplasmic-function sigma factor of *Escherichia coli* interacts with the  $\beta'$  subunit of RNA polymerase', *Journal of Bacteriology*, 185(6), pp. 1796–1802. doi: 10.1128/JB.185.6.1796-1802.2003.
- Maier, U. G. *et al.* (2008) 'Complex chloroplast RNA metabolism: Just debugging the genetic programme?', *BMC Biology*, 6. doi: 10.1186/1741-7007-6-36.
- Maillard, A. P. *et al.* (2014) 'The crystal structure of the anti- $\sigma$  factor CnrY in complex with the  $\sigma$  factor CnrH shows a new structural class of anti- $\sigma$  factors targeting extracytoplasmic function  $\sigma$  factors.', *Journal of molecular biology*, 426(12), pp. 2313–27. doi: 10.1016/j.jmb.2014.04.003.
- Makino, K. *et al.* (2003) 'Genome sequence of *Vibrio parahaemolyticus*: a pathogenic mechanism distinct from that of *V. cholerae*.', *Lancet (London, England)*, 361(9359), pp. 743–9. doi: 10.1016/S0140-6736(03)12659-1.
- Malik, S., Zalenskaya, K. and Goldfarb, A. (1987) 'Competition between sigma factors for core RNA polymerase', *Nucleic Acids Research*, 15(20), pp. 8521–8530. doi: 10.1093/nar/15.20.8521.
- Marcos-Torres, F. J. *et al.* (2016) 'In depth analysis of the mechanism of action of metal-dependent sigma factors: Characterization of CorE2 from *Myxococcus xanthus*', *Nucleic Acids Research*. doi: 10.1093/nar/gkw150.
- Martens, E. C., Koropatkin, N. M., *et al.* (2009) 'Complex glycan catabolism by the human gut microbiota: The bacteroidetes sus-like paradigm', *Journal of Biological Chemistry*, pp. 24673–24677. doi: 10.1074/jbc.R109.022848.
- Martens, E. C., Roth, R., *et al.* (2009) 'Coordinate regulation of glycan degradation and polysaccharide capsule biosynthesis by a prominent human gut symbiont', *Journal of Biological Chemistry*, 284(27), pp. 18445–18457. doi: 10.1074/jbc.M109.008094.
- Martínez-Antonio, A. and Collado-Vides, J. (2003) 'Identifying global regulators in transcriptional regulatory networks in bacteria', *Current Opinion in Microbiology*. Elsevier Ltd, pp. 482–489. doi: 10.1016/j.mib.2003.09.002.
- Mascher, T. (2013) 'Signaling diversity and evolution of extracytoplasmic function (ECF)  $\sigma$  factors', *Current Opinion in Microbiology*, 16(2), pp. 148–155. doi: 10.1016/j.mib.2013.02.001.
- Mascher, T., Hachmann, A. B. and Helmann, J. D. (2007) 'Regulatory overlap and functional redundancy among *Bacillus subtilis* extracytoplasmic function  $\sigma$  factors', *Journal of Bacteriology*. doi: 10.1128/JB.00904-07.
- Mauri, M. and Klumpp, S. (2014) 'A Model for Sigma Factor Competition in Bacterial Cells', *PLoS Computational Biology*. Public Library of Science, 10(10). doi: 10.1371/journal.pcbi.1003845.
- McAllister, W. T. (1993) 'Structure and function of the bacteriophage T7 RNA polymerase (or, the virtues of simplicity).', *Cellular & molecular biology research*, 39(4), pp. 385–91. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/8312975> (Accessed: 6 November 2019).
- Mecasas, J. *et al.* (1993) 'The activity of  $\sigma(E)$ , an *Escherichia coli* heat-inducible  $\sigma$ -factor, is modulated by expression of outer membrane proteins', *Genes and Development*. Cold Spring Harbor Laboratory Press, 7(12 B), pp. 2618–2628. doi: 10.1101/gad.7.12b.2618.
- Mereschowsky, C. (1905) 'Über Natur und Ursprung der Chromatophoren im Pflanzenreiche', *Biologisches Centralblatt*, (25), pp. 593–604.

- Merighi, M. *et al.* (2003) 'The HrpX/HrpY two-component system activates hrpS expression, the first step in the regulatory cascade controlling the Hrp regulon in *Pantoea stewartii* subsp. *stewartii*.', *Molecular plant-microbe interactions : MPMI*, 16(3), pp. 238–48. doi: 10.1094/MPMI.2003.16.3.238.
- Merrick, M. J. (1993) 'In a class of its own? the RNA polymerase sigma factor  $\sigma^{54}$  ( $\sigma^N$ )', *Molecular Microbiology*, 10(5), pp. 903–909. doi: 10.1111/j.1365-2958.1993.tb00961.x.
- Metzger, W. *et al.* (1993) 'Nucleation of RNA chain formation by *Escherichia coli* DNA-dependent RNA polymerase', *Journal of Molecular Biology*, 232(1), pp. 35–49. doi: 10.1006/jmbi.1993.1368.
- Mijakovic, I., Grangeasse, C. and Turgay, K. (2016) 'Exploring the diversity of protein modifications: Special bacterial phosphorylation systems', *FEMS Microbiology Reviews*. Oxford University Press, pp. 398–417. doi: 10.1093/femsre/fuw003.
- Minakhin, L. *et al.* (2001) 'Bacterial RNA polymerase subunit  $\omega$  and eukaryotic polymerase subunit RPB6 are sequence, structural, and functional homologs and promote RNA polymerase assembly', *Proceedings of the National Academy of Sciences of the United States of America*, 98(3), pp. 892–897. doi: 10.1073/pnas.98.3.892.
- Missiakas, D. *et al.* (1997) 'Modulation of the *Escherichia coli* sigmaE (RpoE) heat-shock transcription-factor activity by the RseA, RseB and RseC proteins.', *Molecular microbiology*, 24(2), pp. 355–71. doi: 10.1046/j.1365-2958.1997.3601713.x.
- Mitchell, A. *et al.* (2015) 'The InterPro protein families database: the classification resource after 15 years.', *Nucleic acids research*, 43(Database issue), pp. D213–21. doi: 10.1093/nar/gku1243.
- Mougous, J. D. *et al.* (2007) 'Threonine phosphorylation post-translationally regulates protein secretion in *Pseudomonas aeruginosa*', *Nature Cell Biology*, 9(7), pp. 797–803. doi: 10.1038/ncb1605.
- Mount, S. M. and Chang, C. (2002) 'Evidence for a plastid origin of plant ethylene receptor genes', *Plant Physiology*, 130(1), pp. 10–14. doi: 10.1104/pp.005397.
- Müller, M. *et al.* (2014) 'Deletion of membrane-associated Asp23 leads to upregulation of cell wall stress genes in *Staphylococcus aureus*', *Molecular Microbiology*. doi: 10.1111/mmi.12733.
- Mullis, K. B. (1990) 'The Unusual Origin of the Polymerase Chain Reaction', *Scientific American*. Scientific American, a division of Nature America, Inc., pp. 56–65. doi: 10.2307/24996713.
- Münch, R. *et al.* (2005) 'Virtual Footprint and PRODORIC: An integrative framework for regulon prediction in prokaryotes', *Bioinformatics*, 21(22), pp. 4187–4189. doi: 10.1093/bioinformatics/bti635.
- Murakami, K. S. (2013) 'X-ray crystal structure of *Escherichia coli* RNA polymerase  $\sigma^{70}$  holoenzyme', *Journal of Biological Chemistry*, 288(13), pp. 9126–9134. doi: 10.1074/jbc.M112.430900.
- Murakami, K. S. (2015) 'Structural biology of bacterial RNA polymerase', *Biomolecules*. MDPI AG, pp. 848–864. doi: 10.3390/biom5020848.
- Murakami, K. S., Masuda, S. and Darst, S. A. (2002) 'Structural basis of transcription initiation: RNA polymerase holoenzyme at 4 Å resolution', *Science*, 296(5571), pp. 1280–1284. doi: 10.1126/science.1069594.
- Nagai, H., Yuzawa, H. and Yura, T. (1991) 'Interplay of two cis-acting mRNA regions in translational control of  $\sigma^{32}$  synthesis during the heat shock response of *Escherichia coli*', *Proceedings of the National Academy of Sciences of the United States of America*. National Academy of Sciences, 88(23), pp. 10515–10519. doi: 10.1073/pnas.88.23.10515.
- Naji, S., Grünberg, S. and Thomm, M. (2007) 'The RPB7 orthologue E' is required for transcriptional activity of a reconstituted archaeal core enzyme at low temperatures and stimulates open complex formation', *Journal of Biological Chemistry*, 282(15), pp. 11047–11057. doi: 10.1074/jbc.M611674200.

- Narayanan, A. *et al.* (2018) 'Cryo-EM structure of Escherichia coli 70 RNA polymerase and promoter DNA complex revealed a role of non-conserved region during the open complex formation', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 293(19), pp. 7367–7375. doi: 10.1074/jbc.RA118.002161.
- Nataf, Y. *et al.* (2010) 'Clostridium thermocellum cellulosomal genes are regulated by extracytoplasmic polysaccharides via alternative sigma factors', *Proceedings of the National Academy of Sciences of the United States of America*, 107(43), pp. 18646–18651. doi: 10.1073/pnas.1012175107.
- Nechaev, S. and Severinov, K. (2003) 'Bacteriophage-induced modifications of host RNA polymerase.', *Annual review of microbiology*, 57, pp. 301–22. doi: 10.1146/annurev.micro.57.030502.090942.
- Neidhardt, F. C., Bloch, P. L. and Smith, D. F. (1974) 'Culture medium for enterobacteria', *Journal of Bacteriology*, 119(3), pp. 736–747.
- Nguyen, L.-T. *et al.* (2015) 'IQ-TREE: a fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies.', *Molecular biology and evolution*, 32(1), pp. 268–74. doi: 10.1093/molbev/msu300.
- Nguyen, M. *et al.* (2015) 'HGT-finder: A new tool for horizontal gene transfer finding and application to Aspergillus genomes', *Toxins*. MDPI AG, 7(10), pp. 4035–4053. doi: 10.3390/toxins7104035.
- Nguyen, N. P. D. *et al.* (2015) 'Ultra-large alignments using phylogeny-aware profiles', *Genome Biology*. doi: 10.1186/s13059-015-0688-z.
- Nicolas, P. *et al.* (2012) 'Condition-dependent transcriptome reveals high-level regulatory architecture in Bacillus subtilis', *Science*. American Association for the Advancement of Science, 335(6072), pp. 1103–1106. doi: 10.1126/science.1206848.
- Nizan-Koren, R. *et al.* (2003) 'The Regulatory Cascade That Activates the Hrp Regulon in Erwinia herbicola pv. gypsophilae', *Molecular Plant-Microbe Interactions*. doi: 10.1094/mpmi.2003.16.3.249.
- Ochs, M. *et al.* (1996) 'Surface signaling in transcriptional regulation of the ferric citrate transport system of Escherichia coli: mutational analysis of the alternative sigma factor Fecl supports its essential role in fec transport gene transcription', *MGG Molecular & General Genetics*. Springer Science and Business Media LLC, 250(4), pp. 455–465. doi: 10.1007/bf02174034.
- Olechnovič, K. and Venclovas, C. (2014) 'Voronota: A fast and reliable tool for computing the vertices of the Voronoi diagram of atomic balls.', *Journal of computational chemistry*, 35(8), pp. 672–81. doi: 10.1002/jcc.23538.
- Orphanides, G. and Reinberg, D. (2002) 'A unified theory of gene expression', *Cell*. Cell Press, pp. 439–451. doi: 10.1016/S0092-8674(02)00655-4.
- Orsini, G. *et al.* (1993) 'The asiA gene of bacteriophage T4 codes for the anti- $\sigma$ 70 protein', *Journal of Bacteriology*, 175(1), pp. 85–93. doi: 10.1128/jb.175.1.85-93.1993.
- Ortiz de Ora, L. *et al.* (2018) 'Regulation of biomass degradation by alternative  $\sigma$  factors in cellulolytic clostridia', *Scientific Reports*. doi: 10.1038/s41598-018-29245-5.
- Ouhammouch, M. *et al.* (1995) 'Bacteriophage T4 MotA and AsiA proteins suffice to direct Escherichia coli RNA polymerase to initiate transcription at T4 middle promoters', *Proceedings of the National Academy of Sciences of the United States of America*, 92(5), pp. 1451–1455. doi: 10.1073/pnas.92.5.1451.
- Paget, M. S. (2015) 'Bacterial Sigma Factors and Anti-Sigma Factors: Structure, Function and Distribution.', *Biomolecules*. Multidisciplinary Digital Publishing Institute (MDPI), 5(3), pp. 1245–65. doi: 10.3390/biom5031245.
- Paget, M. S. B. and Helmann, J. D. (2003) 'The  $\sigma$ 70 family of sigma factors', *Genome Biology*. doi: 10.1186/gb-2003-4-1-203.

- Pal, M. and Luse, D. S. (2003) 'The initiation-elongation transition: Lateral mobility of RNA in RNA polymerase II complexes is greatly reduced at +8/+9 and absent by +23', *Proceedings of the National Academy of Sciences of the United States of America*, 100(10), pp. 5700–5705. doi: 10.1073/pnas.1037057100.
- Pal, M., McKean, D. and Luse, D. S. (2001) 'Promoter Clearance by RNA Polymerase II Is an Extended, Multistep Process Strongly Affected by Sequence', *Molecular and Cellular Biology*. American Society for Microbiology, 21(17), pp. 5815–5825. doi: 10.1128/mcb.21.17.5815-5825.2001.
- Pal, M., Ponticelli, A. S. and Luse, D. S. (2005) 'The role of the transcription bubble and TFIIB in promoter clearance by RNA polymerase II', *Molecular Cell*, 19(1), pp. 101–110. doi: 10.1016/j.molcel.2005.05.024.
- Park, J. S., Marr, M. T. and Roberts, J. W. (2002) '*E. coli* transcription repair coupling factor (Mfd protein) rescues arrested complexes by promoting forward translocation', *Cell*. Cell Press, 109(6), pp. 757–767. doi: 10.1016/S0092-8674(02)00769-9.
- Parkinson, J. S. and Kofoid, E. C. (1992) 'Communication Modules in Bacterial Signaling Proteins', *Annual Review of Genetics*. Annual Reviews, 26(1), pp. 71–112. doi: 10.1146/annurev.ge.26.120192.000443.
- Patikoglou, G. A. *et al.* (2007) 'Crystal structure of the Escherichia coli regulator of sigma70, Rsd, in complex with sigma70 domain 4.', *Journal of molecular biology*, 372(3), pp. 649–59. doi: 10.1016/j.jmb.2007.06.081.
- Pedregosa, F. *et al.* (2011) 'Scikit-learn: Machine Learning in {P}ython', *Journal of Machine Learning Research*, 12, pp. 2825–2830.
- Pedregosa, F. *et al.* (2012) 'Scikit-learn: Machine Learning in Python'. Available at: <http://arxiv.org/abs/1201.0490> (Accessed: 5 July 2019).
- Peng, Q. and Yuan, Y. (2018) 'Characterization of a novel phage infecting the pathogenic multidrug-resistant *Bacillus cereus* and functional analysis of its endolysin.', *Applied microbiology and biotechnology*, 102(18), pp. 7901–7912. doi: 10.1007/s00253-018-9219-7.
- Pereira, S. F. F., Goss, L. and Dworkin, J. (2011) 'Eukaryote-Like Serine/Threonine Kinases and Phosphatases in Bacteria', *Microbiology and Molecular Biology Reviews*. American Society for Microbiology, 75(1), pp. 192–212. doi: 10.1128/mnbr.00042-10.
- Pérez, J., Muñoz-Dorado, J. and Moraleda-Muñoz, A. (2018) 'The complex global response to copper in the multicellular bacterium *Myxococcus xanthus*.', *Metallomics: integrated biometal science*, 10(7), pp. 876–886. doi: 10.1039/c8mt00121a.
- Pesole, G. (2008) 'What is a gene? An updated operational definition', *Gene*, pp. 1–4. doi: 10.1016/j.gene.2008.03.010.
- Peters, J. M., Vangeloff, A. D. and Landick, R. (2011) 'Bacterial transcription terminators: The RNA 3'-end chronicles', *Journal of Molecular Biology*, pp. 793–813. doi: 10.1016/j.jmb.2011.03.036.
- Pettersen, E. F. *et al.* (2004) 'UCSF Chimera - A visualization system for exploratory research and analysis', *Journal of Computational Chemistry*. doi: 10.1002/jcc.20084.
- Pinto, D. *et al.* (2018) 'Engineering orthogonal synthetic timer circuits based on extracytoplasmic function  $\sigma$  factors', *Nucleic Acids Research*, 46(14), pp. 7450–7464. doi: 10.1093/nar/gky614.
- Pinto, D. *et al.* (2019) 'Extracytoplasmic Function  $\sigma$  Factors Can Be Implemented as Robust Heterologous Genetic Switches in *Bacillus subtilis*', *iScience*. Elsevier Inc., 13, pp. 380–390. doi: 10.1016/j.isci.2019.03.001.
- Pinto, D. and da Fonseca, R. R. (2020) 'Evolution of the extracytoplasmic function  $\sigma$  factor protein family', *NAR Genomics and Bioinformatics*, 2(1). doi: 10.1093/nargab/lqz026.
- Pinto, D., Liu, Q. and Mascher, T. (2019) 'ECF  $\sigma$  factors with regulatory extensions: The one-

- component systems of the  $\sigma$  universe', *Molecular Microbiology*. doi: 10.1111/mmi.14323.
- Pinto, D. and Mascher, T. (2016) 'The ECF Classification: A Phylogenetic Reflection of the Regulatory Diversity in the Extracytoplasmic Function  $\sigma$  Factor Protein Family', in *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*. doi: 10.1002/9781119004813.ch7.
- Politi, N. *et al.* (2014) 'Half-life measurements of chemical inducers for recombinant gene expression', *Journal of Biological Engineering*, 8(1). doi: 10.1186/1754-1611-8-5.
- Pupov, D. *et al.* (2014) 'Distinct functions of the RNA polymerase  $\sigma$  subunit region 3.2 in RNA priming and promoter escape', *Nucleic Acids Research*. Oxford University Press, 42(7), pp. 4494–4504. doi: 10.1093/nar/gkt1384.
- Puthiyaveetil, S. *et al.* (2008) 'The ancestral symbiont sensor kinase CSK links photosynthesis with gene expression in chloroplasts', *Proceedings of the National Academy of Sciences of the United States of America*, 105(29), pp. 10061–10066. doi: 10.1073/pnas.0803928105.
- Puthiyaveetil, S. and Allen, J. F. (2009) 'Chloroplast two-component systems: Evolution of the link between photosynthesis and gene expression', *Proceedings of the Royal Society B: Biological Sciences*. Royal Society, pp. 2133–2145. doi: 10.1098/rspb.2008.1426.
- Puthiyaveetil, S., Ibrahim, I. M. and Allen, J. F. (2013) 'Evolutionary rewiring: A modified prokaryotic gene-regulatory pathway in chloroplasts', *Philosophical Transactions of the Royal Society B: Biological Sciences*, 368(1622). doi: 10.1098/rstb.2012.0260.
- Quesada, J. M. *et al.* (2016) 'The activity of the pseudomonas aeruginosa virulence regulator  $\sigma$ VreI is modulated by the anti- $\sigma$  factor VreR and the transcription factor PhoB', *Frontiers in Microbiology*. Frontiers Media S.A., 7(AUG). doi: 10.3389/fmicb.2016.01159.
- Rajasekar, K. V. *et al.* (2016) 'The anti-sigma factor RsrA responds to oxidative stress by reburying its hydrophobic core', *Nature Communications*. Nature Publishing Group, 7. doi: 10.1038/ncomms12194.
- Rausell, A. *et al.* (2010) 'Protein interactions and ligand binding: from protein subfamilies to functional specificity.', *Proceedings of the National Academy of Sciences of the United States of America*, 107(5), pp. 1995–2000. doi: 10.1073/pnas.0908044107.
- Remmert, M. *et al.* (2012) 'HHblits: Lightning-fast iterative protein sequence searching by HMM-HMM alignment', *Nature Methods*, 9(2), pp. 173–175. doi: 10.1038/nmeth.1818.
- Rhee, H. S. and Pugh, B. F. (2012) 'Genome-wide structure and organization of eukaryotic pre-initiation complexes', *Nature*. Nature Publishing Group, 483(7389), pp. 295–301. doi: 10.1038/nature10799.
- Rhodium, V. A. *et al.* (2013) 'Design of orthogonal genetic switches based on a crosstalk map of  $\sigma$ s, anti- $\sigma$ s, and promoters', *Molecular Systems Biology*. doi: 10.1038/msb.2013.58.
- Rhodium, V. A. and Mutalik, V. K. (2010) 'Predicting strength and function for promoters of the Escherichia coli alternative sigma factor, sigmaE.', *Proceedings of the National Academy of Sciences of the United States of America*, 107(7), pp. 2854–9. doi: 10.1073/pnas.0915066107.
- Roeder, R. G. and Rutter, W. J. (1969) 'Multiple forms of DNA-dependent RNA polymerase in eukaryotic organisms', *Nature*, 224(5216), pp. 234–237. doi: 10.1038/224234a0.
- Rombel, I. *et al.* (1998) 'The bacterial enhancer-binding protein NtrC as a molecular machine', in *Cold Spring Harbor Symposia on Quantitative Biology*. Cold Spring Harbor Laboratory Press, pp. 157–166. doi: 10.1101/sqb.1998.63.157.
- Rousseeuw, P. J. (1987) 'Silhouettes: A graphical aid to the interpretation and validation of cluster analysis', *Journal of Computational and Applied Mathematics*, 20(C), pp. 53–65. doi: 10.1016/0377-0427(87)90125-7.
- Runner, V. M., Podolny, V. and Buratowski, S. (2008) 'The Rpb4 Subunit of RNA Polymerase II

- Contributes to Cotranscriptional Recruitment of 3' Processing Factors', *Molecular and Cellular Biology*. American Society for Microbiology, 28(6), pp. 1883–1891. doi: 10.1128/mcb.01714-07.
- Saba, J. *et al.* (2019) 'The elemental mechanism of transcriptional pausing', *eLife*. NLM (Medline), 8. doi: 10.7554/eLife.40981.
- Sainsbury, S., Bernecky, C. and Cramer, P. (2015) 'Structural basis of transcription initiation by RNA polymerase II', *Nature Reviews Molecular Cell Biology*. Nature Publishing Group, pp. 129–143. doi: 10.1038/nrm3952.
- Sainsbury, S., Niesser, J. and Cramer, P. (2013) 'Structure and function of the initially transcribing RNA polymerase II-TFIIB complex', *Nature*, 493(7432), pp. 437–440. doi: 10.1038/nature11715.
- Salah Ud-Din, A. I. M. and Roujeinikova, A. (2017) 'Methyl-accepting chemotaxis proteins: a core sensing element in prokaryotes and archaea', *Cellular and Molecular Life Sciences*. Birkhauser Verlag AG, pp. 3293–3303. doi: 10.1007/s00018-017-2514-0.
- Salgado, P. S. *et al.* (2006) 'The structure of an RNAi polymerase links RNA silencing and transcription', *PLoS Biology*, 4(12), pp. 2274–2281. doi: 10.1371/journal.pbio.0040434.
- Salomon, D. *et al.* (2013) 'Vibrio parahaemolyticus Type VI Secretion System 1 Is Activated in Marine Conditions to Target Bacteria, and Is Differentially Regulated from System 2', *PLoS ONE*, 8(4). doi: 10.1371/journal.pone.0061086.
- Sambrook, J. and Russell, D. W. (2006) 'The Inoue Method for Preparation and Transformation of Competent *E. coli*: "Ultra-Competent" Cells', *Cold Spring Harbor Protocols*. Cold Spring Harbor Laboratory, 2006(1), p. pdb.prot3944. doi: 10.1101/pdb.prot3944.
- Sang, T. P., Kang, C. M. and Husson, R. N. (2008) 'Regulation of the SigH stress response regulon by an essential protein kinase in Mycobacterium tuberculosis', *Proceedings of the National Academy of Sciences of the United States of America*, 105(35), pp. 13105–13110. doi: 10.1073/pnas.0801143105.
- Schindler, D. *et al.* (2016) 'Design and Assembly of DNA Sequence Libraries for Chromosomal Insertion in Bacteria Based on a Set of Modified MoClo Vectors', *ACS Synthetic Biology*. American Chemical Society, 5(12), pp. 1362–1368. doi: 10.1021/acssynbio.6b00089.
- Schneider, G. J. and Hasekorn, R. (1988) 'RNA polymerase subunit homology among cyanobacteria, other eubacteria and archaeobacteria.', *Journal of bacteriology*, 170(9), pp. 4136–4140. doi: 10.1128/jb.170.9.4136-4140.1988.
- Schultz, S. G. and Solomon, A. K. (1961) 'Cation transport in Escherichia coli. I. Intracellular Na and K concentrations and net cation movement.', *The Journal of general physiology*, 45, pp. 355–369. doi: 10.1085/jgp.45.2.355.
- Schulz, S. *et al.* (2015) 'Elucidation of Sigma Factor-Associated Networks in Pseudomonas aeruginosa Reveals a Modular Architecture with Limited and Function-Specific Crosstalk', *PLOS Pathogens*. Edited by V. T. Lee, 11(3), p. e1004744. doi: 10.1371/journal.ppat.1004744.
- Schumacher, M. A. *et al.* (2018) 'The crystal structure of the RsbN- $\sigma$ BldN complex from Streptomyces venezuelae defines a new structural class of anti- $\sigma$  factor.', *Nucleic acids research*, 46(14), pp. 7405–7417. doi: 10.1093/nar/gky493.
- Schweer, J., Türkeri, H., Link, B., *et al.* (2010) 'AtSIG6, a plastid sigma factor from Arabidopsis, reveals functional impact of cpCK2 phosphorylation', *Plant Journal*, 62(2), pp. 192–202. doi: 10.1111/j.1365-313X.2010.04138.x.
- Schweer, J., Türkeri, H., Kolpack, A., *et al.* (2010) 'Role and regulation of plastid sigma factors and their functional interactors during chloroplast transcription - recent lessons from Arabidopsis thaliana.', *European journal of cell biology*, 89(12), pp. 940–6. doi: 10.1016/j.ejcb.2010.06.016.
- Seipke, R. F., Patrick, E. and Hutchings, M. I. (2014) 'Regulation of antimycin biosynthesis by the orphan ECF RNA polymerase sigma factor  $\sigma$  AntA', *PeerJ*. doi: 10.7717/peerj.253.
- Sengupta, S., Prajapati, R. K. and Mukhopadhyay, J. (2015) 'Promoter escape with bacterial two-

- component  $\sigma$  factor suggests retention of  $\sigma$  region two in the elongation complex', *Journal of Biological Chemistry*. American Society for Biochemistry and Molecular Biology Inc., 290(47), pp. 28575–28583. doi: 10.1074/jbc.M115.666008.
- Shepard, W. *et al.* (2007) 'The crystal structure of Rv0813c from *Mycobacterium tuberculosis* reveals a new family of fatty acid-binding protein-like proteins in bacteria', *Journal of Bacteriology*, 189(5), pp. 1899–1904. doi: 10.1128/JB.01435-06.
- Shimizu, M. *et al.* (2010) 'Sigma factor phosphorylation in the photosynthetic control of photosystem stoichiometry', *Proceedings of the National Academy of Sciences of the United States of America*, 107(23), pp. 10760–10764. doi: 10.1073/pnas.0911692107.
- Shukla, J. *et al.* (2014) 'Structural basis for the redox sensitivity of the *Mycobacterium tuberculosis* SigK-RskA  $\sigma$ -anti- $\sigma$  complex.', *Acta crystallographica. Section D, Biological crystallography*, 70(Pt 4), pp. 1026–36. doi: 10.1107/S1399004714000121.
- Sievers, F. *et al.* (2011) 'Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega', *Molecular Systems Biology*. doi: 10.1038/msb.2011.75.
- Sievers, F. and Higgins, D. G. (2014) 'Clustal Omega, Accurate Alignment of Very Large Numbers of Sequences', in Russell, D. J. (ed.) *Multiple Sequence Alignment Methods*. Totowa, NJ: Humana Press, pp. 105–116. doi: 10.1007/978-1-62703-646-7\_6.
- Sigurdsson, S., Dirac-Svejstrup, A. B. and Svejstrup, J. Q. (2010) 'Evidence that Transcript Cleavage Is Essential for RNA Polymerase II Transcription and Cell Viability', *Molecular Cell*, 38(2), pp. 202–210. doi: 10.1016/j.molcel.2010.02.026.
- Šiková, M. *et al.* (2019) 'The torpedo effect in *Bacillus subtilis*:  $\sigma$  RNase J1 resolves stalled transcription complexes', *The EMBO Journal*. doi: 10.15252/embj.2019102500.
- Sineva, E., Savkina, M. and Ades, S. E. (2017) 'Themes and variations in gene regulation by extracytoplasmic function (ECF) sigma factors', *Current Opinion in Microbiology*. doi: 10.1016/j.mib.2017.05.004.
- Song, N., Sedgewick, R. D. and Durand, D. (2007) 'Domain architecture comparison for multidomain homology identification', in *Journal of Computational Biology*, pp. 496–516. doi: 10.1089/cmb.2007.A009.
- Sorenson, M. K., Ray, S. S. and Darst, S. A. (2004) 'Crystal structure of the flagellar  $\sigma$ /anti- $\sigma$  complex  $\sigma$ 28/FlgM reveals an intact  $\sigma$  factor in an inactive conformation', *Molecular Cell*, 14(1), pp. 127–138. doi: 10.1016/S1097-2765(04)00150-9.
- Stamatakis, A. (2014) 'RAxML version 8: A tool for phylogenetic analysis and post-analysis of large phylogenies', *Bioinformatics*. Oxford University Press, 30(9), pp. 1312–1313. doi: 10.1093/bioinformatics/btu033.
- Stancik, I. A. *et al.* (2018) 'Serine/Threonine Protein Kinases from Bacteria, Archaea and Eukarya Share a Common Evolutionary Origin Deeply Rooted in the Tree of Life.', *Journal of molecular biology*, 430(1), pp. 27–32. doi: 10.1016/j.jmb.2017.11.004.
- Staroń, A. *et al.* (2009) 'The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF)  $\sigma$  factor protein family', *Molecular Microbiology*, 74(3), pp. 557–581. doi: 10.1111/j.1365-2958.2009.06870.x.
- Staroń, A. and Mascher, T. (2010) 'General stress response in  $\alpha$ -proteobacteria: PhyR and beyond', *Molecular Microbiology*. doi: 10.1111/j.1365-2958.2010.07336.x.
- Steinegger, M. and Söding, J. (2017) 'MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets.', *Nature biotechnology*, 35(11), pp. 1026–1028. doi: 10.1038/nbt.3988.
- Stolzer, M. *et al.* (2015) 'Event inference in multidomain families with phylogenetic reconciliation', *BMC Bioinformatics*. BioMed Central Ltd., 16(14). doi: 10.1186/1471-2105-16-S14-S8.
- Studholme, D. J. and Buck, M. (2000) 'The biology of enhancer-dependent transcriptional regulation

- in bacteria: insights from genome sequences', *FEMS Microbiology Letters*, 186(1), pp. 1–9. doi: 10.1111/j.1574-6968.2000.tb09074.x.
- Sweetser, D., Nonet, M. and Young, R. A. (1987) 'Prokaryotic and eukaryotic RNA polymerases have homologous core subunits', *Proceedings of the National Academy of Sciences of the United States of America*, 84(5), pp. 1192–1196. doi: 10.1073/pnas.84.5.1192.
- Tam, C. and Missiakas, D. (2005) 'Changes in lipopolysaccharide structure induce the  $\sigma$ E-dependent response of *Escherichia coli*', *Molecular Microbiology*, 55(5), pp. 1403–1412. doi: 10.1111/j.1365-2958.2005.04497.x.
- Thakur, K. G., Joshi, A. M. and Gopal, B. (2007) 'Structural and biophysical studies on two promoter recognition domains of the extra-cytoplasmic function  $\sigma$  factor  $\sigma$ C from *Mycobacterium tuberculosis*', *Journal of Biological Chemistry*. doi: 10.1074/jbc.M606283200.
- Tran, K. and Gralla, J. D. (2008) 'Control of the timing of promoter escape and RNA catalysis by the transcription factor IIB fingertip', *Journal of Biological Chemistry*, 283(23), pp. 15665–15671. doi: 10.1074/jbc.M801439200.
- Tran, N. T. *et al.* (2019) 'Defining the regulon of genes controlled by  $\sigma$  E, a key regulator of the cell envelope stress response in *Streptomyces coelicolor*', *Molecular Microbiology*. doi: 10.1111/mmi.14250.
- Treviño-Quintanilla, L., Freyre-González, J. and Martínez-Flores, I. (2013) 'Anti-Sigma Factors in *E. coli*: Common Regulatory Mechanisms Controlling Sigma Factors Availability', *Current Genomics*. Bentham Science Publishers Ltd., 14(6), pp. 378–387. doi: 10.2174/1389202911314060007.
- Tsirigos, K. D. *et al.* (2015) 'The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides', *Nucleic Acids Research*. doi: 10.1093/nar/gkv485.
- Tsubery, H. *et al.* (2000) 'Structure–Function Studies of Polymyxin B Nonapeptide: Implications to Sensitization of Gram-Negative Bacteria #', *Journal of Medicinal Chemistry*, 43(16), pp. 3085–3092. doi: 10.1021/jm0000057.
- Typas, A. and Hengge, R. (2006) 'Role of the spacer between the -35 and -10 regions in sigma promoter selectivity in *Escherichia coli*', *Molecular Microbiology*, 59(3), pp. 1037–1051. doi: 10.1111/j.1365-2958.2005.04998.x.
- Uzman, A. *et al.* (2000) 'Molecular Cell Biology (4th edition) New York, NY, 2000, ISBN 0-7167-3136-3', *Biochemistry and Molecular Biology Education*, 29, p. Section 1.2 The Molecules of Life. doi: 10.1016/S1470-8175(01)00023-6.
- Vecchione, S. and Fritz, G. (2019) 'CRIMoClo plasmids for modular assembly and orthogonal chromosomal integration of synthetic circuits in *Escherichia coli*', *Journal of Biological Engineering*, 13(1), p. 92. doi: 10.1186/s13036-019-0218-8.
- Velkov, T. *et al.* (2013) 'Pharmacology of polymyxins: New insights into an 'old class of antibiotics'', *Future Microbiology*, 8(6), pp. 711–724. doi: 10.2217/fmb.13.39.
- Vellanoweth, R. L. and Rabinowitz, J. C. (1992) 'The influence of ribosome-binding-site elements on translational efficiency in *Bacillus subtilis* and *Escherichia coli* in vivo', *Molecular Microbiology*, 6(9), pp. 1105–1114. doi: 10.1111/j.1365-2958.1992.tb01548.x.
- Vingadassalom, D. *et al.* (2005) 'An unusual primary sigma factor in the Bacteroidetes phylum', *Molecular Microbiology*, 56(4), pp. 888–902. doi: 10.1111/j.1365-2958.2005.04590.x.
- Vuthoori, S. *et al.* (2001) 'Domain 1.1 of the  $\sigma$ 70 subunit of *Escherichia coli* RNA polymerase modulates the formation of stable polymerase/promoter complexes', *Journal of Molecular Biology*. Academic Press, 309(3), pp. 561–572. doi: 10.1006/jmbi.2001.4690.
- Waterhouse, A. *et al.* (2018) 'SWISS-MODEL: Homology modelling of protein structures and complexes', *Nucleic Acids Research*. Oxford University Press, 46(W1), pp. W296–W303. doi: 10.1093/nar/gky427.

- Weber, E. *et al.* (2011) 'A modular cloning system for standardized assembly of multigene constructs', *PLoS ONE*, 6(2). doi: 10.1371/journal.pone.0016765.
- Wecke, T. *et al.* (2012) 'Extracytoplasmic function  $\sigma$  factors of the widely distributed group ECF41 contain a fused regulatory domain', *MicrobiologyOpen*. doi: 10.1002/mbo3.22.
- Wei, Z. *et al.* (2019) 'Alternative  $\sigma$ I/anti- $\sigma$ I factors represent a unique form of bacterial  $\sigma$ /anti- $\sigma$  complex', *Nucleic acids research*. NLM (Medline), 47(11), pp. 5988–5997. doi: 10.1093/nar/gkz355.
- Weigt, Martin *et al.* (2009) 'Identification of direct residue contacts in protein-protein interaction by message passing.', *Proceedings of the National Academy of Sciences of the United States of America*. doi: 10.1073/pnas.0805923106.
- Weigt, M. *et al.* (2009) 'Identification of direct residue contacts in protein-protein interaction by message passing', *Proceedings of the National Academy of Sciences*. doi: 10.1073/pnas.0805923106.
- Weiss, S. B. and Gladstone, L. (1959) 'A Mammalian System for The Incorporation of Cytidine Triphosphate into Ribonucleic Acid', *Journal of the American Chemical Society*, 81(15), pp. 4118–4119. doi: 10.1021/ja01524a087.
- Weiss, S. B. and Nakamoto, T. (1961) 'Net synthesis of ribonucleic acid with a microbial enzyme requiring deoxyribonucleic acid and four ribonucleoside triphosphates.', *The Journal of biological chemistry*, 236, pp. PC18-20. Available at: <http://www.ncbi.nlm.nih.gov/pubmed/13784098> (Accessed: 5 November 2019).
- Werner, F. and Grohmann, D. (2011) 'Evolution of multisubunit RNA polymerases in the three domains of life', *Nature Reviews Microbiology*, pp. 85–98. doi: 10.1038/nrmicro2507.
- Wiegand, S. *et al.* (2019) 'Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology', *Nature Microbiology 2019*. Nature Publishing Group, pp. 1–15. doi: 10.1038/s41564-019-0588-1.
- Wilbur, W. J. and Lipman, D. J. (1983) 'Rapid similarity searches of nucleic acid and protein data banks.', *Proceedings of the National Academy of Sciences of the United States of America*, 80(3), pp. 726–30. doi: 10.1073/pnas.80.3.726.
- Wilson, M. J. and Lamont, I. L. (2006) 'Mutational analysis of an extracytoplasmic-function sigma factor to investigate its interactions with RNA polymerase and DNA', *Journal of Bacteriology*. doi: 10.1128/JB.188.5.1935-1942.2006.
- Wu, H. *et al.* (2019) 'The role of C-terminal extensions in controlling ECF  $\sigma$  factor activity in the widely conserved groups ECF 41 and ECF 42', *Molecular Microbiology*. doi: 10.1111/mmi.14261.
- Xue, Y. *et al.* (2008) 'GPS 2.0, a tool to predict kinase-specific phosphorylation sites in hierarchy', *Molecular and Cellular Proteomics*, 7(9), pp. 1598–1606. doi: 10.1074/mcp.M700574-MCP200.
- Yan Ning Zhou, Walter, W. A. and Gross, C. A. (1992) 'A mutant  $\sigma$ 32 with a small deletion in conserved region 3 of  $\sigma$  has reduced affinity for core RNA polymerase', *Journal of Bacteriology*, 174(15), pp. 5005–5012. doi: 10.1128/jb.174.15.5005-5012.1992.
- Yanamandra, S. S. *et al.* (2012) 'Role of the Porphyromonas gingivalis extracytoplasmic function sigma factor, SigH', *Molecular Oral Microbiology*. doi: 10.1111/j.2041-1014.2012.00643.x.
- Yang, J. *et al.* (2014) 'The I-TASSER suite: Protein structure and function prediction', *Nature Methods*. Nature Publishing Group, pp. 7–8. doi: 10.1038/nmeth.3213.
- Yeats, C., Finn, R. D. and Bateman, A. (2002) 'The PASTA domain: A  $\beta$ -lactam-binding domain', *Trends in Biochemical Sciences*. Elsevier Ltd, pp. 438–440. doi: 10.1016/S0968-0004(02)02164-3.
- Yokoyama, S. *et al.* (2002) 'Crystal structure of a bacterial RNA polymerase holoenzyme at 2.6 Å resolution', *Nature*, 417(6890), pp. 712–719. doi: 10.1038/nature752.
- Yoshimura, M. *et al.* (2004) 'Interaction of Bacillus subtilis extracytoplasmic function (ECF) sigma factors with the N-terminal regions of their potential anti-sigma factors.', *Microbiology (Reading)*,

*England*), 150(Pt 3), pp. 591–9. doi: 10.1099/mic.0.26712-0.

Yu, Y. *et al.* (2012) ‘Putative type VI secretion systems of *Vibrio parahaemolyticus* contribute to adhesion to cultured cell monolayers’, *Archives of Microbiology*, 194(10), pp. 827–835. doi: 10.1007/s00203-012-0816-z.

Zhang, G. and Darst, S. A. (1998) ‘Structure of the *Escherichia coli* RNA polymerase  $\alpha$  subunit amino-terminal domain’, *Science*, 281(5374), pp. 262–266. doi: 10.1126/science.281.5374.262.

Zhang, J. and Landick, R. (2016) ‘A Two-Way Street: Regulatory Interplay between RNA Polymerase and Nascent RNA Structure’, *Trends in Biochemical Sciences*. Elsevier Ltd, pp. 293–310. doi: 10.1016/j.tibs.2015.12.009.

Zhang, Y. *et al.* (2012) ‘Structural basis of transcription initiation’, *Science*. American Association for the Advancement of Science, 338(6110), pp. 1076–1080. doi: 10.1126/science.1227786.

Zhang, Z. and Hendrickson, W. A. (2010) ‘Structural Characterization of the Predominant Family of Histidine Kinase Sensor Domains’, *Journal of Molecular Biology*, 400(3), pp. 335–353. doi: 10.1016/j.jmb.2010.04.049.

Zimmermann, L. *et al.* (2018) ‘A Completely Reimplemented MPI Bioinformatics Toolkit with a New HHpred Server at its Core.’, *Journal of molecular biology*, 430(15), pp. 2237–2243. doi: 10.1016/j.jmb.2017.12.007.

## Supplementary tables

**Table S3.1. Description of ECF groups** in terms of taxonomic origin, predicted target promoter motif, percentage of overlap with original groups, presence of C- or N-terminal extensions and domains therein, percentage of sequences with transmembrane helices, other features of the ECF protein, presence and type of the associated anti- $\sigma$  factor, number of transmembrane helices of the associated anti- $\sigma$  factor and percentage of anti- $\sigma$  factors that contain this number of helices according to the consensus prediction of TopCons (Tsirigos et al., 2015), position where the associated anti- $\sigma$  factor is encoded relative to the ECF coding sequence, presence of other conserved regulatory elements in the genetic neighborhoods of members of a group, elements conserved in the genetic neighborhood, where conservation indicates its presence in more than 75% of the genetic neighborhoods. In most of the cases, these data were extracted exclusively from ECFs from organisms tagged as “representative” or “reference” in NCBI (<https://www.ncbi.nlm.nih.gov>), selecting only RefSeq assemblies when both RefSeq and GenBank representative/reference assemblies are available. However, few ECFs come from this type of organisms are available for some ECF groups and the analysis was done in all the members of the group. These cases are adequately labeled in the “Non-Ref/Rep” column.

General features			ECF features					Regulation				
ECF group	Taxonomy	Predicted target promoter	Original group	C terminal extension and domain	N-terminal extension	TM helix(es) ECF	Other sequence features	Putative anti- $\sigma$ factor and domain	TM helix(es) putative anti- $\sigma$	Position	Other	Non-Ref/Rep?
ECF201	Firmicutes (100%)	-	-	-	-	-	-	-	-	-	-	-
ECF46	Planctomycetes (44.44%), Proteobacteria (37.04%), Verrucomicrobia (18.56%)	-	ECF46 (100.0%)	-	-	-	2C in $\sigma_4$ , C in $\sigma_2$ (s4)	FecR + Concanavalin A-like lectin/glucanases superfamily	1 (92.59%)	1	-	-
ECF202	-	-	-	-	-	-	-	-	-	-	-	-
ECF203	Actinobacteria (100%)	-	-	-	-	-	-	-	-	-	TetR repressor	-
ECF204	Actinobacteria (100%)	GGAACC-Xn-CGGTGTA	-	-	-	-	-	CAS, DUF3040, RskA-like, DUF4367	1 (80.36%)	1	-	-
ECF205	Actinobacteria (93.75%), Proteobacteria (6.25%)	-	-	SnoaL-like (9.38%)	-	-	-	-	-	-	-	-
ECF114	Bacteroidetes (99.11%)	-	ECF114 (36.83%)	-	-	-	-	-	-	-	-	-
ECF28	Proteobacteria (99.34%), Cyanobacteria (0.66%)	-	ECF28 (91.24%)	-	-	-	-	DUF3379	1 (93.21%)	1	-	-
ECF206	Acidobacteria (89.29%), Proteobacteria (10.71%)	-	ECF26 (63.83%)	-	-	-	C in $\sigma_4$ (s2)	ZAS, RskA-like	1 (51.85%)	1	-	-
ECF207	Proteobacteria (97.89%), Verrucomicrobia (1.05%), Planctomycetes (1.05%)	GGAATAAA-Xn-GTC	ECF26 (93.49%)	~20aa s1 and s6	~50aa in s11	-	C in $\sigma_2$ in some subgroups	ZAS, RskA-like	1 (46.81%)	1	-	-
ECF12	Actinobacteria (88.76%), Bacteroidetes (6.91%), Proteobacteria (1.75%), Chlorobi (1.24%), Ignavibacteriae (0.31%), Planctomycetes (0.21%), Firmicutes (0.10%), Calditrichaeota (0.10%)	-	ECF12 (86.12%), ECF26 (3.5%)	-	-	-	-	ZAS	0 (97.09%)	1, others are possible	-	-
ECF15	Proteobacteria (99.86%), Cyanobacteria (0.14%)	GGAAC-Xn-CATT	ECF15 (88.52%), ECF26 (1.76%)	~30 aa	~60aa s2 and some s1	-	CGC in $\sigma_4$	CAS	0 (90.32%)	Variable	Anti-anti $\sigma$ factor fused to response regulator, histidine kinase	-
ECF26	Proteobacteria (99.87%), Bacteroidetes (0.13%)	GGAATAAA-Xn-GTT	ECF26 (94.35%)	-	~5-20aa in some	-	-	ZAS	1 (85.83%)	1	-	-

subgroups, conserved													
ECF208	Spirochaetes (100%)	-	-	-	-	-	-	-	CAS	1 (56%)	1	-	-
ECF209	Actinobacteria (100%)	CCGTGAACC-Xn-CCGACTGT	-	-	-	-	-	C in $\sigma_4$	CAS	4 (66.66%)	1	-	-
ECF58	Planctomycetes (100%)	-	ECF58 (100.0%)	-	-	-	-	C in $\sigma_2$	-	-	-	-	-
ECF54	Actinobacteria (100%)	GTATCAG-Xn-CTCC	ECF54 (100.0%)	-	~10 aa in s2, s6 and s5	-	-	2C in $\sigma_4$ in some subgroups	-	-	-	-	Subtilase, carboxypeptidase regulatory-like domain, CHAT domain and tetratricopeptide repeats
ECF122	Actinobacteria (100%)	CTCAC-Xn-CGTCTAC	ECF122 (94.99%)	-	-	-	-	C in $\sigma_2$ , slightly shorter linker	-	-	-	-	-
ECF36	Actinobacteria (100%)	GTC-Xn-GTTCCCG	ECF36 (48.17%)	-	DUF2275 with Zinc-finger (58.33); ~10aa in s6,s10,s5,s1	3.05 in s3	-	CGC in $\sigma_4$ of some subgroups	ZAS, DUF2275, CAS	4 (41.85%)	-1	-	-
ECF57	Planctomycetes (100%)	-	ECF57 (95.8%)	homeodomain-like domain in 8.33% of the proteins from s1 and a WD40-like beta propeller repeat in 21.05% of the contexts from s2	~20aa longer	~2	-	C in $\sigma_2$ and $\sigma_4$	-	-	-	-	-
ECF210	Proteobacteria (100%)	-	-	-	-	-	-	-	-	-	-	-	-
ECF211	Actinobacteria (100%)	-	-	-	-	-	-	C in $\sigma_2$ (s1)	CAS	6 (75%)	1	-	-
ECF118	Actinobacteria (100%)	TGTGAC-Xn-AACC	ECF118 (70.48%)	~32aa (2/3 s9)	-	-	-	C in $\sigma_2$	ZAS	6 (96.03%)	1	-	-
ECF212	Firmicutes (100%)	TGAAC-Xn-TGTATA	-	-	-	-	-	-	CAS	1 (100%)	1	-	-
ECF33	Proteobacteria (100%)	TGAACCTTT-Xn-AC	ECF33 (44.11%)	-	~15aa in some ECFs	-	-	-	ZAS	1 (73.64%)	1	-	-
ECF213	Chloroflexi (8.33%), Planctomycetes (33.33%) and Firmicutes (58.33%)	CCAACA-Xn-CGTTATCTA	-	-	-	-	-	-	CAS, RskA-like	1 (50%)	1	-	-
ECF17	Actinobacteria (100%)	TGAACC-Xn-CGT	ECF17 (85.82%)	-	-	-	-	S/T rich N-terminus of s1	ZAS	1 (83.3%)	1	-	-
ECF11	Proteobacteria (100%)	GTGATC-Xn-CGTA	ECF11 (83.58%)	-	-	-	-	-	ChrR Cupin-like domain	0 (100%)	1	-	-
ECF214	Firmicutes (24.16%), Bacteroidetes (75.56%)	-	ECF10 (1.03%)	-	-	-	-	2C in $\sigma_4$ of s1	RskA-like	1 (94.33%)	1	-	-
ECF18	Proteobacteria (100%)	TGCATCTT-Xn-CGTA	ECF18 (95.33%)	-	-	-	-	2C in $\sigma_4$	RskA-like	1 (96.64%)	1	-	-
ECF19	Actinobacteria (86.21%), Cyanobacteria (5.65%), Firmicutes (1.95%), Chloroflexi	AT-Xn-CG or TG	ECF19 (56.46%), ECF126	-	-	-	-	2C in $\sigma_4$ in s1, s2, s6, s11 and s13	RskA-like, ChrR-like (s19), ZAS fused to the N-terminal	1 RskA (79.71%), 0 ZAS and	1	-	Putative anti-anti $\sigma$ in -1 of s7 (94.44%)

	(0.51%), Proteobacteria (1.65%), Verrucomicrobia (0.51%), Deinococcus-Thermus (1.54%), Acidobacteria (1.03%) and Planctomycetes (0.62%) and Gemmatimonadetes (0.10%)		(3.66%), ECF34 (2.72%), ECF18 (1.0%)					domain of a mycothiol malenylpyruvate isomerase (s7, s18), none (s4)	ChrR (19.49%)			and s18 (40%)	
ECF61	Planctomycetes (100%)	-	ECF61 (100.0%)	-	-	-	Longer $\sigma_{4.1}$	-	-	-	-	Protein kinase (83.33%)	-
ECF215	Candidatus	-	-	-	-	-	1/2 C in $\sigma_2$ , C in $\sigma_4$ some sequences	CAS, DUF5472	1 (63.89%)	1	-	-	done with no-ref/no-rep
ECF216	Proteobacteria (100%)	-	-	~30 aa with TMH in s2	-	18% in s2	~12 extra aa in s2 (before $\sigma_2$ )	ZAS (no Pfam domain in s2)	1 (69.23%)	1	-	-	-
ECF62	Planctomycetes (100%)	-	ECF62 (92.59%)	-	-	-	-	-	-	-	-	Protein kinase, fused to WD in s1 and s2, zinc finger in s1, s2 and s3	-
ECF16	Proteobacteria (97.33%), Spirochaetes (2.67%)	TTC-Xn-TAC or TAAC	ECF16 (96.5%)	-	-	-	-	DUF1109	6 (95.8%)	1	-	2C facing the periplasm in anti- $\sigma$	-
ECF40	Actinobacteria (100%)	-	ECF40 (98.37%)	-	-	-	2C in $\sigma_2$	RsdA-like	1 (50%)	1	-	-	-
ECF132	Actinobacteria (100%)	-	ECF132 (80.22%)	-	-	-	-	CAS	1 (58.14%)	1	-	-	-
ECF123	Actinobacteria (100%)	-	ECF123 (97.48%)	~20-30 aa in s1 and s3	~10 aa in s2	-	-	CAS	1 (73.33%)	1	-	-	-
ECF53	Acidobacteria (100%)	TGTTTATC-Xn-TCTCC	ECF53 (100.0%)	~350 aa, putative zinc-finger (30%) + tm	-	1	-	-	-	-	-	-	-
ECF217	Planctomycetes (100%)	-	ECF59 (100.0%)	-	-	-	-	-	-	-	-	Protein kinase In +1	-
ECF52	Actinobacteria (100%)	-	ECF52 (100.0%)	Zinc-finger domain (all) and a carbohydrate-binding domain such as NPCBM/NEW2 domain (s1) or ricin-type beta-trefoil lectin domain-like (s4)	-	1 or 2 (0 in s8)	CC in $\sigma_2$ (s1, s4)	-	-	-	-	-	-
ECF127	Actinobacteria (100%)	GTAACCC -35	ECF127 (100.0%)	-	-	-	-	Rieske [2Fe-2S] + TAT signal peptide in 26.84% (s1) and 40% (s2)	0 (100%)	1	-	-	-
ECF125	Actinobacteria (100%)	AAAA-Xn-CGTCGGA	ECF125 (100.0%)	-	-	-	C in $\sigma_2$ , CX4C in $\sigma_4$ (s1) or C in $\sigma_4$ (s2)	Glyoxalase/bleomycin resistance protein/ dioxygenase	0 (100%)	1	-	-	-
ECF218	Acidobacteria (100%)	Subgroup dependent	ECF49 (33.49%), ECF47	~40aa s1, ~30aa s7, ~20aa s17	-	-	-	RskA-like, DUF4349(s8), CAS	1 (74.33%)	1	-	-	-

			(29.27%), ECF50 (28.48%)										
ECF59	Planctomycetes (100%)	TC in -35	ECF59 (100.0%)	-	-	-	~6 amino acid longer linker	-	-	-	-	-	Protein kinase in +1
ECF48	Actinobacteria (100%)	-	ECF48 (14.0%)	Putative zinc-finger (60 (s2) to 400aa (s4))	-	0 (s2) to ~1 TM (s1, s3 and s4)	C in $\sigma_2$ of some sequences, 3C in $\sigma_2$ (s1)	-	-	-	-	-	-
ECF131	Actinobacteria (100%)	TGTCA-Xn-CGCAC	ECF131 (60.71%)	-	-	-	2C in $\sigma_2$ (s3), C in $\sigma_4$ (s2)	CAS	1 (58.33%)	-1	-	-	TetR repressors encoded in neighborhood of s1
ECF128	Actinobacteria (100%)	CCAACC-Xn-GGCGTC	ECF128 (91.59%)	-	-	-	C in $\sigma_2$ (s1)	CAS, BNR/Asp-box repeat, sortilin, photosynthesis system II assembly factor YCF48	1 (67.16%)	1	-	-	-
ECF130	Actinobacteria (100%)	-	ECF130 (73.85%)	-	-	-	-	-	-	-	-	-	-
ECF219	Actinobacteria (100%)	-	-	-	-	-	lack of $\sigma_{2.1}$ (s2 and s3), G rich linker	CAS s2	1 (66.67%) s2	1 s2	-	-	-
ECF43	Proteobacteria (100%)	GTC-Xn-CAG	ECF43 (100.0%)	-	-	-	ECF $\sigma$ factor domain. S/T in first part of $\sigma_{2.2}$	-	-	-	-	-	Protein kinase in +1
ECF42	Actinobacteria (70.66%), Proteobacteria (22.49%), Bacteroidetes (2.30%), Firmicutes (1.21%), Acidobacteria (1.13%), Cyanobacteria (0.77%), Planctomycetes (0.44%), Verrucomicrobia (0.4%), Chloroflexi (0.24%), Spirochaetes (0.16%), Armatimonadetes (0.08%), Nitrospirae (0.08%)	TGTCGAT-Xn-CGTC	ECF42 (99.9%)	~200aa Tetratricopeptide repeat	-	-	Conserved C or CC in $\sigma_4$	-	-	-	-	-	YCII domain in -1
ECF220	Proteobacteria (100%)	-	-	~0 to 400aa	-	1 s1	C in C-terminus of s1	-	-	-	-	-	Two transmembrane proteins in +1 and -1 of s1
ECF221	Actinobacteria (100%)	CGAACGTTT C-Xn-GCGTCxTAG	-	-	-	-	Positively charged N-terminus	DUF4131, DUF2694, RskA-like, CAS	1 (72.55%)	1	-	-	-
ECF222	Actinobacteria (100%)	-	-	-	-	-	-	CAS	1 (64.29%)	1	-	-	-
ECF51	Actinobacteria (100%)	GCAACCG-Xn-GCGTGTC	ECF51 (97.9%)	s5 (15.38%) PQQ domain fused	~50aa s15	s5 (15.35%, PQQ domain fused)	-	CAS	1 (70.73%)	1	-	-	Putative anti-anti $\sigma$ factor in -1 of s9
ECF38	Actinobacteria (100%)	AACC-Xn-TC	ECF38 (73.09%), ECF39 (0.34%)	-	~30aa (s4), ~20aa (s1), ~40aa (s3)	-	-	CAS, RskA-like	1 (67.46%)	1	-	-	-

ECF39	Actinobacteria (100%)	CAACC-Xn-CGTC	ECF39 (93.15%)	~20aa s3 with conserved ERCAA motif, ~50aa s25	~70aa s1, ~20aa s2	-	-	CAS, RskA-like, DUF3040, PASTA, WD40	1 (72.31%)	1	s1, s3 associated to 2CS, s2 associated to RR only	-
ECF223	Firmicutes (71.43%), Fusobacteria (28.57%)	AATAAA-Xn-TT	-	-	-	-	-	CAS, RskA-like, PepSY	1 (69.05%)	1	-	-
ECF224	Candidatus	-	-	-	-	-	-	CAS	2 (76.92%)	1	-	done with no-ref/no-rep
ECF225	Proteobacteria (100%)	AAACTTTTT-Xn-CGACT	-	-	-	-	C in $\sigma_2$	ZAS, CAS	1 (18.18%)	1	-	-
ECF226	Acidobacteria (55.56%), Planctomycetes (33.33%), Verrucomicrobia (11.11%)	-	-	-	-	-	-	ZAS, CAS, DUF3520, von Willebrand	1 (100%)	1	-	-
ECF227	Firmicutes (97.17%), Bacteroidetes (2.83%)	AAAAC-Xn-TTATAT	-	~40aa s9	-	-	2C in $\sigma_2$ in several subgroups	CAS	6 (79.16%)	1	ABC transporters	-
ECF31	Firmicutes (100%)	TGAAC-Xn-CGT	ECF31 (91.1%)	-	-	-	-	DUF2207, DUF5345	2 (84.35%)	1	2TM helix in +2, ABC transporters and N-terminal domain of a phospholipase D-nuclease	-
ECF228	Bacteroidetes (99.74%)	-	-	-	-	-	-	-	-	-	2CS stabilization in the only described member, 2CSs not significantly conserved in genetic context	-
ECF35	Proteobacteria (100%)	ACCC-Xn-CGT or TAACCCG-Xn-CGTCT	ECF35 (6.13%)	-	-	-	2C in $\sigma_4$ is some subgroups	CAS, ZAS, tetratricopeptide repeats	1 (67.68%)	1	-	-
ECF229	Spirochaetes (100%)	ATTC in -35	-	-	-	-	C in $\sigma_2$ (s1)	FecR	1 (68.42%)	1	-	-
ECF230	Firmicutes (100%)	AAACTATTT-Xn-TACGAATAT A	-	-	-	-	-	CAS, ABC transporter?	6 (73.33%)	1	-	-
ECF231	Bacteroidetes (96.88%), Acidobacteria (3.13%)	TGTAACCTT in -35	-	-	-	-	-	ZAS with HEAT repeat	1 (96.88%)	1	Adhesine	-
ECF232	Firmicutes (100%)	TGTT-Xn-CGT	-	-	-	-	2C in $\sigma_4$ (s4), 2C in $\sigma_2$ (s2)	CAS with DUF4367 or WD40	1 (85.11%)	1	ABC transporter	-
ECF233	Candidatus	-	-	-	-	-	-	CAS	1 (54.54%)	1	-	done with no-ref/no-rep
ECF234	Firmicutes (100%)	GATA-Xn-ACAA (s2)	-	-	-	-	C in $\sigma_2$	-	-	-	2CS, ABC transporter	-

ECF121	Actinobacteria (100%)	-	ECF121 (73.36%)	-	~30--130aa	-	C in $\sigma_4$	CAS	1 (27.17%)	1	-	-
ECF116	Firmicutes (100%)	TGAAAC-Xn-CGTCTAAT	ECF116 (25.24%)	-	-	-	1/2 C in $\sigma_4$ of s1 and s2	CAS (sporulation and spore germination, DUF)	1 (71.37%)	1	2CS in s1 and s2	-
ECF235	Firmicutes (99.74%) Actinobacteria (0.26%)	TGTAAC -35	ECF110 (12.01%), ECF124 (11.8%), ECF108 (10.79%), ECF40 (0.11%)	s41 (unknown domain but no representative sequence)	s13 (short unknown domain)	-	C in $\sigma_2$ and $\sigma_4$ in some cases	ZAS, $\sigma$ factor regulator	1 (77.79%), 4 (6.55%)	1	Another $\sigma$ factor (s15)	-
ECF120	Bacteroidetes (100%)	TGTAACAAA-Xn-TCGTCAT	ECF120 (20.53%)	-	-	-	s1, s13 pair of Cys in $\sigma_2$ and $\sigma_4$	ZAS, CAS, DUF3379	1 (71.22%)	1	DUF4252, sometimes fused to an auto-transporter adhesin head GIN domain, peptidase S41 (Flavobacteriales)	-
ECF237	Proteobacteria (42.86%), Cyanobacteria (42.86%) Actinobacteria (11.43%), and Chloroflexi (2.86%)	-	-	s1 (Homeodomain-like domain 10%), s5 (unknown)	s4 (unknown)	-	Pair of C (or CC-C) in $\sigma_2$ (except s2), C in $\sigma_4$ (s4 and s5)	-	-	-	Killing trait (Proteobacteria)	-
ECF238	Proteobacteria (56.37%), Bacteroidetes (27.94%), Firmicutes (6.37%), Actinobacteria (4.90%), Acidobacteria (2.94%), Nitrospirae (0.49%), Spirochaetes (0.49%), Chloroflexi (0.49%).	-	ECF24 (84.87%), ECF44 (13.09%)	-	~20 aa (negative charge) in s10, s18 and s13	-	multiple C in $\sigma_2$ and $\sigma_4$ , C-rich C-terminus, C in linker	-	-	-	-	-
ECF239	Bacteroidetes (99.22%)	AACA or GACA-Xn-TAC	ECF10 (95.18%)	-	-	-	2C in $\sigma_2$ of s3, s1, s5, s6, s9, s4 and 3C in $\sigma_2$ of s2	FecR-like fused to a DUF4974	1 (96.88%)	1	TonB-dependent receptor	-
ECF240	Bacteroidetes (99.89%)	-	ECF10 (82.76%)	s106, s155, s180 with unknown conserved domain ~55aa	s28, s64 unknown domain ~25aa	In several subgroups	C in $\sigma_2$ and/or $\sigma_4$ in several subgroups	FecR-like	1 (87.70%)	1, -1 (s1, s22), +2 (s4)	TonB-dependent receptor fused to a carboxypeptidase regulatory domain, carbohydrate-binding outer-membrane protein (susD-like)	-
ECF241	Bacteroidetes (68.28%), Proteobacteria (24.14%), Acidobacteria (6.21%), Spirochaetes (0.69%)	CAATA-Xn-TCT	-	-	-	-	Conserved linker	CAS (heavy-metal resistance)	2 (57.93%), 1 (27.58%) (2 consecutive	1 (s2, s3, s4, s6, s8, s9, s10) -1 (s1, s5, s7)	-	-

										e TMH in N-terminus in MSA)				
ECF242	Proteobacteria (44.19%) and Spirochaetes (55.81%)	-	-	-	-	-	C in $\sigma_2$ (s1) or $\sigma_4$ (s2, s3)	FecR-like fused to a DUF4974 and DUF4880 in some cases	1 (100%)	1	-	-	TonB-dependent receptors (no in Spirochaetes)	-
ECF243	Proteobacteria (100%)	-	ECF05 (45.1%), ECF07 (16.62%), ECF06 (6.67%), ECF09 (5.5%), ECF08 (3.14%), ECF10 (0.39%)	s172, s125 (FecR protein 100%), s57 (homeodomain 100%) short extensions (20-30aa) in s62, s51, s1, s77, s66, s23, s59, s47, s134, s3, s38, s53 (subgroup dependent)	~40 aa s134	1 (s172, s125, s57)	GC or CG in $\sigma_2$ (not always), non-conserved 2,3, different types of linker (some G-rich), C end of $\sigma_4$ (not always)	FecR-like sometimes fused to DUF4974 and DUF4880	1 (82.33%)	1	-	-	TonB-dependent receptors in +2	-
ECF244	Firmicutes (100%)	GAGA-Xn-CGTC	ECF20 (0.52%)	-	-	-	C in $\sigma_2$ of s4, G-rich in C-terminus (except s5)	DUF4179 + zinc-finger	1 (89.13%)	1	-	-	-	-
ECF245	Firmicutes (100%)	TGAAAC-Xn-CGTAT	ECF20 (2.53%), ECF30 (1.78%)	-	-	-	1/2 C in $\sigma_2$	bactofilin + zinc-finger	1 (97.69%)	1	-	-	-	-
ECF246	Spirochaetes (93.75%), Firmicutes (6.25%), Candidatus (non-representatives)	-	-	-	-	-	C in $\sigma_2$ (no s2)	CAS	3 (16.36%), 2 (31.26%)	1	-	-	2CS	done with no-ref/no-rep
ECF247	Chloroflexi (87.5%), unclassified bacteria (12.5%)	-	-	-	-	-	1/2 C in $\sigma_2$ , one 100% conserved	CAS, sometimes with zinc-binding	2 (62.5%)	1	-	-	Bactofilin in +2	-
ECF248	Bacteroidetes (98.39%) and Gemmatimonadetes (0.81%)	-	-	-	-	-	-	-	-	-	-	-	Glycosyl transferase family 2 in -1	-
ECF249	Firmicutes (41.81%), Bacteroidetes (27.68%), Acidobacteria (9.6%), Actinobacteria (4.52%), Chlorobi (4.52%), Planctomycetes (7.91%), Proteobacteria (0.56%) and Verrucomicrobia (3.39%).	CGCAACTTT-Xn-GTT	ECF04 (2.63%), ECF10 (2.19%), ECF52 (0.44%), ECF20 (0.44%)	s14 (Putative zinc-finger)	s5 (unknown domain)	-	C in $\sigma_2$ (s3)	ZAS	2 (43.21%), only s1 1 (38.21%)	1	-	-	-	-
ECF102	Bacteroidetes (66.17%), Gammaproteobacteria (33.83%)	-	ECF102 (61.39%)	-	-	0.43 in s5	2 (3) C in $\sigma_2$ (s1)	CAS	0 (90.91%)	-2	-	-	Mechanosensing with OprF, CfrX and CmpX at least in s1	-
ECF250	Firmicutes (100%)	TGTCAC or TGTCCTTT-Xn-GTTATATA	-	-	-	-	2C in $\sigma_2$	DUF4367, DUF4179	1 (82.67%)	1	-	-	-	-

ECF251	Bacteroidetes (98%), Ignavibacteriae (1%)	GAC-Xn- GGT(T/A)ACA	-	-	-	-	C in $\sigma_2$ and $\sigma_4$ (s1)	DUF2207, CAS	2 (38%), 1 (60%)	1	-	-
ECF23	Firmicutes (100%)	TGATAG-Xn- CGTATTA	ECF23 (55.79%)	-	-	-	C in $\sigma_2$ after DAED	DUF4179, DUF4367, CAS	1 (88.7%)	1	-	-
ECF252	Bacteroidetes (42.11%), Proteobacteria (38.60%), Spirochaetes (19.30%)	AAA-Xn-TTG	-	-	s1 (short)	-	2C in $\sigma_2$ (s2) C in $\sigma_4$ (s5)	RseA-like, CAS	2 (49.12%) 1 (38.6%)	1	-	-
ECF253	Proteobacteria (100%)	TTTGAAGGG- Xn-CGTCTAA	-	-	-	-	S/T rich C- terminus	ZAS, RskA-like	1 (66.67%)	1	-	-
ECF105	Firmicutes (100%)	TG(A/T)AGGG -Xn-CGTCTAT	ECF105 (12.33%)	-	-	-	-	ZAS, CAS	1 (85.11%)	1	-	-
ECF254	Firmicutes (100%)	-	-	-	-	-	-	DUF4367	1 (77.78%)	2	-	-
ECF255	Bacteroidetes (98.13%)	TATGGAT-Xn- GCATCT	ECF119 (16.8%)	-	-	-	2C in $\sigma_2$ , C in $\sigma_4$ , extended N- term + charged	CAS (TonB protein C-terminal + Carboxypeptase regulatory-like domain)	1 (78.5%)	1	-	-
ECF22	Bacteroidetes (78.81%), Proteobacteria (18.14%), Verrucomicrobia (1.19%), Planctomycetes (1.02%), Acidobacteria (0.68%)	TGTGATTTT- Xn- GCGAAT(C/A) AT	ECF22 (88.1%)	-	-	-	C in $\sigma_2$	DUF2207, RseA- like, CAS	4 (48.06%)	1	-	-
ECF256	Firmicutes (100%)	TGAAAC-Xn- CGTTTCAT	ECF31 (1.32%)	-	-	-	-	-	-	-	-	-
ECF107	Firmicutes (100%)	-	ECF107 (98.96%)	-	-	-	-	CAS	2 (100% of +1), 3 (100% of +2)	1 and 2	2 anti- $\sigma$ factors	-
ECF257	Firmicutes (100%)	CTTACA-Xn- TGTA- XnTGAAG	-	-	-	-	1/2 C in $\sigma_2$	-	-	-	ABC transporters	-
ECF258	Firmicutes (100%)	TGAACC-Xn- AATATA	-	-	-	-	C in $\sigma_2$	DUF4179, CAS	1 (77.5%)	1	-	-
ECF259	Firmicutes (100%)	GAACC-Xn-T- rich	-	-	-	-	3C in $\sigma_4$ , longer C-terminus	DUF4179	1 (94.6%)	1	-	-
ECF260	Firmicutes (100%)	CGAAC-Xn- GTATA	-	-	-	-	C in $\sigma_4$	DUF4179, CAS	1 (66.67%)	1	-	-
ECF261	Firmicutes (100%)	(T/C)GAAC- Xn-AATATA	-	-	-	-	-	DUF3600, CAS	1 (75%)	1	-	-
ECF262	Firmicutes (100%)	-	-	All but s4, various domains	-	0-1	Pair of C in $\sigma_4$ (s3), C in $\sigma_2$ (s2, s4)	-	-	-	Other ECF262 in +1 or -1	-
ECF103	Firmicutes (100%)	TGTCACAA- Xn-TCT	ECF103 (83.78%)	-	-	-	-	ZAS, DUF4179	1 (84%)	1	-	-
ECF106	Firmicutes (100%)	-	ECF106 (33.2%)	-	-	-	-	DUF4367	1 (97.14%)	1	-	-
ECF263	Proteobacteria (89.8%), Verrucomicrobia (8.16%), Planctomycetes (2.04%)	TGCAC(A/C)- Xn-CGTC	-	230(s1,s2,s4), 330aa(s3) with fatty acid desaturase (2 cases in s4 and one in s2)	-	5 (s1, s2, s4), 7 (s3)	C in $\sigma_2$	DUF2007	1 (81.25%)	-1	-	-
ECF264	Firmicutes (100%)	GTAACCTT- Xn- GGCACATT	-	~350 aa with DUF1835	-	-	C in $\sigma_2$	-	-	-	-	-

ECF265	Firmicutes (95.80%), Actinobacteria (3.50%)	AAC-Xn-CGTC	ECF30 (46.6%), ECF55 (5.84%), ECF112 (0.51%)	-	-	-	longer linker (s7,s4,s5,s24, s19, s17, s14), C in $\sigma_2$	DUF4179, RskA- like, CAS	1 (63.79%)	1	PadR transcriptional repressor and RodA in s4 and s7	-
ECF266	Firmicutes (100%)	TGCAACA- Xn-ACTCT	-	-	-	-	-	Beta propeller domain	1 (84.38%)	1	-	-
ECF30	Firmicutes (99.86%)	TGCAACA or TGAAACTTT- Xn-CGTC or CTCTAAT	ECF30 (94.68%)	s80 DUF4901 (100%) DUF4192 (100%)	-	-	C in $\sigma_2$ (exceptions), GG in C- terminus (exceptions)	DUF4179, ZAS, DUF4163, DUF3298, RskA	1 (91.02%)	1	-	-
ECF267	Proteobacteria (100%)	-	-	-	-	-	G D/E rich C- terminus	FecR + TPR	1 (100%)	1	Protein kinase in non- conserved position	-
ECF268	Acidobacteria (100%)	GGGAAC-Xn- GGTGT	-	-	-	-	~10 extra aa in N and C- terminus	ZAS	1 (100%)	1	-	-
ECF269	Proteobacteria (100%)	-	-	-	-	-	2C in $\sigma_4$	CAS	2 (69.23%)	-1	-	-
ECF02	Proteobacteria (100%)	GAACCTT-Xn- GTCT	ECF02 (99.2%)	-	-	-	C in $\sigma_4$	RseA-like	1 (~100%)	1	PDZ domain, RseB, RseC	-
ECF32	Proteobacteria (100%)	-	ECF32 (100.0%)	-	-	-	S/T rich N- terminus	-	-	-	2CS	-
ECF25	Cyanobacteria (50%), Firmicutes (11.88%), Proteobacteria (15.84%), Spirochaetes (s3, 2.97%), Fibrobacteres (0.5%), Nitrospirae (0.5%), Deinococcus-Thermus (6.44%), Chloroflexi (6.44%), Gemmatimonadetes (1%), Ignavibacteriae (1%), Acidobacteria (1%), Bacteroidetes (1.98%), Calditrichaeota (0.5%) Bacteroidetes (95.44%), Proteobacteria (1.27%), Verrucomicrobia (2.28%), Kiritimatiellaeota (0.25%), Chlorobi (0.25%)	GGAAC-Xn- GTC	ECF25 (30.6%), ECF20 (0.19%)	-	-	-	C in $\sigma_2$ in some subgroups	ZAS, DUF4384, DUF4349	1 (74.75%), 2 in s14, s16, s18 and s19; 3 in s20	1	-	-
ECF03	Proteobacteria (1.27%), Verrucomicrobia (2.28%), Kiritimatiellaeota (0.25%), Chlorobi (0.25%)	TTAAACC-Xn- TC	ECF03 (89.41%)	~30aa s21 (unknown domain)	-	-	C in $\sigma_2$ of some subgroups	DUF4179, CAS	1 (62.03%)	1	-	-
ECF270	Proteobacteria (37.5%), Planctomycetes (12.5%), Ignavibacteriae (25%), Nitrospirae (25%)	-	-	s2 (putative zinc-finger)	-	-	Pair of C in $\sigma_2$ (except s1)	ZAS	0 (100%)	1	-	-
ECF271	Chloroflexi (81.82%) Acidobacteria (9.09%)	GCTGCTG-Xn- GXTCA	ECF03 (2.5%)	-	-	-	CxCx34C in $\sigma_2$ of most s1	ZAS	0 (100%)	1	-	-
ECF29	Proteobacteria (83.64%), Bacteroidetes (16.36%)	GGGAACCT- Xn- CATCCAAT or TTAGAT-Xn- GTTACA (s18, s14)	ECF29 (88.98%)	~30 aa with RCE/D motif	-	-	RCE/D motif in C-terminus extension	-	-	-	-	-

ECF272	Proteobacteria (97%), Gemmatimonadetes (3%)	TTTAAACCTT T-Xn- CG(A/T)CTAA	-	-	~10 aa longer N-terminus	-	C in $\sigma_2$ of some subgroups, longer N- terminus	RskA-like	1 (87%)	1	Putative adhesine	-
ECF273	Proteobacteria (100%)	AGTATAGG- Xn-CC	-	-	~10-20 S/T/P rich	-	-	RskA-like, CAS	1 (80%)	1	PDZ domain +2	-
ECF274	Bacteroidetes (100%)	CAACCTTT- Xn-CTCTTT	-	-	-	-	-	RskA-like, CAS	1 (81.25%)	1	-	-
ECF275	Bacteroidetes (99.61%)	TGxAAC-Xn- GTC	-	-	-	-	C in C- terminus, truncated proteins in some cases (different criteria to choose starting M?)	RskA-like, CAS, DUF3379	1 (51.95%)	1	-	-
ECF21	Bacteroidetes (99.62%)	GCAACC-Xn- CGTCT	ECF21 (62.35%)	-	~380 aa with outer membrane beta-barrel in s12 and s29, this protein is the anti- $\sigma$ factor in the rest	~1 (s12, s29)	C in $\sigma_2$	Outer membrane beta-barrel	1 (73.31%)	1	-	-
ECF111	Cyanobacteria (68.75%), Thermotogae (14.06%), Proteobacteria (10.94%), Acidobacteria (4.69%) Fibrobacteres (1.56%)	TTGAACCT- Xn-GTCTAAA	ECF111 (15.25%)	-	-	-	CXnCG in $\sigma_2$ of some subgroups	CAS	1 (64.91%)	1	-	-
ECF14	Actinobacteria (93.53%), Proteobacteria (3.07%), Armatimonadetes (0.65%)	CTACTGG -35	ECF14 (81.93%)	-	some in s1 and s6 (~50 aa, unknown domain)	-	C in $\sigma_4$ (no s6)	ZAS	1 (52.14%), 0 (47%)	1	Anti- $\sigma$ factor must be phosphorylated by PknB to be degraded in M. tuberculosis, O- methyltransfer ase in -1	-
ECF276	Firmicutes (60.98%), Proteobacteria (29.27%), Planctomycetes (9.76%)	AAC-Xn-TCT	ECF30 (5.04%)	-	~10 aa longer N-terminus, s5 (non- representatives )	-	C in $\sigma_2$	ZAS	2 (56.1%), 1 (29.27%)	1	-	-
ECF277	Acidobacteria (71.43%), Bacteroidetes (14.29%), Proteobacteria (14.29%)	-	-	-	s1 (~70aa) and s4 (~40aa)	-	1 (s4, s3) or 2 C (s1) in $\sigma_2$	ZAS, RskA-like, HEAT (s1), CAS	1 (82.86%)	1	PDZ domain (s1 and s5)	-
ECF278	Proteobacteria (83.33%), Acidobacteria (16.67%)	GTGAC-Xn- GTATA	-	-	-	-	-	Glycogen recognition site, ZAS	1 (100%)	1	-	-
ECF279	Bacteroidetes (100%)	TTGCAA-Xn- CGTCTAA	-	-	-	-	C in $\sigma_4$	DUF5056	2 (52.5%)	1	-	-
ECF280	Proteobacteria (100%)	-	-	-	-	-	Cx3CxGxG in $\sigma_4$	CAS	2 (73.33%)	1	ECF in -3 or -4 of s1 and s3	-
ECF281	Firmicutes (41.44%), Actinobacteria (23.42%), Deinococcus-Thermus (18.92%), Proteobacteria (16.22%)	GGAACCTT-Xn- GTCTAA	ECF20 (10.45%), ECF113 (0.85%)	-	-	-	C in $\sigma_2$ and/or $\sigma_4$ in some subgroups	ZAS (+ DUF4349)	1 (79.69%)	1	-	-

ECF282	Actinobacteria (100%)	-	-	-	-	-	C in $\sigma_2$	-	-	-	AA in C-terminus of s2 that is targeted by ClpXP for inhibition	-
ECF283	Actinobacteria (100%)	CCTT-Xn-AG	-	-	-	-	S/T rich linker	-	-	-	Protein kinase in -1 (or +2 in s5)	-
ECF284	Actinobacteria (100%)	GAATCCTTT-Xn-CTCTT	-	-	-	-	C in $\sigma_2$	AphC/TSA family, CAS	2 (57.14%)	1	-	-
ECF285	Proteobacteria (100%)	AA-Xn-CCTCT	-	-	-	-	2C in $\sigma_2$	DUF2892	2 (67.39%)	-1	-	-
ECF286	Actinobacteria (100%)	GAC-Xn-TC	ECF20 (2.2%)	-	~10 aa longer, S/T and D/E	-	C in $\sigma_2$ and C in $\sigma_4$ in s1 and s2	-	-	-	3.33 Asp23 per ECF	-
ECF287	Actinobacteria (100%)	TCG (-35)	-	~80aa (unknown function)	-	-	2C in $\sigma_4$ , C in C-terminus extension	-	-	-	TetR/MerR transcriptional regulators	-
ECF288	Firmicutes (100%)	TGTCACA-Xn-TGTCTAAT	-	~100aa CGx2GxGxCxC motif	-	-	C-rich $\sigma_4$ with Cx6Cx5C and C-terminal with Cx7CxC, S/T rich $\sigma_2$	-	-	-	DUF2461	-
ECF289	Proteobacteria (100%)	CCCAAGGA-Xn-CGTCTxA	ECF20 (42.27%)	-	-	-	A-rich N-terminal and beginning of $\sigma_{4,2}$ , C in $\sigma_2$	DUF3520 + von Willebrand factor	1 (87.34%)	1	-	-
ECF27	Actinobacteria (100%)	-	ECF27 (90.2%), ECF20 (0.06%)	s1 G-tract extension	s10	-	C in $\sigma_2$ and $\sigma_4$	ZAS, RskA-like, CAS	1 (78.65%)	1	-	-
ECF290	Proteobacteria (97.64%), Planctomycetes (2.36%)	ACGG-Xn-CGT (only s1 and s2)	ECF20 (95.41%)	-	s3 (~10aa) (unknown domain)	-	C in $\sigma_2$	CAS, RskA-like	1 (42.54%)	1	Heavy-metal resistance protein with 1 TM (no in s3)	-
ECF37	Proteobacteria (99.32%), Nitrospirae (0.68%)	TGTCAACC-Xn-CGTT	ECF37 (93.68%)	-	-	-	2C in $\sigma_4$ in s1, C in $\sigma_4$ in the rest	DUF3619, RskA-like	1 (69.1%)	1	DUF3106	-
ECF291	Proteobacteria (98.45%), Cyanobacteria (1.55%)	-	ECF20 (33.76%)	-	-	-	C in $\sigma_2$ , sometimes a pair	CnrY-like, CAS	1 (81.4%)	-2	cnrYHXCBAT controlling metal resistance efflux pump	-
ECF292	Actinobacteria (100%)	-	-	-	-	-	C in $\sigma_2$ (s2) or $\sigma_4$ (s1)	-	-	-	2.08 Asp23 per ECF	-
ECF293	Proteobacteria (53.03%), Bacteroidetes (29.65%), Actinobacteria (16.88%), Verrucomicrobia (0.22%), Spirochaetes (0.22%)	-	ECF13 (43.98%), ECF101 (11.35%), ECF117 (9.46%)	-	-	-	Extended linker, 3C in $\sigma_4$ in most of the subgroups	ZAS	0 (97.92%)	1	-	-
ECF294	Proteobacteria (96.15%), Acidobacteria (3.95%)	-	-	~120aa snoaL_2 domain	-	-	C in $\sigma_2$ in most, 3C in $\sigma_4$ in some	-	-	-	-	-
ECF115	Firmicutes (100%)	ATC-Xn-CGT	ECF115 (40.3%)	~70aa (several domains) ~130aa in s3 (snoaL_2)	-	-	C in $\sigma_2$ , sometimes 2C	-	-	-	Clp protease	-
ECF295	Firmicutes (78.57%), Actinobacteria (21.43%)	-	ECF56 (100.0%)	~120aa snoaL_2 domain	-	-	2C in $\sigma_2$ , 2C in C-terminus	-	-	-	activator of Hsp90 ATPase	-

				extension	homolog 1
ECF56	Actinobacteria (96.15%), Proteobacteria (2.86%), Chloroflexi (0.66%), Gemmatimonadetes (0.22%), Acidobacteria (0.11%)	-	ECF56 (99.89%)	~120aa snoaL_2 (115.49%)	extended linker, 2C in $\sigma_2$
ECF41	Actinobacteria (78.48%), Proteobacteria (14.49%), Firmicutes (Bacillales) (4.88%), Acidobacteria (78.48%), Chloroflexi (0.35%), Cyanobacteria (0.49%), Verrucomicrobia (0.11%), Bacteroidetes (0.21%), Planctomycetes (0.07%), Gemmatimonadetes (0.04%), Armatimonadetes (0.04%)	TGCAACC- Xn-CGTT	ECF41 (99.09%)	~120aa snoaL_2 domain	

**Table S3.2. Distribution of the ECFs of new phyla across the ECF classification.** The number (labeled as N) of ECFs and ECF groups where these ECFs cluster is shown for all the species associated to a phylum and for the set of “representative” and “reference” genomes, this is, assemblies tagged as “representative” or “reference” in NCBI (<https://www.ncbi.nlm.nih.gov>), selecting only RefSeq assemblies when both RefSeq and GenBank representative/reference assemblies are available (labeled as rep/ref in the table).

Phylum	N species	N species (rep/ref)	N ECFs	N ECFs (rep/ref)	ECF groups	ECF groups (rep/ref)	ECF groups where they contribute > 10%	ECF groups where they contribute > 10% (rep/ref)
Aquificae	10	2	10	2				
Armatimonadetes	27	2	370	50	ECF12, ECF14, ECF22, ECF281, ECF41, ECF42, ECF56	ECF14, ECF41, ECF42		
Balneolaeota	6	0	138	0	ECF12, ECF22, ECF228, ECF231, ECF248, ECF249, ECF25, ECF41, ECF42			
Caldiserica	3	1	13	4				
Calditrichaeota	5	1	39	8	ECF12, ECF248, ECF25, ECF252	ECF12, ECF25		
Chlamydiae	4	0	16	0	ECF03, ECF12, ECF21, ECF235, ECF236, ECF279, ECF30			
Chrysiogenetes	3	2	9	6				
Deferribacteres	13	6	24	11				
Dictyoglomi	4	2	8	4				
Elusimicrobia	61	2	294	3	ECF14, ECF22, ECF281, ECF293			
Fibrobacteres	38	2	89	4	ECF111, ECF25	ECF111, ECF25		
Gemmatimonadetes	44	2	531	45	ECF12, ECF19, ECF207, ECF22, ECF248, ECF25, ECF270, ECF272, ECF273, ECF291, ECF293, ECF41, ECF42, ECF56	ECF19, ECF248, ECF25, ECF272, ECF41, ECF42, ECF56	ECF273	
Ignavibacteriae	47	2	501	14	ECF03, ECF12, ECF241, ECF248, ECF249, ECF25, ECF251, ECF252, ECF270, ECF272, ECF277, ECF293, ECF42	ECF12, ECF25, ECF251, ECF270	ECF270	ECF270
Kiritimatiellaeota	2	1	4	2	ECF03	ECF03		
Lentisphaerae	14	1	184	42	ECF03			
Nitrospirae	15	1	31	5	ECF111, ECF12			
Nitrospirae	55	4	195	28	ECF02, ECF111, ECF12, ECF127, ECF238, ECF25, ECF270, ECF293, ECF37, ECF42	ECF238, ECF25, ECF270, ECF37, ECF42	ECF270	ECF270
Synergistetes	4	1	4	1				
Tenericutes	36	7	72	11				
Verrucomicrobia	94	17	1081	312	ECF03, ECF19, ECF207, ECF22, ECF226, ECF249, ECF263, ECF276, ECF293, ECF33, ECF41, ECF42, ECF46	ECF03, ECF19, ECF207, ECF22, ECF226, ECF249, ECF263, ECF293, ECF41, ECF42, ECF46	ECF226, ECF46	ECF226, ECF46
Candidatus Abawacabacteria	1	0	2	0				
Candidatus Acetothermia	1	0	2	0				
Candidatus Adlerbacteria	13	0	25	0	ECF202, ECF246		ECF202	

Candidatus Amesbacteria	34	0	43	0	ECF215, ECF246	ECF215
Candidatus Aminicenantes	10	0	114	0	ECF231	
Candidatus Atribacteria	3	0	3	0	ECF246	
Candidatus Azambacteria	31	0	50	0	ECF224	ECF224
Candidatus Beckwithbacteria	14	0	27	0		
Candidatus Berkelbacteria	5	0	5	0	ECF246	
Candidatus Blackburnbacteria	9	0	13	0		
Candidatus Brennerbacteria	3	0	9	0		
Candidatus Buchananbacteria	19	0	31	0	ECF224, ECF233, ECF246	
Candidatus Campbellbacteria	10	0	17	0	ECF202, ECF246	
Candidatus Chisholmbacteria	5	0	6	0		
Candidatus Cloacimonetes	3	1	4	1		
Candidatus Coatesbacteria	1	0	2	0		
Candidatus Collierbacteria	23	0	25	0	ECF246	
Candidatus Colwellbacteria	10	0	15	0		
Candidatus Curtissbacteria	11	0	13	0	ECF246	
Candidatus Dadabacteria	3	0	11	0		
Candidatus Daviesbacteria	24	0	35	0	ECF215	
Candidatus DeLongbacteria	1	0	2	0		
Candidatus Desantisbacteria	3	0	4	0		
Candidatus Doudnabacteria	33	0	86	0	ECF246	
Candidatus Edwardsbacteria	6	0	29	0		
Candidatus Eisenbacteria	2	0	19	0	ECF272	
Candidatus Falkowbacteria	36	0	92	0	ECF224, ECF233, ECF246	ECF233
Candidatus Firestonebacteria	5	0	37	0	ECF208	
Candidatus Fischerbacteria	1	0	9	0		
Candidatus Fraserbacteria	1	0	2	0		
Candidatus Giovannonibacteria	63	0	122	0	ECF202, ECF224	ECF224
Candidatus Glassbacteria	3	0	15	0		
Candidatus Gottesmanbacteria	45	0	67	0	ECF215, ECF224, ECF246	ECF215
Candidatus Gracilibacteria	1	0	1	0		
Candidatus	1	0	25	0	ECF12, ECF25, ECF293, ECF56	

<b>Handelsmanbacteria</b>						
<b>Candidatus Harrisonbacteria</b>	9	0	16	0		
<b>Candidatus Hydrogenedentes</b>	1	0	3	0		
<b>Candidatus Jacksonbacteria</b>	12	0	32	0	ECF224, ECF233, ECF246	
<b>Candidatus Jorgensenbacteria</b>	15	0	21	0		
<b>Candidatus Kaiserbacteria</b>	57	0	105	0	ECF202, ECF246	ECF202
<b>Candidatus Kerfeldbacteria</b>	9	0	24	0	ECF246	
<b>Candidatus Komeilibacteria</b>	10	0	24	0	ECF246	
<b>Candidatus Kryptonina</b>	36	0	161	0	ECF03, ECF12, ECF248	
<b>Candidatus Kuenenbacteria</b>	8	0	13	0	ECF246	
<b>Candidatus Latescibacteria</b>	1	0	9	0	ECF249, ECF252	
<b>Candidatus Levybacteria</b>	30	0	31	0	ECF246	
<b>Candidatus Liptonbacteria</b>	13	0	24	0	ECF246	
<b>Candidatus Llyodbacteria</b>	14	0	24	0	ECF202, ECF246	ECF202
<b>Candidatus Magasanikbacteria</b>	56	0	143	0	ECF224, ECF233, ECF246	ECF233
<b>Candidatus Margulisbacteria</b>	4	0	17	0		
<b>Candidatus Marinimicrobia</b>	3	0	13	0	ECF249, ECF25	
<b>Candidatus Microgenomates</b>	3	0	5	0	ECF246	
<b>Candidatus Moranbacteria</b>	52	0	127	0	ECF246	
<b>Candidatus Nealsobacteria</b>	16	0	25	0	ECF246	
<b>Candidatus Niyogibacteria</b>	3	0	4	0		
<b>Candidatus Nomurabacteria</b>	72	0	121	0	ECF215, ECF246	
<b>Candidatus Omnitrophica</b>	57	0	127	0	ECF111, ECF25	
<b>Candidatus Pacebacteria</b>	13	0	21	0	ECF246	
<b>Candidatus Parcubacteria</b>	14	0	29	0	ECF202, ECF224, ECF246	
<b>Candidatus Peregrinibacteria</b>	64	0	154	0	ECF246	
<b>Candidatus Portnoybacteria</b>	13	0	19	0	ECF233, ECF246	
<b>Candidatus Raymondbacteria</b>	7	0	43	0		
<b>Candidatus Riflebacteria</b>	2	0	18	0		
<b>Candidatus Roizmanbacteria</b>	67	0	101	0	ECF246	
<b>Candidatus Rokubacteria</b>	41	0	320	0	ECF12, ECF207, ECF25, ECF270, ECF273, ECF276, ECF42, ECF56	ECF270
<b>Candidatus Ryanbacteria</b>	20	0	42	0	ECF224, ECF246	

Candidatus Saccharibacteria	7	0	7	0	ECF246	
Candidatus Schekmanbacteria	8	0	28	0	ECF25	
Candidatus Shapirobacteria	10	0	23	0		
Candidatus Spechtbacteria	9	0	10	0		
Candidatus Staskawiczbacteria	40	0	64	0	ECF246	
Candidatus Sungbacteria	20	0	37	0		
Candidatus Tagabacteria	2	0	2	0		
Candidatus Taylorbacteria	38	0	73	0	ECF224, ECF246	
Candidatus Tectomicrobia	4	0	60	0	ECF12, ECF237, ECF33	
Candidatus Terrybacteria	9	0	9	0		
Candidatus Uhrbacteria	66	0	117	0	ECF233, ECF246	
Candidatus Veblenbacteria	4	0	8	0		
Candidatus Vogelbacteria	7	0	17	0	ECF246	
Candidatus Wallbacteria	1	0	5	0		
Candidatus Wildermathbacteria	23	0	27	0		
Candidatus Wirthbacteria	1	0	2	0		
Candidatus Woesebacteria	70	0	105	0	ECF215, ECF224, ECF246	ECF215
Candidatus Wolfbacteria	28	0	63	0	ECF224, ECF246	
Candidatus Woykebacteria	8	0	11	0		
Candidatus Yanofskybacteria	47	0	83	0	ECF224, ECF246	
Candidatus Yonathbacteria	7	0	14	0	ECF246	
Candidatus Zambryskibacteria	48	0	88	0	ECF246	
candidate division CPR1	2	0	2	0	ECF246	
candidate division CPR2	1	0	1	0		
candidate division CPR3	7	0	25	0		
candidate division Hyd24-12	3	0	6	0	ECF12	
candidate division KD3-62	1	0	4	0		
candidate division NC10	5	0	27	0		
candidate division WOR-3	4	0	24	0		
candidate division WWE3	28	0	28	0	ECF246	
candidate division Zixibacteria	13	0	66	0	ECF12, ECF248, ECF249, ECF25, ECF252	

