986–1005 *Nucleic Acids Research, 2021, Vol. 49, No. 2* *Published online 4 January 2021*
doi: 10.1093/nar/gkaa1229

Expansion and re-classification of the extracytoplasmic function (ECF) σ factor family

Delia Casas-Pastor[1,†], Raphael R. Müller[2,†], Sebastian Jaenicke[2], Karina Brinkrolf[2], Anke Becker[1], Mark J. Buttner[3], Carol A. Gross[4], Thorsten Mascher[5], Alexander Goesmann[2] and Georg Fritz[6,*]

[1]Center for Synthetic Microbiology (SYNMIKRO), Philipps Universität Marburg, Germany, [2]Bioinformatics and Systems Biology, Justus-Liebig-Universität, Giessen, Germany, [3]Department of Molecular Microbiology, John Innes Centre, Norwich Research Park, Norwich NR4 7UH, UK, [4]Department of Microbiology and Immunology, University of California, San Francisco, San Francisco, CA 94158, USA; Department of Cell and Tissue Biology, University of California, San Francisco, San Francisco, CA 94158, USA; California Institute of Quantitative Biology, University of California, San Francisco, San Francisco, CA 94158, USA, [5]Institute of Microbiology, Technische Universität Dresden, Germany and [6]School of Molecular Sciences, The University of Western Australia, Perth, Western Australia 6009, Australia

Received October 23, 2020; Revised December 01, 2020; Editorial Decision December 04, 2020; Accepted December 07, 2020

ABSTRACT

Extracytoplasmic function σ factors (ECFs) represent one of the major bacterial signal transduction mechanisms in terms of abundance, diversity and importance, particularly in mediating stress responses. Here, we performed a comprehensive phylogenetic analysis of this protein family by scrutinizing all proteins in the NCBI database. As a result, we identified an average of ∼10 ECFs per bacterial genome and 157 phylogenetic ECF groups that feature a conserved genetic neighborhood and a similar regulation mechanism. Our analysis expands previous classification efforts ∼50-fold, enriches many original ECF groups with previously unclassified proteins and identifies 22 entirely new ECF groups. The ECF groups are hierarchically related to each other and are further composed of subgroups with closely related sequences. This two-tiered classification allows for the accurate prediction of common promoter motifs and the inference of putative regulatory mechanisms across subgroups composing an ECF group. This comprehensive, high-resolution description of the phylogenetic distribution of the ECF family, together with the massive expansion of classified ECF sequences and an openly accessible data repository called 'ECF Hub' (https://www.computational.bio.uni-giessen.de/ecfhub), will serve as a powerful hypothesis-generator to guide future research in the field.

INTRODUCTION

Bacterial homeostasis is achieved through signal transduction mechanisms that connect the extracellular medium with the cytoplasm. Extracytoplasmic function σ factors (ECFs) are the core components of one of the major signal transduction mechanisms in bacteria in terms of abundance and importance of the stress responses they mediate (1). As members of the σ^{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 (2,3). Although ECFs are generally less prevalent than one-component systems (1CS) and two-component systems (2CS), previous studies revealed a large ECF abundance, with an average of six ECFs per bacterial genome (1), a large diversity, with >90 phylogenetic groups (1,4–6), and a diverse range of activation mechanisms (7). However, the range of bacterial genomes analyzed in these studies was limited and the ECF subgroup could be more diverse and abundant than previously thought.

Members of the σ^{70} family are modular proteins composed of up to four core domains ($\sigma_{1–4}$) that are classified into four groups. While group 1 represents the essential, full-length version of the σ^{70} family, σ^{70} groups 2–4, also known as alternative σ factors, are usually non-essential and represent truncations of this general structure (8,9). ECFs (or group 4 σ^{70}s) are the most minimalistic members of σ^{70} since they only contain the σ_2 and σ_4 domains, essential for contact with the RNAP and transcription initiation (8). The functions of both domains are well differentiated: while σ_4 initiates the first step of promoter recognition by bind-

*To whom correspondence should be addressed. Tel: +61 8 6488 3329; Fax: +61 8 6488 1058; Email: georg.fritz@uwa.edu.au
†The authors wish it to be known that, in their opinion, the first two authors should be regarded as Joint First Authors.

ing to the –35 element (10), σ_2 is responsible for recognition and melting of the –10 element (11). In addition, both domains are connected via a highly variable linker, which has recently been shown to assist contact with RNAP similar to the $\sigma_{3.2}$ domain of group 1 σ^{70}s (12–14).

Given that bacteria typically contain several alternative σ factors, it is key that under non-stimulating conditions ECF activity is kept low, preventing undesired activation of their cognate target genes. The most common mode of ECF activity regulation is via the sequestration of the ECF by a transmembrane anti-σ factor, which keeps its cognate ECF in an inactive state unless the anti-σ factor undergoes stimulus-induced proteolytic degradation (3); however, a plethora of other ECF regulatory mechanisms exist, including sequestration by cytoplasmic anti-σ factors, undergoing stimulus-induced conformational changes (15,16), partner-switching mechanisms via phosphorylation of ECF-mimicking anti-anti-σ factors (17), C-terminal protein extensions fused to the ECF (18,19), direct phosphorylation of the σ factor by serine/threonine kinases (20), as well as control of ECF transcription via two-component systems (21,22).

Many of these regulatory mechanisms were first predicted by genomics approaches (reviewed in (7)). Indeed, research of the ECF family has been heavily dependent on bioinformatic tools from the date of their first description (23). ECFs are especially suitable for the application of prediction tools since they usually autoregulate their own expression, facilitating the identification of their target promoter. Moreover, they are usually co-encoded with regulators of their activity and/or with genes regulated by the ECF. Accordingly, the first bioinformatic classification of the ECF σ factor family grouped ECFs from <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 conserved genomic neighborhood (1). Altogether, this work proposed that it is possible to predict the contact of an ECF σ factor with its target promoter, its regulatory mechanism and its target genes from sequence information alone. Following studies expanded the number of phylogenetic groups by focusing on particularly ECF-rich bacterial phyla, such as Actinobacteria or Planctomycetes. Initially, nine planctomycetal (4) and 100 actinobacterial (5) genomes were analyzed for their ECF repertoire, again identifying correlations between primary protein sequence and function. More recently, a comprehensive study on the isolation, cultivation and genomics of 79 new planctomycetal species also lead to a significant expansion of the ECF diversity within this phylum. Almost 6000 ECFs were identified in 150 planctomycetal genomes with an average of 40 ECFs per genome. This diversity included 30 newly described ECF groups, including some with altogether novel signaling mechanisms (24).

While these initial ECF classification studies helped to understand the large diversity of ECFs across the tree of life, they so far addressed ECFs from a limited number of genomes and/or focused on specific phyla (1,4,5,24). Based on the relatively sparse sequence basis, some of the initially defined ECF groups featured natural limitations in that they either defined groups with only very few (<10) proteins—so-called 'minor' groups (ECF100, ECF102, etc.) (1)—or in that they clustered divergent sequences into groups that share only few unifying characteristics (e.g. ECF01, ECF10, ECF20) (1,4). However, the explosion of annotated sequences in databases, not only from re-sequenced species but also from new species of underrepresented phyla, enables a more comprehensive view of the ECF family.

In the light of the above-mentioned limitations of the initial studies, in this work we searched for ECF σ factors in all available genomes and metagenomes of the National Center for Biotechnology Information (NCBI) database, thereby expanding the number of ECF proteins 50-fold. We clustered the new ECFs into 2380 subgroups with a high degree of sequence conservation. Subgroups were further grouped into 157 ECF groups according to genetic context conservation and their putative mode of regulation. As a result, we defined 22 novel ECF groups with no significant similarity to previously described ECF groups. The conservation of the subgroups facilitated downstream *in silico* analyses such as prediction of conserved target promoter elements, conserved protein domains in the genetic neighborhood, and putative anti-σ factors. Even though the large amount of information collected for each of the 157 ECF groups only allows us to focus on the most interesting findings, we provide an extensive compendium of all the information gathered for each group in the Supplementary Material and via a newly developed web-platform, ECF Hub, which serves as a central community resource allowing researchers to browse our findings with additional visualizations, cross references, and statistics. This wealth of data represents a comprehensive resource to both computational and experimental researchers and helps guiding the characterization of ECF σ factors of unknown function.

MATERIALS AND METHODS

General bioinformatic tools

The data was processed with custom scripts written in Python when nothing else is stated (available upon request from the corresponding author). Multiple Sequence Alignments (MSAs) were generated by Clustal Omega 1.2.3 (25) and were visualized in CLC Main Workbench 7.7.2 (QIAGEN®). HMMER suite 3.1b2 (26) was utilized for generating and employing Hidden Markov Models (HMMs). HMMs were produced by hmmbuild (26) and proteins were searched against HMMs using hmmscan. The maximum-likelihood phylogenetic tree of ECF subgroups and bootstrap values were retrieved by IQ-TREE 1.5.5 with default options and automatic model selection (27). Phylogenetic trees were visualized using iTOL (28). Transmembrane helix predictions were carried out with TopCons (29) and PRED_TMBB2 for outer membrane proteins (30).

Nomenclature

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 (Figure 4). 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.

ECF Hub

To ease future research, we implemented the ECF Hub (https://www.computational.bio.uni-giessen.de/ecfhub) as a novel, centralized resource (i) to browse the extended collection of ECFs, (ii) to visualize individual ECF characteristics, (iii) to assign arbitrary user-submitted protein sequences to ECF groups and (iv) to provide a community-driven portal to collect and present the most recent findings about ECFs (see Supplementary Text S1 for details). Briefly, our data was transformed into an organized relational database scheme representing ECF groups, ECF subgroups, and ECF σ factors (Supplementary Figure S7). The ECF Hub web interface was established as the primary channel for human-computer and computer–computer communication to provide convenient access to the underlying ECF database. Moreover, to maintain and curate the underlying data by biological specialists with granted privileges, a user/admin interface was established. For the maintenance of the web application by bioinformatics experts, a command-line interface was realized. Since ECF σ factors are encoded in close vicinity of related proteins, the inclusion of genetic neighborhood information is a promising strategy to improve the extrapolation of protein functions and regulators of ECF activity. At the ECF Hub, we implemented a dynamic genome viewer, which retrieves and displays the genomic context information for ECF σ factors. For each feature in the context, detailed information, like annotated Pfam domains or gene products, is additionally provided via tooltips (Supplementary Figure S9). ECF Hub provides access to all the taxa and their ancestors of the NCBI taxonomy, where ECF σ factors could be identified. We included taxa from the ranks 'species', 'subspecies', and 'no rank' if there were genome assemblies with ECF σ factor sequences associated with them. Additionally, we included their ancestors with the following ranks: 'superkingdom', 'phylum', 'class', 'order', 'family', 'genus' and 'species'. For each taxonomic entity, a representative genome was determined (Supplementary Figure S8). The taxonomic identifiers were inherited from the original taxonomy by NCBI and were received via their E-utilities (https://www.ncbi.nlm.nih.gov/books/NBK25501/).

At the ECF Hub, scientists can do their own analyses for identifying and classifying protein sequences with (i) a one-click-solution on the ECF Hub website, (ii) using the ecf_classify tool hosted on GitHub (http://ecfclassify.computational.bio) and (iii) with HMMER and our supporting data hosted and citable at Zenodo (https://doi.org/10.5281/zenodo.3672544). The workflow for the first two options is described in Supplementary Figure S6.

We simplified the access to the hosted data by implementing multiple protocols for receiving data and files. First, processed data are accessible either via machine-friendly protocols like REST-based interfaces or can be downloaded in standard formats like JSON or CSV. Second, access to annotated protein sequences is provided in standard biological data formats, allowing users to obtain all assigned sequences either individually or for all ECF groups as a whole. For the conserved parts of ECF groups/subgroups and their promoter regions, HMMs are offered for download. Third, most of our figures and visualizations can be directly downloaded within the web interface. Furthermore, we offer sequence logos in high-quality PDF format for promoter regions as well as for all sequences of a group/subgroup. The MSAs were generated with MUSCLE and the sequence logos with Weblogo 3.0. With the implemented search interfaces, our users are able to fast lookup of arbitrary terms in the database: (i) A general search feature allows for querying, e.g. protein names or ECF group accessions in relevant database tables. (ii) An advanced search interface allows applying various filters to the database queries. The search results always redirect the users to content specific web pages within the ECF Hub, individually customized for the ECF researcher's demands. Moreover, we implemented the ECF Hub's TaXplorer (Taxonomy Explorer, Supplementary Figure S10) for browsing and exploring specific taxa, their lineage and their ECF associations in interactive file-explorer-like taxonomic trees. Users can inspect ECFs based on the taxonomic lineage of their origin organism and benefit from accumulated computations such as overrepresented phylum of an ECF group or the number of ECF sigma factors present in, e.g. the genus Bacillus. Via a search mask, an individual taxon can be selected as the root for a currently displayed taxonomic subtree. Subsequently, the displayed tree can be expanded or widened by selecting either the direct children or the parent of a taxon. By filtering for ECF groups, it is possible to display the selected ECF group distributions for all selected taxa, either for the representative or the non-redundant ECF σ factors. Furthermore, the taxonomic distribution of present ECF σ factors can be interactively browsed with sunburst plots. Finally, the ECF Hub provides a detailed inspection of genome assemblies. The relevant meta-information of an assembly is shown in a clear overview in addition to the present ECFs with their genetic neighborhood displayed in a genome viewer (10 genes up- and downstream of the target ECF).

For curating and discussing our database, the implemented community portal allows users to propose additional literature (Supplementary Figure S11), comment on main entities, and present their work on ECF σ factors in a blog-like stories section. Moreover, new database features can be proposed and publicly discussed. Curators, a small group of selected users, can modify the groups' descriptions, review and accept literature proposals, and attach references to related items (Supplementary Figure S12).

ECF identification and classification

One of the main computational foci of ECF research is the identification and classification of (new) protein sequences as ECF σ factor proteins. Our ECF Hub and the underlying software package ecf_classify now allow the discovery of potentially new ECF σ factors and the prediction of their functionality with regards to the new classification schema. As an initial step, we classify a sequence as ECF σ factor candidate based on the alignment with the general ECF HMM using the HMMER3 software suite (http://hmmer.janelia.org). For all sequences showing sufficient similarity to the generic ECF model, the classification is further verified by checking for the presence of the Pfam domains 'sigma2' (PF04542) and 'sigma4'/'sigma4_2' (PF04545, PF08281) and for the absence of Pfam domain 'sigma3' (PF04539). To further exclude sequences that might bear a cryptic σ_3 domain in between the σ_2 and σ_4 domains, a maximum distance of 50 amino acids is allowed between σ_2 and σ_4. A sequence is considered as an affirmed ECF if it passes all the above-mentioned checks (Supplementary Figure S6). Further analyses are subsequently performed for each confirmed ECF in order to assign it to the correct ECF group and subgroup. The conserved part of the sequence, i.e. the region covered by Pfam domains σ_2 and σ_4, is compared to the HMMs of all ECF groups and subgroups. Based on the HMMER bit scores, the best fitting ECF groups and subgroups are assigned: Initially, the number of ECF groups/subgroups is reduced to those which produce sufficiently proper alignments for the protein sequence. For this, the HMMER bit score is evaluated with two cut-offs for each ECF group and subgroup: the trusted and the noise cut-off. The trusted cut-off is the minimum bit score a confirmed member of an ECF group/subgroup has achieved, while the noise cut-off is the maximum bit score of all foreign ECF σ factors in the classification against that ECF group/subgroup. ECF groups/subgroups are afterward used in the second step if the bit score of the query protein sequence against it exceeds the trusted cut-off. If no group/subgroup exceeds its threshold, the noise cut-off is considered. For the second step, logistic regression curves are fitted in order to determine the probability that a protein belongs to an ECF group/subgroup. If there are probabilities higher than the ROC-optimized probability threshold, the ECF group/subgroup with the highest probability exceeding this threshold is assigned to the protein sequence (see Supplementary Figure S5). The statistical values required for the classification of new ECFs against ECF groups, subgroups and original groups are given in Supplementary Table S5. The described classification workflow is also provided for offline use; it can be downloaded as a command-line tool called 'ecf_classify' at http://ecfclassify.computational.bio or directly used as a Docker container.

RESULTS

Rationale of this study

Previous ECF classification efforts (1,4,5) were based on 495 genomes and identified a total of 3554 ECFs, which were classified into 94 ECF groups (summarized in (6)). Upon initiation of this work (February, 2017), the NCBI database contained 180 909 genomes, of which 4106 were bacterial genomes tagged as 'reference' or 'representative', ignoring GenBank assemblies when a RefSeq assembly was available for the same genome (see https://www.ncbi.nlm.nih.gov/refseq/about/prokaryotes/). This suggests that the presently available number and diversity of ECF sequences might be much larger than previously observed. To expand the library of ECFs, we here performed a comprehensive ECF search in the NCBI database and hierarchically reclassified the resulting proteins. First, we clustered protein sequences into fine-grained ECF subgroups with a high degree of sequence similarity, and then we aggregated subgroups into coarse-grained groups that share a common genetic neighborhood and a putative type of anti-σ factor. The similarity among the ECFs contained in groups allowed the identification of common putative target promoter motifs and ECF regulators. These hypotheses were confirmed whenever experimental reports on members of the group exist.

The number of ECFs is 50-fold larger than in the original ECF classification

To identify novel ECFs, we first extracted the sequences from all previous ECF classification efforts (1,4,5), aligned them and created a general Hidden Markov Model (HMM) for the ECF core region, including the linker between σ_2 and σ_4, but excluding any potential protein domains fused N- or C-terminally to the ECF (Figure 1A). To discriminate ECFs from other σ factors, we first scored this generic ECF HMM against two sets of training sequences—true ECFs from the original classification and a set of σ factors from groups 1, 2 and 3 that additionally contain domains σ_3 and σ_1 in some cases. This allowed us to define a threshold score that maximizes true positive ECFs (Figure 1B; *green*) while minimizing the number of false positive σ factors (Figure 1B; *red*). We then selected the non-redundant protein sequences from the NCBI database, for which the generic ECF HMM yielded scores higher than this threshold (Figure 1C). As further quality controls, we filtered for sequences containing the Pfam domains σ_2 and σ_4 but lacking the σ_3 domain, and discarded proteins with poorly defined amino acidic residues, such as X or J. This resulted in a library of 177 910 non-redundant ECF sequences. Some of the candidate ECFs included in this list clustered together with group 3 σ factors, indicating the presence of a cryptic σ_3 domain, which prompted us to remove them from the ECF library. This left us with 177 341 non-redundant ECFs, accounting for a ∼50-fold expansion over the original ECF classification (1,4,5) (Figure 1C). The full list of ECFs extracted during this study can be found in Supplementary Table S1.

Next, we analyzed the taxonomic origin of this expanded ECF library to determine the typical number of ECF numbers found in individual bacterial phyla. To enable such statistics, we focused on the subset of complete genomes of non-metagenomic origin, classified as NCBI 'reference genomes' or 'representative genomes', thereby mitigating bias towards heavily sequenced species. Analysis of 12 539 ECFs extracted from 1234 of these genomes showed that the taxonomic distribution of species became

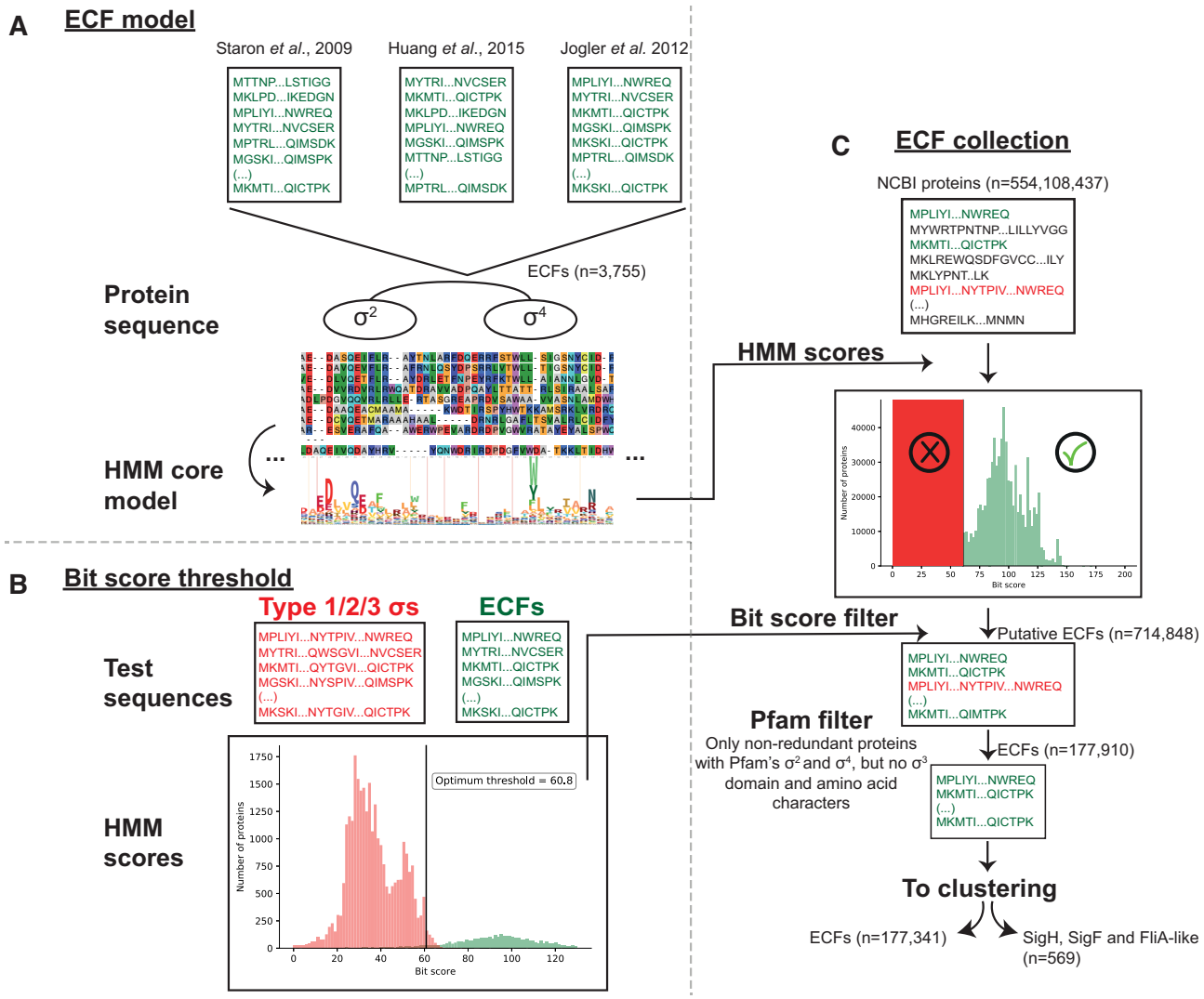

Figure 1. ECF retrieval pipeline. (**A**) We collected and aligned ECF sequences from previous classification efforts (1,4,5) and built an HMM from the area containing σ_2, linker and σ_4 regions. (**B**) In order to define a HMMER bit-score threshold for ECF extraction, we used the ECFs from (A) as positives and the σ factors containing a σ_3 domain in the Pfam database as negatives. We scored positives and negatives using the HMM model from (A) and derived a threshold that produced the largest specificity and sensitivity in the classification process. (**C**) We used the HMM model from (A) to score all proteins from NCBI as per February 2017, using as threshold the bit-score defined in (B). Putative ECFs without σ_2 or σ_4 domain, or with σ_3 domain, or proteins with characters that do not denote amino acids, were discarded. The final set of non-redundant ECFs includes 177 910 proteins.

more diverse than in the original classification efforts (Figure 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 came together with an increase in the number of species from underrepresented phyla, such as Bacteroidetes and Cyanobacteria (Figure 2A; *Genomes*). In addition, 19 new ECF-containing phyla emerged (Supplementary Table S3, *New phyla*). 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 (Figure 2A; *ECFs*). This is not the case for Cyanobacteria and Acidobacteria, which contribute a smaller percentage of ECFs than in the original library (Figure 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 libraries (1,4,5) to about ten ECFs per genome in the new library (Figure 2B). Confirming the findings of previous reports (1,5), the number of ECFs per genome is directly proportional to genome size (Supplementary Figure S1), although the average number of ECFs per genome depends on the phyla of origin (Figure 2B). Bacteroidetes and Actinobacteria have the greatest abundance of ECFs, with an average of 22.5 and 17.7 ECFs per genome, respectively (Figure 2B). Phyla with a lower abundance of ECFs

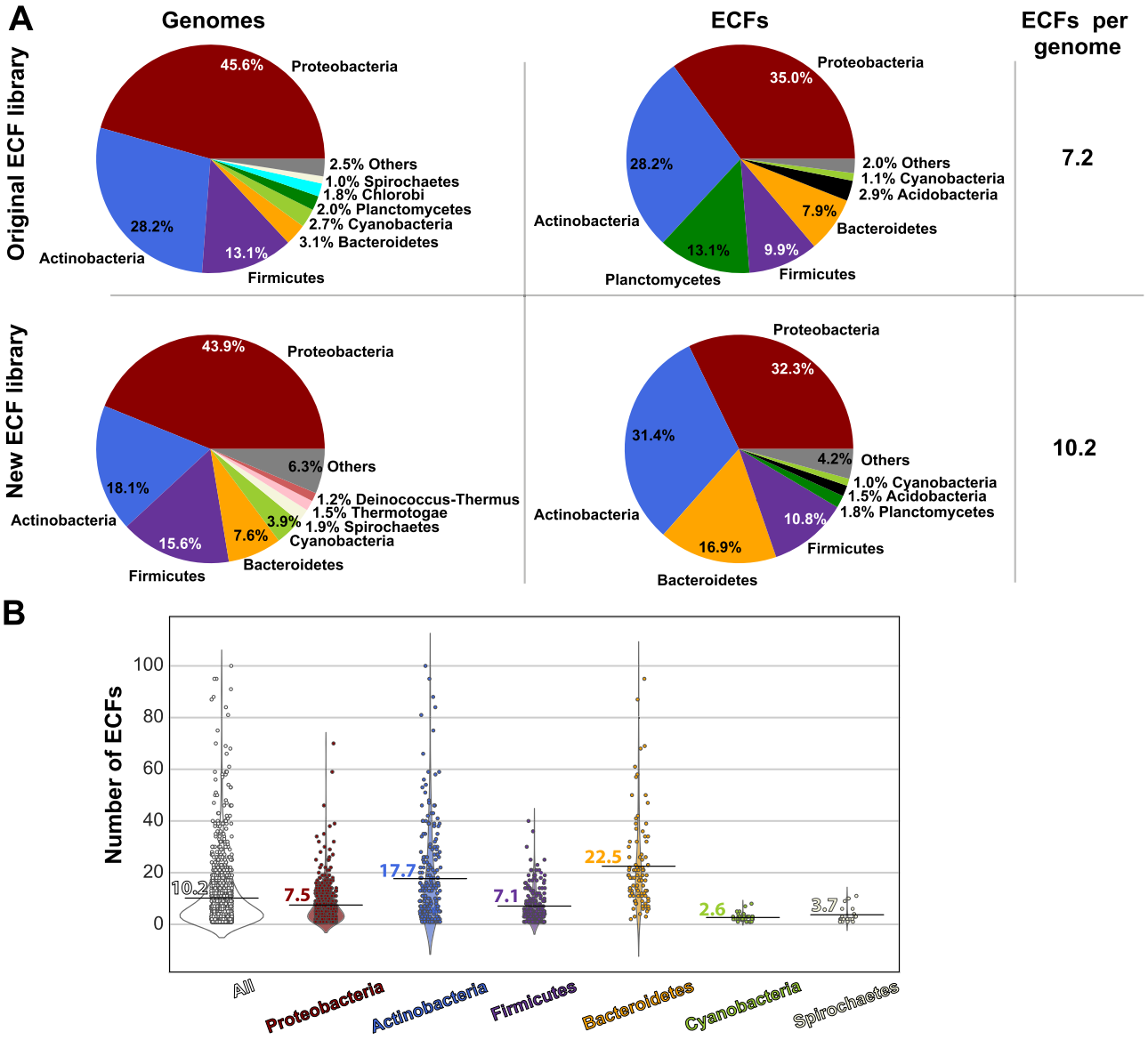

Figure 2. Taxonomic analysis of the ECF library. (**A**) Taxonomic composition of the input genomes, ECFs and average number of ECFs per genome in the original ECF classification (1,4,5) and in this work. For the data of this work, we only included ECFs and genomes from complete and non-metagenomic assemblies tagged as 'representative' or 'reference' in NCBI (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/), selecting RefSeq assemblies when both RefSeq and GenBank assemblies are available for the same genome. (**B**) Number of ECFs per genome for phyla with >20 complete genomes available. Average number of ECFs per genome is shown.

include Cyanobacteria and Spirochaetes, with an average of 2.7 and 3.7 ECFs per genome, respectively (Figure 2B). Firmicutes and Proteobacteria contain an intermediate number of ECFs, 7.1 and 7.5, respectively (Figure 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 (5).

ECF classification 2.0

The wealth of new proteins identified in the initial library expansion prompted us to reclassify ECF σ factors into groups according to protein sequence similarity. To this end, we first subjected the 177 910 protein sequences of the new ECF library to the rapid MMSeqs2 clustering algorithm (31), followed by a quality control that bisects the resulting clusters until the maximum pairwise *k*-tuple distance between sequences was ≤0.60 (Figure 3A, see Supplementary Text S1 and Supplementary Figure S4 for more details). Clusters with ≤10 proteins were discarded to ensure high sequence coherence within clusters, while preventing an explosion of small clusters with limited statistical relevance (Figure 3A). This procedure yields a total of 2380 ECF subgroups, which harbor a median of 22 non-redundant proteins (Figure 3D). Subgroups capture 77.3% of the proteins, similar to the statistics in the original classification (1) (Supplementary Table S1). Permutation tests on sub-

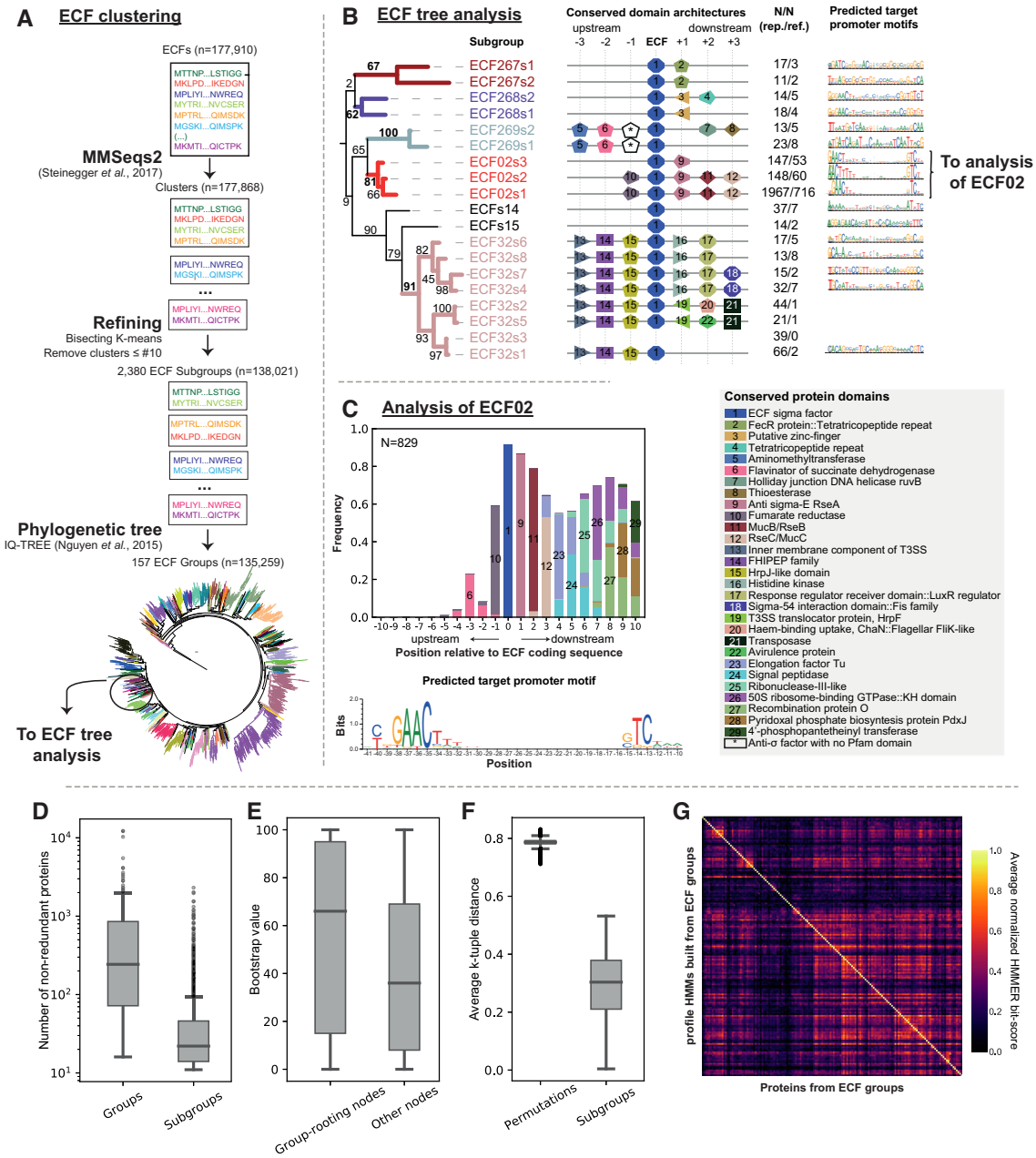

Figure 3. ECF clustering pipeline. (**A**) The ECF clustering pipeline starts with non-redundant ECF σ factor sequences stripped to their σ_2 and σ_4 domains, which were clustered using MMSeqs2 and refined using bisecting *K*-means until the maximum intra-cluster distance was ≤0.6. Subgroups with less than 10 sequences were not further considered. The consensus sequences of the resulting subgroups were hierarchically clustered, resulting in the ECF σ factor phylogenetic tree, which was used as the basis for the ECF group definition (see Supplemental material for details). (**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 (>50% conservation) in the genetic neighborhoods of subgroups that form monophyletic clades was used as a criterion for the ECF group definition. 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 Supplemental material. Subgroups with non-self-regulated ECFs do not feature a conserved promoter motif (see ECF32 description). (**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 >20% of the proteins encoded in a certain position are shown. Note that the architecture frequency might be underestimated due to the presence of higher scoring overlapping domains that interfere with the automatic domain identification (see Supplemental material: ECF group analysis). The predicted target promoter motif for ECF02 is also shown and has been confirmed for several members of ECF02 (see description of ECF02). (**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). See Supplementary Figure S2 for the complete version of this graph.

groups showed that the average *k*-tuple distance is significantly lower (two-tailed Student's *t*-test; *P*-value < 1e–16) in our subgroups as compared to random clusters of the same size distribution, indicating that subgroups are well defined (Figure 3F).

Then, we computed a phylogenetic tree based on the consensus sequence of each subgroup. This tree helps to identify the evolutionary relationship between the ECF subgroups (Figure 3A, bottom; Figure 4). As outgroups we included sequences with a low-scoring σ_3 domain, as well as the consensus sequence of all σ_3-containing proteins in Pfam, the latter of which we selected as root of the tree. Not surprisingly, proteins with a low-scoring σ_3 domain clustered at the base of the tree (Figure 4) and formed three groups with significant similarity to the sporulation σ factor SigF from Firmicutes and Actinobacteria, the flagellum biosynthesis σ factor FliA and the stationary phase σ 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 (ECF) σ factors (9) and account for the quality of our clustering approach. Other sequences with σ_3 domain remained unclassified (0.18% of the unclassified sequences).

To identify subgroups with common characteristics, we 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-σ factor type (Figure 3B and supplemental material for more details). As a result, 76.0% of the ECFs were captured in groups, displaying a median group size of 243 non-redundant proteins (Figure 3D). The assignment of ECFs to ECF groups and subgroups can be found in Supplementary Table S1. 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-σ factor RseA, and the regulators RseB and RseC, suggesting that ECFs in these subgroups feature the same mode of regulation as RpoE from *E. coli* (part of ECF02s1). Likewise, the subgroups aggregated into group ECF32 display strong conservation with a 2CS and a large number of genes encoding a type III secretion system (T3SS) (Figure 3B). 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 (1). 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 (Supplementary Table S2).

To provide a systematic overview on the conserved genomic context in each ECF group, we analyzed the frequencies of genes with a conserved protein domain architecture encoded up- and downstream of the ECF (Figure 3C). For group ECF02, for instance, this reveals that downstream of the regulators RseA-C there is enrichment of genes encoding translation regulators (e.g. EF-Tu) (Figure 3C). However, despite the overall conservation of the genomic context within an ECF group, we often find subgroup-specific traits with respect to the positioning and the specific type of conserved genes (Supplementary Text S2; Table S2), clearly indicating that the definition of ECF subgroups is highly relevant to the biological function of an ECF.

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 (1,32). We found overrepresented promoter motifs in many groups, e.g. ECF02, while others did not show significant motifs, e.g. ECF32 (Figure 3B and C), consistent with observations that the latter are not auto-regulated (21). 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 (Figure 3B). However, as for the conservation of the genomic context, we sometimes find subgroup-specific putative target promoters (e.g. in group ECF30 and others in supplementary text), 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 high bootstrap support scores at the rooting branches of ECF groups (Figure 3B and E), indicating that ECF groups are robust with respect to re-sampling. To further check the performance of the new classification approach, we tested whether HMMs built from ECF groups and subgroups were capable of faithfully classifying ECF sequences from their own groups (Figure 3G and Supplementary Figure S2). This showed that ECF proteins 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 that they allow for the classification of novel ECF σ factors in the future.

ECF classification 2.0 refines original and identifies novel ECF groups

As a proof of concept, we compared the original ECF classification and the classification presented in this work. To this end, we classified the new ECFs gathered here against the original classification. This showed a broad degree of correlation between the different classification approaches (Figure 4, ring #1). Accordingly, for these groups of high coherence we maintained the original group. 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 1 and Supplementary Figure S3).

One case of an extremely scattered original group is ECF01 (Supplementary Figure S3). This group was al-

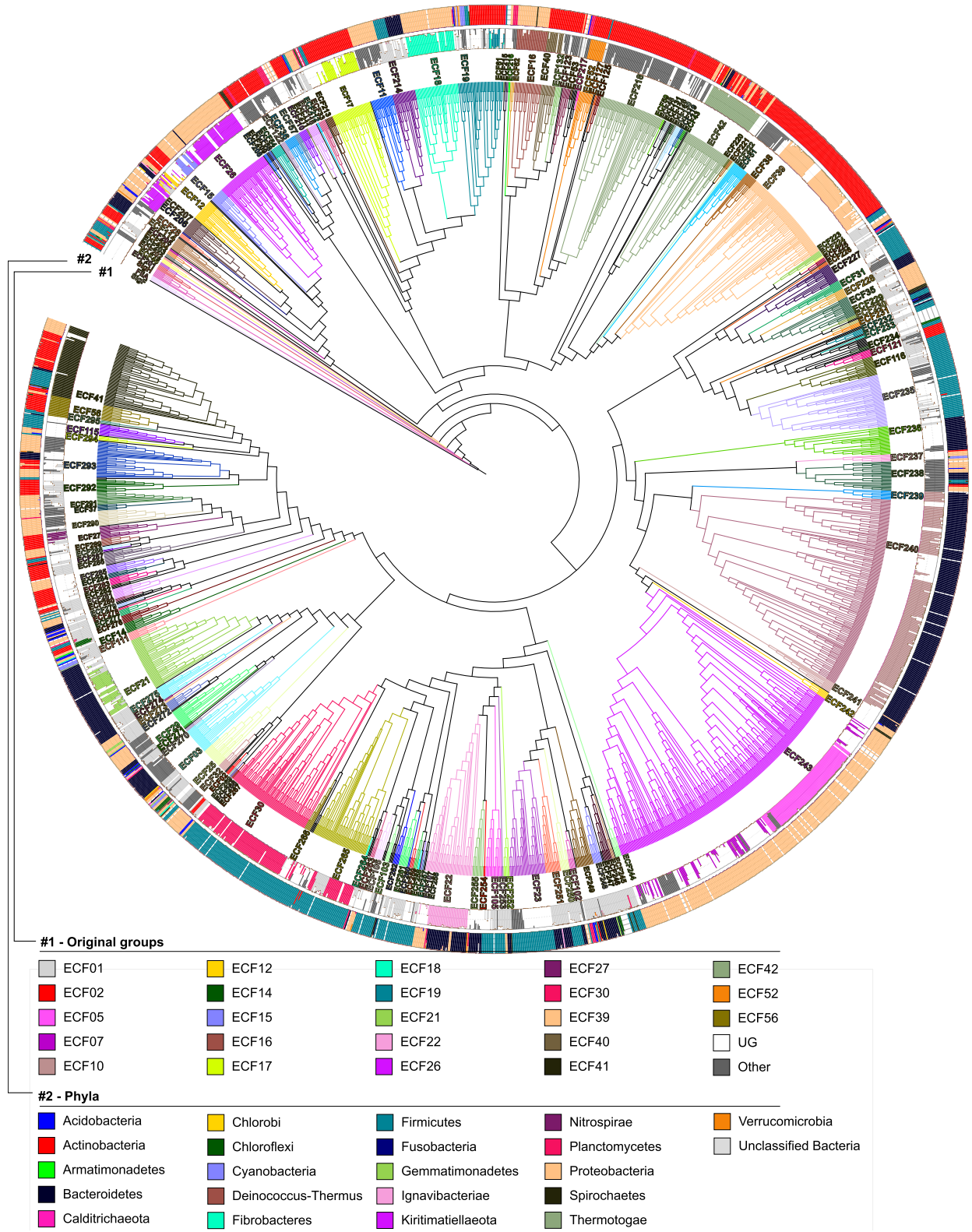

Figure 4. ECF σ factor tree. Phylogenetic tree of the consensus sequences of ECF subgroups. Clades are colored and named according to their group. Ring #1 shows which ECF groups the clusters would have been assigned to, if they were based on the original ECF classification. Original ECF groups with <1% sequences are shown under 'Other'. Ring #2 shows the phylogenetic origin of the majority of ECFs in a given subgroup.

Table 1. Rearrangements of original ECF groups. Equivalence between original and new groups is shown. Further information supporting this table can be found in Supplementary Figure S3. 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.

Rearrangments of original ECF groups	
Original ECF groups	**New ECF groups**
Merged (21)	
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
Small percentages (3)	
ECF04	ECF249
ECF113	ECF281
ECF119	ECF255
Ungrouped (5)	
ECF45	None
ECF60	None
ECF104	None
ECF109	None
ECF129	None
Scattered (3)	
ECF01	Many (see Figure S2)
ECF10	Many (see Figure S2)
ECF20	Many (see Figure S2)

ready considered highly diverse in the first ECF classification (1) and, based on the relatively unspecific HMM model of this group, it acquired more sequences in subsequent classification efforts (4,5). As a result, we did not consider the proteins from ECF01 for the nomenclature of the ECF groups in this work. Another highly scattered original group is ECF20 (Supplementary Figure S3). ECF20 is present in four main groups of our classification: ECF281, ECF289, ECF290 and ECF291 (Supplementary Table S2). 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) (33). Nevertheless, the anti-σ factors encoded in the genetic context of members of these groups differ. ECF281 features a zinc finger-containing anti-σ factor downstream of the ECF coding sequence (position +1), while in the case of ECF289 this protein contains a DUF3520 domain fused to a von Willebrand factor; ECF290 contains a RskA-like anti-σ factor, and, lastly, ECF291 contains a CnrY-like anti-σ factor in position -2 (Supplementary Text S2, Table S2). Based on this anti-σ factor diversity, it seems likely that the 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, only groups ECF239 and ECF240 receive most of the proteins of the original ECF10. 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 (1).

The high occurrence of new groups that combine several original groups is probably due to the incorporation of new sequences that bridge previously isolated ECF groups. Indeed, this possibility was considered in the original ECF classification (1). One example of a merged group is ECF243, which constitutes the largest group of the new classification and contains the proteins previously associated to original FecI-like groups ECF05 to ECF09 (Supplementary Figure S3). Another example of a merged group is ECF238, which contains sequences from the original groups ECF24 and ECF44 (Supplementary Figure S3, Table S2). Members of ECF238 contain a cysteine-rich C-terminal extension of approximately 20 amino acids (Supplementary Text S2), 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) (34).

What is likely the most interesting contribution of the new classification are the entirely new groups. We found 22 new groups that could not be assigned to any original group (Table 2). From the 16 new groups with 10 or more proteins from representative/reference organisms, 10 share a conserved genetic neighborhood with putative anti-σ factors. A special case of these is ECF241, which is located in the FecI-like clade and represents an evolutionary intermediate between groups ECF240 (derived from original ECF10) and the iron uptake FecI-like group ECF242. Instead of the canonical FecR-like anti-σ factor from FecI-like groups, members of ECF241 are encoded in proximity to a conserved two-transmembrane helix protein that in some cases hits the Pfam model for heavy-metal resistance proteins (Pfam: PF13801). The N-terminus of this protein is too short to feature a typical ASD. However, a MSA of this protein with the ASDs of canonical FecR-like anti-σ factors revealed that a putative, divergent ASD might be located in the C-terminal cytoplasmic part of the conserved protein. To our knowledge, this is the first time an anti-σ 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 (Supplementary Text S2, Table S2). 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 below.

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, we described the new findings concerning the 22 new ECF groups with no significant 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-

Table 2. Description of 22 new groups that have no sequences with significant similarity to any original group. The description shows the number of non-redundant ECFs, the number of ECFs from organisms tagged as 'representative' or 'reference' in NCBI (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/), excluding GenBank assemblies when an equivalent RefSeq assembly exists, the taxonomic origin where members of each group are present, their putative regulator and other special traits. Taxonomic origin of groups where no representative/reference members are available are marked with '-'. Groups were regulators were not

New groups with no homology to any original ECF group					
ECF group (22)	No. of non-redundant proteins	No. from rep/ref organisms (*)	Taxonomic origin	Regulator	Special traits
ECF201	69	13	Firmicutes (100%)	-	Closest ECF group to type III sigma factors
ECF202	35	none	-	-	
ECF208	79	25	Spirochaetes (100%)	Putative anti-sigma factor	Associated to glycosyl transferases fused to IDEAL domains
ECF210	139	5	Proteobacteria (100%)	-	
ECF215	49	none	-	-	
ECF216	46	13	Proteobacteria (100%)	Putative anti-sigma factor	
ECF219	88	20	Actinobacteria (100%)	Putative anti-sigma factor	Lack of sigma2.1 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 AS 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 *et al.*, 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	

(*) Representative and reference organisms are defined by NCBI (https://www.ncbi.nlm.nih.gov/genome/browse#!/prokaryotes/). Only RefSeq assembly is considered if both RefSeq and GenBank assemblies are available for the same genome.

ure 5 and the full description of the groups is available in Supplementary Text S2.

ECF σ factors feature diverse, often multi-layered, modes of regulation

Given the large diversity of the ECF σ factor family, it is essential to focus on individual groups in order to extract conclusions concerning their biological function, regulation and DNA binding site. Genetic neighborhoods of ECFs typically contain anti-σ factors. However, other regulatory elements might be substituting them, ranging from fused C-terminal extensions, to two-component systems and serine/threonine protein kinases (1,7,20). Here, we provide an overview of the different modes of regulation present across the ECF groups. Their comprehensive description can be found in Supplementary Text S2 and in Supplementary Table S2.

Most of the ECF groups (114 out of 157) contain a putative anti-σ factor, as defined by (i) the presence of Pfam domains of known anti-σ factors, (ii) detectable similarity to anti-σ factors of the original classification (1) and (iii) presence of transmembrane helices (see Supplemental material for details). This anti-σ factor is typically encoded in position +1 from the ECF coding sequence. A list of putative anti-σ factors identified in this study can be found in Supplementary Table S4. In most of the cases, the putative anti-σ factor does not match any Pfam domain of experimentally addressed anti-σ factors. In order to decipher common types of anti-σ factors present across the ECF tree, we built HMMs from their conserved cytoplasmic area. Searching the proteins encoded 10 CDSs up- and downstream of the ECF coding sequence, we found that most of the anti-σ

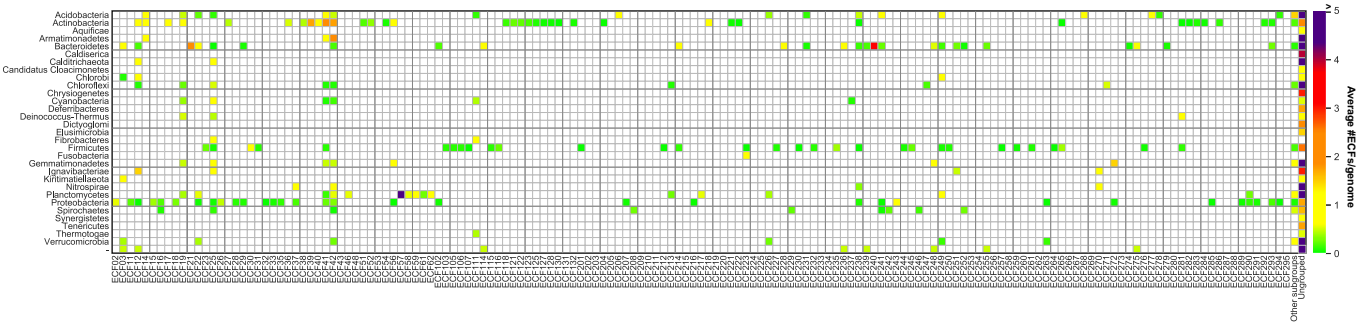

Figure 5. ECF abundance in different phyla. The heatmap shows the average number of ECFs from a certain ECF group in a certain phylum. We also show ECFs that are grouped against subgroups that are not part of groups and ECFs that remain ungrouped. Underrepresented phyla are rich in the latter category. These values were calculated using the set of ECFs present in complete, non-metagenomic genomes from 'reference' or '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 '-'.

factors are group-specific (Figure 6F). This agrees with the orthogonality observed between anti-σ factors of different groups (32). Exceptions are the clade that contains ECF222, ECF51, ECF38 and ECF39, groups regulated by FecR-like anti-σ factors (ECF239, ECF240, ECF242 and ECF243), groups regulated by RskA-like anti-σ factors, anti-σ factors with a putative zinc-finger, and anti-σ factors with a DUF4179, almost exclusively present in Firmicutes (Figure 6F). The number of transmembrane helices on putative anti-σ factors is usually one (82 ECF groups), followed by two (14 groups), six (five groups), four (three groups) and three (one group) (Figure 6E). Anti-σ factors predicted as soluble are present in 10 ECF groups (Figure 6E). However, given that our analysis can only identify soluble anti-σ factors with detectable similarity to described anti-σ factors, it is likely that other soluble anti-σ factor variants may exist. Additionally, it is not guaranteed that all the new putative anti-σ factors function as such, given the vast diversity, lack of sequence conservation and lack of studies confirming their function.

ECF107 contains two putative anti-σ factors, which could be part of the same protein complex or compete for binding the ECF (6), thereby illustrating the complexity and diversity of anti-σ factor mediated regulation. Another interesting example is ECF102. The only described member of ECF102, SigX from *P. aeruginosa*, has been suggested to have a role in mechanosensing (35). SigX is part of a seven-gene operon which includes a mechanosensitive ion channel (CmpX), a putative anti-σ factor (CfrX) and an outer membrane porin (OprF) (35). Even though original reports hypothesized that the regulation of SigX is carried out by the putative anti-σ factor CfrX (6), new reports suggest that its regulatory mechanism is more complex and involves also CmpX and OprF (35). We observed that these proteins are conserved in ECF102s1. Moreover, the mechanosensitive ion channel is conserved in ECF102s2 and ECF102s5. A similar case, in which proteins in addition to the anti-σ factor are required for ECF regulation, is ECF31. The only characterized member of ECF31, SigY from *B. subtilis* (subgroup 1), requires both the anti-σ factor YxlC, encoded in +1, and YxlD, encoded in +2, for its regulation, presumably forming a protein complex with the ECF (36). YxlCD homologs are conserved across ECF31.

The second most common regulatory mode of ECFs is the presence of C-terminal extensions (19 groups) (Figure 6B), which is typically correlated with the lack of putative anti-σ factors (Figure 6F). 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 (37). An interesting exception is ECF205, which also has a SnoaL-like extension but is located in proximity to the base of the ECF tree (Figure 4), indicating that more factors, in addition to its C-terminal extension, determine the sequence conservation of this group. We newly 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 50 aa, but we found that some groups contain short C-terminal extensions that are difficult to identify when only looking at average 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 (∼20 aa) C-terminal extension and also in the linker in some instances. One of the described members of ECF238, CorE2 from *Myxococcus xanthus* (ECF238s15), is known to be activated by Cd^{2+} and Zn^{2+} via this cysteine-rich C-terminal domain (34,38). Another characterized member of ECF238 is SigZ from *Bacillus subtilis* (subgroup 9). SigZ is not regulated by any anti-σ factor and the studies about its function are very limited (39). However, association of SigZ with group ECF238 suggests that the two cysteines present in its C-terminus could have a functional role. Another group with

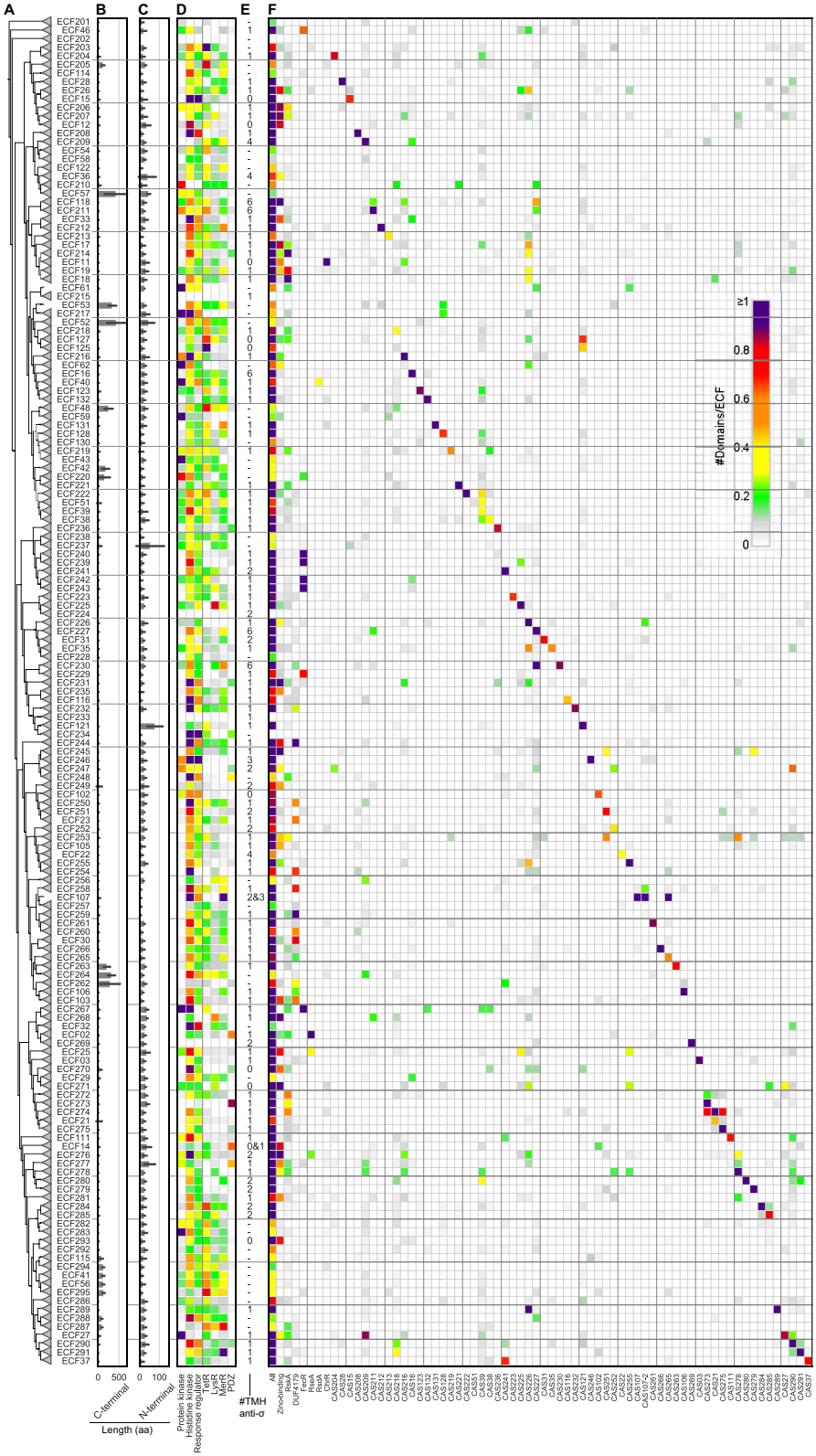

Figure 6. Genetic context analysis of ECF groups. (**A**) Schematic representation of the ECF σ factor tree. (**B, C**) Bar plot with the average length after σ_4 domain (C-terminal) or before σ_2 domain (N-terminal), respectively. Error bars indicate standard deviation. (**D**) Average number of regulatory domains in genetic neighborhoods per ECF. (**E**) Number of predicted transmembrane helices of the putative anti-σ factor encoded in the genetic neighborhood of groups. (**F**) Average number of anti-σ factor domains per ECF, predicted in the genetic neighborhood of ECF group members.

a short C-terminal extension is ECF29, which contains a conserved RCE/D motif in its extra ∼30 aa and lacks any putative anti-σ factor. Unfortunately, no member of ECF29 has been experimentally addressed.

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 (Figure 6C). This extension has been studied in BldN (subgroup 1) from *Streptomyces coelicolor*, where it is proteolytically degraded in order to yield its mature ECF, which then is subject to anti-σ factor regulation (40). Subgroups from several groups contain N-terminal extensions (Supplementary Table S2). For instance, members of ECF36s4 lack a discernable anti-σ factor and their N-terminal extension has been proposed to inhibit DNA contact in the uninduced state in SigC from *M. tuberculosis* (41). Alternatively, the N-terminal extension of two members of ECF12s1, σ^R from *S. coelicolor* and SigH from *Mycobacterium smegmatis*, generates an unstable isoform produced from an earlier start codon upon exposure to thiol oxidants (42). This makes σ^R susceptible towards σ^R-activated ClpP1/P2 proteases and thus implements a negative feedback loop that contributes to turning off the stress response (42).

Other putative regulators of σ factor activity that we often found in the conserved genetic neighborhood of ECFs were serine/threonine protein kinases (Figure 6D). ECF σ factors of five original groups have been hypothesized to be directly phosphorylated by a protein kinase (ECF43 and ECF59-ECF62 (1,4)). We added to the list of protein kinase-associated groups ECF217, ECF267 and ECF283 (Figure 6D). 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. One reason could be the divergent σ_2 domain observed in members of this group (20). Protein kinase-related ECF groups typically lack co-encoded anti-σ factors (Figure 6D), consistent with the notion that direct phosphorylation of the σ factor regulates their activity (20). The only exception is group ECF267, which may be regulated by a putative FecR-like anti-σ factor with tetratricopeptide repeats. Given that ECFs from group ECF267 are very distant from the FecI-like clade (ECF239–ECF243) (Figure 4), it seems possible that this putative anti-σ 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 σ factor. Whether the anti-σ factor and/or the protein kinase regulate the activity of members of ECF267 is unclear.

Four groups contain 2CSs in their genetic context. These regulators can co-occur in conjunction with anti-σ 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 2CSs. On one hand, members of ECF15 are regulated by a partner-switching that involves an anti-anti-σ factor fused to the response regulator of a 2CSs (17). Instead, the response regulator of members of ECF246 is fused to a transcriptional regulator, suggesting that either the putative anti-σ factor or the 2CS regulate ECF246 activity. For members of ECF32, it was shown that their 2CS indirectly regulates their transcription (21,43,44). In the case of ECF39, a 2CS is directly regulating their transcription (45,46). Similar direct or indirect regulation of the 2CS over the ECF expression could also occur in members of ECF234, given the absence of a putative anti-σ factor and the fusion of the response regulator to a transcriptional regulator. The physiological function of ECF234 could be related to an ABC transporter present in its genetic context.

Some ECF groups contain conserved transcriptional regulators in their genetic contexts, such as TetR-like repressors, which appear in groups with anti-σ factors (ECF125) and, remarkably, in ECF203, which lacks any obvious regulator (Figure 6D). 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 judging by the lack of a conserved predicted target promoter motif (Supplementary Table S2). Other transcriptional regulators include LysR- and MerR-like repressors, which appear in several ECF groups associated with anti-σ factors (Figure 6D).

A total of 16 ECF groups are not linked to any of the above-mentioned regulators (Figure 7), inspiring us to predict novel, putative regulators of ECF activity. So far, only three of these 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 2CS PorXY (47). It is possible that other members of ECF228 have a similar regulation. In the case of the novel group ECF282, σ^{AntA} from *Streptomyces albus* (ECF282s2) is regulated at the level of transcription and might be target of ClpXP proteolysis (48). Indeed, homologs of σ^{AntA} have been considered a new group of ECF σ factors that control the expression of antimycins (48). However, the C-terminal AA dipeptide, suggested as target of ClpXP proteolysis (48), is only present in members of subgroup 2. In ECF114, SigH from *Porphyromonas gingivalis* (ECF114s4) is induced upon exposure to O_2 and promotes aerotolerance and hemin uptake (49). SigH has been speculated to be regulated at a transcriptional level (49).

Given that the genetic neighborhood of the remaining 13 groups does not feature canonical ECF regulators (Figure 7), we speculated about their putative function. 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, and both options should be considered in downstream experimental analyses. Interestingly, we found new putative regulators/targets of regulation of the original groups ECF54 and ECF130. ECF54 is encoded in close proximity to a protein with a 4Fe–4S cluster, whereas ECF130 is encoded in proximity to a helix-turn-helix (HTH) containing protein, which could be involved in the transcriptional control of members of ECF130 (Figure 7). Similarly, members of ECF201 are encoded in proximity to HTH proteins (Figure 7). Interestingly, members of ECF237 share genetic neighborhood with several 'killing

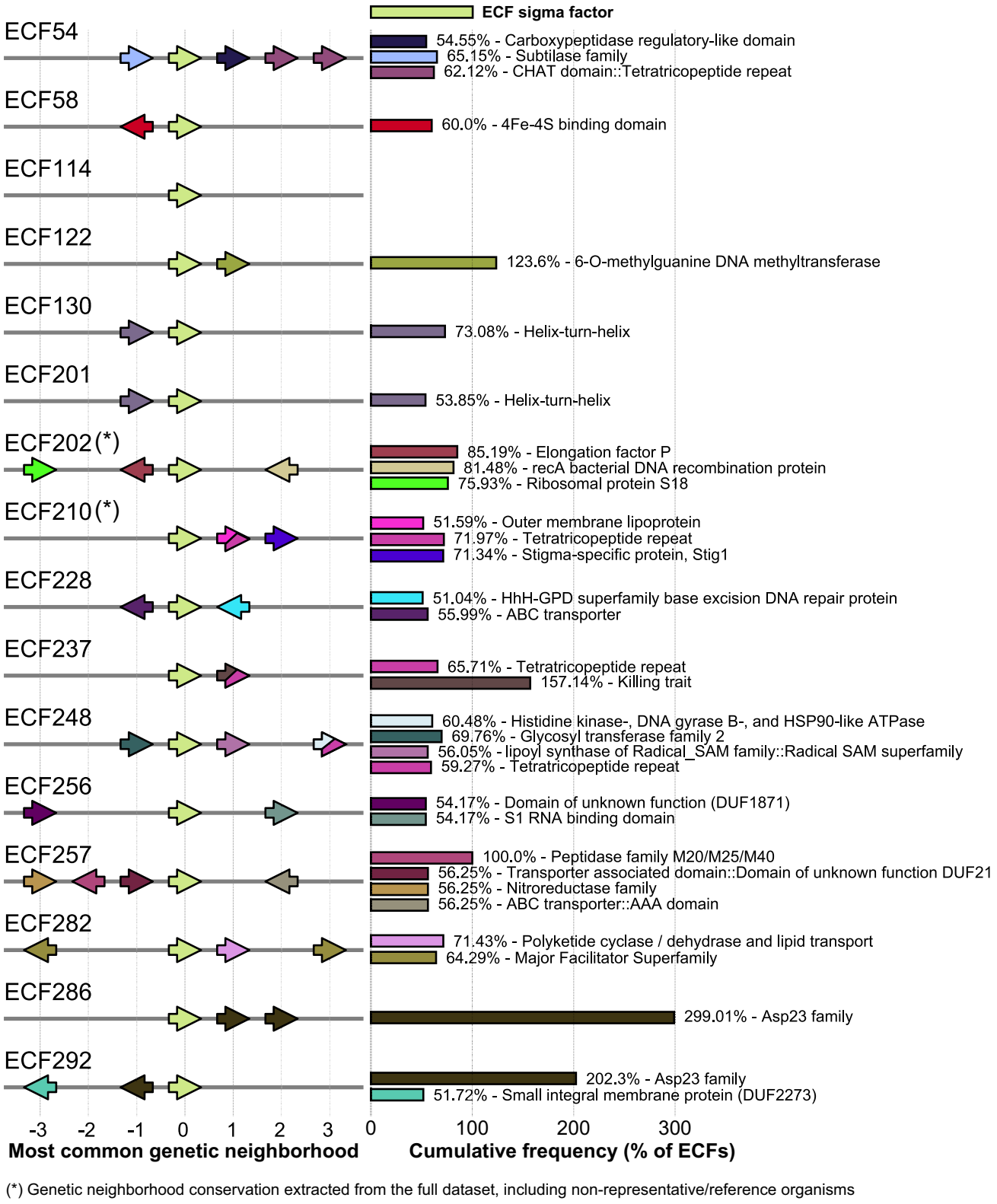

Figure 7. Genetic neighborhood of ECF groups that lack a canonical regulator. The left side shows the typical positions of genes encoding a certain protein domain architecture (present in >50% of the genetic contexts). Only positions ±3 from the ECF coding sequence are displayed. The direction of the 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. Only for groups ECF202 and ECF210 (marked with stars) sequences deriving from non-representative organisms were included, since these ECF groups contain less than 10 proteins in representative organisms.

trait' proteins with homology to RebB (Pfam: PF11757) (Figure 7). Given the absence of conservation for the rest of the proteins from the R-body, the cellular structure that RebB ensembles (50), and the presence of several copies of RebB, it is possible that these proteins have an alternative function, not related to R-body assembly. Lastly, members of ECF286 and ECF292 share genetic neighborhood with several copies of Asp23 proteins (Pfam: PF03780) (Figure 7). Asp23 is one of the most abundant proteins of *Staphylococcus aureus* and its deletion leads to upregulation of the cell wall stress response (51). Therefore, Asp23 proteins could be acting as a new type of anti-σ factor that regulate the activity of members of ECF286 and ECF292.

The ECF Hub portal enables convenient access to the novel ECF classification

The ECF Hub web portal (Figure 8) was developed as a central resource facilitating convenient access to the new ECF classification scheme (https://www.computational.bio.uni-giessen.de/ecfhub). It enables easy access to the preexisting data enriched with additional contextual information as well as supports the classification and assignment of user-supplied sequences. At the ECF Hub, scientists can inspect and examine the distribution of ECF σ factors in different taxa, or get insights into the taxonomic distribution within a certain ECF group. The ECF Hub provides robust search interfaces for easy access to all stored data, as well as the possibility of exporting the analysis results in a variety of standard formats or images. The ECF classification process can be performed directly on the ECF Hub web page, or, e.g. for large amounts of input sequences or confidential data, with the ecf_classify tool (http://ecfclassify.computational.bio), which supports reproducible offline use. The underlying models and supporting files are permanently available at Zenodo (https://doi.org/10.5281/zenodo.3672544). The resulting annotations are comparable or even superior to those obtained by the former classification tool ECFfinder (Figure 9).

DISCUSSION

This work unifies, 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, are further hierarchically clustered into 157 ECF groups, defined on the basis of a common genetic neighborhood, which suggests a similar mode of regulation. As part of the *in silico* characterization of ECF groups, we predicted their putative regulators, their target promoter motifs and their most likely function (Supplementary Text S2 and Table S2). These predictions are biologically meaningful in that they correctly reflect results of experimentally studied members, whenever available. We additionally developed the ECF Hub as a supporting platform. The ECF Hub allows users to browse these data and allows them to analyze their own ECFs. 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.

The new ECF groups are monophyletic clades of the ECF phylogenetic tree, that are 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 (4,5) and in two phyla, Actinobacteria and Planctomycetes, respectively. Within our ECF tree, these two groups are direct neighbors with a bootstrap support value of 17, indicating a significant protein similarity between them. None of them has a putative anti-σ factor or any other clear regulator, and they contain different elements in their genetic context (Figure 7). These results suggest that ECF58 and ECF54 have the same origin, but they evolved independently in Actinobacteria and Planctomycetes, acquiring the regulation of 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. Moreover, the great enrichment in phylogenetically diverse proteins allows for the application of *in silico* prediction tools for individual groups. These types of analyses are only possible in large protein datasets with enough protein diversity, as shown for groups ECF41 and ECF42 (37).

As part of the description of ECF groups, we analyzed their most likely regulators and the types of putative anti-σ factors encoded in their genetic neighborhood (Figure 6). Most of the predicted anti-σ factors are highly specific for their own groups (Figure 6F). 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-σ factor sequences. The general lack of the same type of anti-σ factors in neighboring groups reflects their large diversity and their specificity, which has been exploited for the construction of orthogonal genetic circuits (32). Anti-σ factors are not the only genes conserved in the genetic context of ECF σ factors. In this study, we identified the ECF groups associated to other known ECF regulators such a C-terminal and N-terminal extensions, two-component systems, serine/threonine kinases (7), and other regulators such as TetR repressors (Figure 6).

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-σ factors and N-terminal extensions, some members of ECF12 are regulated by both anti-σ factors and alternative promoters that generate an unstable longer versions of the ECF σ factor (42) and members of ECF18 and ECF19 are not only regulated by RskA-like anti-σ factors, but also by a pair of conserved cysteine residues known to form a disulphide bridge that senses oxidative stress in SigK from *M. tuberculosis* (ECF19s1) (52). While these regulatory layers have only been deciphered for a few well-studied ECF σ factors, they might point towards a broader means of regulation also implemented in additional ECF groups. For instance, several ECF groups feature conserved cysteine residues potentially able to form disulphide bridges (Supplementary Table S2), and members of ECF267 contain both a FecR-like anti-σ factor and a conserved protein kinase in their genomic neighborhood. Given their multi-layered regulation, abundance and diver-

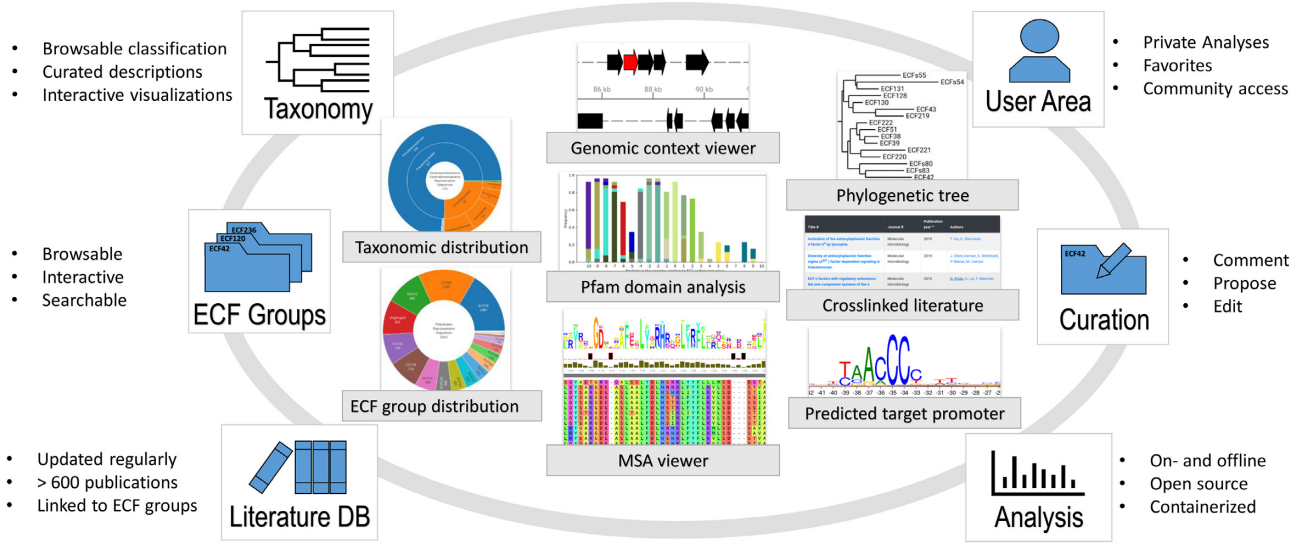

Figure 8. Schematic overview of the ECF Hub's capabilities. ECF Hub enables access to novel classification, which can be interactively explored based on taxonomy, ECF groups and literature. For this purpose, a variety of high-quality visualizations and statistics are provided. With the ECF Hub, scientist can upload and process own protein sequences in order to detect and classify ECFs. Moreover, the ECF Hub serves as a collaborative platform where users are able to comment on existing content or propose new features. Finally, registered users have their own private area for analyses, favorites, and community interaction.

sity, we suggest that ECF σ factors have higher signal integration capabilities than previously anticipated.

With an average of approx. 10 ECFs per genome, these regulators are more abundant than previously thought (1). Confirming previous reports (1,5,53), we find that the number of ECFs is proportional to genome size (Supplementary Figure S1), with species thriving in diverse environments typically featuring larger genomes that provide them with the ability to sense and respond to a 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 (53). 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 ECF groups acquired exclusively in *S. cellulosum* So0157–2, we found ECF03 (one extra member), ECF26 (one extra member), ECF41 (two extra members) and ECF56 (one extra member). 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 (Supplementary Table S2), and could have been acquired by horizontal gene transfer. However, this protein is not overexpressed under alkaline conditions (53), 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 (53) 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.

The search of ECF σ factors presented in this work has some limitations related to the quality filters that we applied during the ECF retrieval. These filters limit the diversity of the extracted sequences, while ensuring that the collected proteins function as real ECF σ factors. In particular, we noticed that two main types of ECF σ factors could not be captured, namely, ECF σ factors from phages and ECFs whose conserved σ_2 and σ_4 domains are divergent. σ factors of phage origin have been described in literature (see review (54)); nevertheless, they are usually divergent from traditional σ factors (54), in some cases incorporating alternative domains replacing σ core domains. For instance, in *Bacillus* phage vB_BceM-HSE3, the ECF Gp17 contains a double zinc ribbon domain (Pfam: PF12773) in the position of the σ_2 domain, while the generic Pfam domain for σ_4 is not found at all (55). Similarly, the σ factors Gp01 and Gp103 contain only σ_2 domain or no Pfam domain, respectively (55). Another reason for the lack of phage proteins in the present work is that viral genomes are usually not annotated in NCBI (56) and did not enter the ECF search in most of the cases. Other types of ECF-like σ factors not included in the current version are ECFs whose σ_4 (e.g. SigI from *Bacillus subtilis*) or σ_2 domain (such as VP0055 from *Vibrio parahaemolyticus* or ComX from *Streptococcus pneumoniae*) do not hit their Pfam models. A special example of this are σ^I-like ECFs, which contain a $\sigma_{I\text{-}C}$ domain instead of a canonical σ_4 domain (57). These ECFs are involved in the synthesis of components of the cellulosome in cellulolytic clostridia (57). Attempts to classify these proteins against the current ECF classification were unsuccessful. The group with the highest probability of containing σ^I-like ECFs is ECF201 (probability = 1.12e–19), the outermost group of the ECF classification, indicating that σ^I-

			ECF Hub		ECFfinder	
RefSeq Assembly Accession	**Species (Strain)**	**Phylum**	**Identified**	**Classified**	**Identified**	**Classified**
GCF_000005845.2	Escherichia coli (str. K-12 substr. MG1655)	Proteobacteria	2	100,00%	2	100,00%
GCF_000006765.1	Pseudomonas aeruginosa (PAO1)	Proteobacteria	19	100,00%	19	89,47%
GCF_000006965.1	Sinorhizobium meliloti (1021)	Proteobacteria	11	100,00%	11	90,91%
GCF_000007565.2	Pseudomonas putida (KT2440)	Proteobacteria	19	100,00%	19	89,47%
GCF_000007985.2	Geobacter sulfurreducens (PCA)	Proteobacteria	1	100,00%	1	100,00%
GCF_000009365.1	Alcanivorax borkumensis (SK2)	Proteobacteria	2	100,00%	2	100,00%
GCF_000011705.1	Burkholderia mallei (ATCC 23344)	Proteobacteria	11	100,00%	11	100,00%
GCF_000018865.1	Chloroflexus aurantiacus (J-10-fl)	Chloroflexi	9	66,67%	9	88,89%
GCF_000064305.2	Flavobacterium psychrophilum (JIP02/86)	Bacteroidetes	7	100,00%	7	71,43%
GCF_000146165.2	Shewanella oneidensis (MR-1)	Proteobacteria	5	100,00%	5	100,00%
GCF_000184745.1	Variovorax paradoxus (EPS)	Proteobacteria	17	100,00%	18	72,22%
GCF_000195955.2	Mycobacterium tuberculosis (H37Rv)	Actinobacteria	10	100,00%	10	80,00%
GCF_000215745.1	Klebsiella aerogenes (KCTC 2190)	Proteobacteria	3	66,67%	3	33,33%
GCF_000238395.3	Pseudoalteromonas arctica (A 37-1-2)	Proteobacteria	8	87,50%	9	55,56%
GCF_000318015.1	Bordetella bronchiseptica (253)	Proteobacteria	12	100,00%	12	83,33%

Figure 9. Comparison between ECFfinder and ECF Hub assignment for selected genomes. Selected genomes were processed with the ECFfinder website and the ECF Hub classification tool. Left: ECF predictions obtained from ECFfinder and ECF Hub are generally in accordance. Right: The ECF Hub, which incorporates the new classification scheme, enables a larger fraction of ECFs to be classified.

like ECFs are distant from canonical ECFs and might have evolved in parallel to them from group 3 σ^{70} factors. Lowering the stringency in our extraction pipeline would allow to study these non-canonical ECFs in further depth. The strategy to follow could be similar to work made on ECF43, the group of VP0055 from *Vibrio parahaemolyticus*, which was expanded from the 69 non-redundant proteins identified in this work, to more than 900 members of ECF43 (20). Likewise, a similar approach could be used to identify and expand knowledge around bacterial ECF factors with split domains, such as *B. subtilis* SigO-RsoA, which contain the σ_2 and σ_4 domains in two separate polypeptides (58,59).

The average number of identified ECF σ factors per genome varies for different groups and bacterial phyla (Figure 5). In this analysis it turned out that the number of unclassified ECFs per organism is larger in bacterial phyla underrepresented in biological databases (Figure 5). In the future, clustering strategies could specifically target these proteins, which are likely too diverse and scarce to be clustered with the currently available dataset. We also found that some ECF groups are particularly enriched in certain phyla. For instance, we observed an average of 5.3 copies of ECF57 per planctomycetal genome, and an average of 3.4 copies of ECF240 per Bacteroidetes genome (Figure 5). In these cases, a question that remains unsolved is 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 group ECF10, have been associated to carbohydrate scavenging in Bacteroidetes (1). Even though no member has been experimentally addressed to date, it is possible that the different members of ECF240 present in the same genome are involved in the uptake of different carbohydrates. A similar case occurs in the proteobacterial group ECF243, which merges original FecI-like groups ECF05–09, and is in charge of iron uptake (1,60). We 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 in 3.4 proteins per organism on average. Interestingly, only 8.9% of the organisms contain ECF243's 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-σ factors, which in turn, detect the presence of iron-siderophore complexes from different FecA transporters (see (60) for a review). Future analyses have to answer whether the different members of the same ECF group in the same genome have acquired different specificities and whether this specificity is a general feature of ECF σ factors.

In summary, the updated ECF classification presented in this work serves as a detailed 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 regulators of ECF activity, conserved proteins encoded in the same genetic neighborhood, and predicted target promoter motifs (Supplementary Text S2, Table S2, and the ECF Hub as online resource). Collectively, this information allows prediction of the potential

function of the members of the group, which is verified by experimentally described members whenever they are available (Supplementary Text S2). Moreover, our hierarchical two-level classification provides a broad sequence collection with an appropriate degree of similarity (or variability) required for *in silico* prediction tools that employ sequence variation-based algorithms, such as co-variation-based prediction of protein-protein interactions or structural predictions.

DATA AVAILABILITY

The ECFHub platform is publicly available at https://www.computational.bio.uni-giessen.de/ecfhub. In addition, the ecf_classify tool is hosted on GitHub (http://ecfclassify.computational.bio) and the hidden Markov models are hosted as citable supporting data at Zenodo (https://doi.org/10.5281/zenodo.3672544).

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

Author contributions: D.C.-P. developed the algorithms and analyzed the data with the help of G.F.. R.R.M., S.J., K.B. and A.G. developed the ECFHub platform. All authors designed the study and wrote the paper. A.G. and G.F. supervised the study.

FUNDING

German Federal Ministry of Education and Research (BMBF); Biotechnology and Biological Sciences Research Council (BBSRC) via the joint ERASynBio project ECF-express [BMBF grant 031L0010B to A.B., A.G., G.F., 031L0010A to T.M., BBSRC grant BB/N006852/1 to M.J.B.]; BBSRC Institute Strategic Programme Grant [BB/J004561/1 to the John Innes Centre]; project 'Bielefeld-Gießen Center for Microbial Bioinformatics-BiGi' [031A533 to R.R.M., S.J., K.B.] within the German Network for Bioinformatics Infrastructure (de.NBI). Funding for open access charge: The University of Western Australia.
Conflict of interest statement. None declared.

REFERENCES

1. Staroń,A., Sofia,H.J., Dietrich,S., Ulrich,L.E., Liesegang,H. and Mascher,T. (2009) The third pillar of bacterial signal transduction: classification of the extracytoplasmic function (ECF) σ factor protein family. *Mol. Microbiol.*, **74**, 557–581.
2. Helmann,J.D. (2002) The extracytoplasmic function (ECF) sigma factors. *Adv. Microb. Physiol.*, **46**, 47–110.
3. Paget,M.S.B. and Helmann,J.D. (2003) The σ70 family of sigma factors. *Genome Biol.*, **4**, 203.
4. Jogler,C., Waldmann,J., Huang,X., Jogler,M., Glöckner,F.O. and Mascher,T. (2012) Identification of proteins likely to be involved in morphogenesis, cell division, and signal transduction in planctomycetes by comparative genomics. *J. Bacteriol.*, **194**, 6419–6430.
5. Huang,X., Pinto,D., Fritz,G. and Mascher,T. (2015) Environmental sensing in Actinobacteria: a comprehensive survey on the signaling capacity of this phylum. *J. Bacteriol.*, **197**, 2517–2535.
6. Pinto,D. and Mascher,T. (2016) The ECF classification: a phylogenetic reflection of the regulatory diversity in the extracytoplasmic function σ factor protein family. In: *Stress and Environmental Regulation of Gene Expression and Adaptation in Bacteria*. John Wiley & Sons, Inc., Hoboken, NJ, Vol. **1**, pp. 64–96.
7. Mascher,T. (2013) Signaling diversity and evolution of extracytoplasmic function (ECF) σ factors. *Curr. Opin. Microbiol.*, **16**, 148–155.
8. Paget,M.S. (2015) Bacterial sigma factors and anti-sigma factors: structure, function and distribution. *Biomolecules*, **5**, 1245–1265.
9. Lonetto,M., Gribskov,M. and Gross,C.A. (1992) The σ70 family: sequence conservation and evolutionary relationships. *J. Bacteriol.*, **174**, 3843–3849.
10. Lane,W.J. and Darst,S.A. (2006) The structural basis for promoter -35 element recognition by the group IV σ factors. *PLoS Biol.*, **4**, 1491–1500.
11. Campagne,S., Marsh,M.E., Capitani,G., Vorholt,J.A. and Allain,F.H.T. (2014) Structural basis for -10 promoter element melting by environmentally induced sigma factors. *Nat. Struct. Mol. Biol.*, **21**, 269–276.
12. Li,L., Fang,C., Zhuang,N., Wang,T. and Zhang,Y. (2019) Structural basis for transcription initiation by bacterial ECF σ factors. *Nat. Commun.*, **10**, 1153.
13. Lin,W., Mandal,S., Degen,D., Cho,M.S., Feng,Y., Das,K. and Ebright,R.H. (2019) Structural basis of ECF-σ-factor-dependent transcription initiation. *Nat. Commun.*, **10**, 710.
14. Fang,C., Li,L., Shen,L., Shi,J., Wang,S., Feng,Y. and Zhang,Y. (2019) Structures and mechanism of transcription initiation by bacterial ECF factors. *Nucleic Acids Res.*, **47**, 7094–7104.
15. Campbell,E.A., Greenwell,R., Anthony,J.R., Wang,S., Lim,L., Das,K., Sofia,H.J., Donohue,T.J. and Darst,S.A. (2007) A conserved structural module regulates transcriptional responses to diverse stress signals in bacteria. *Mol. Cell*, **27**, 793–805.
16. Rajasekar,K.V, Zdanowski,K., Yan,J., Hopper,J.T.S., Francis,M.-L.R., Seepersad,C., Sharp,C., Pecqueur,L., Werner,J.M., Robinson,C.V *et al.* (2016) The anti-sigma factor RsrA responds to oxidative stress by reburying its hydrophobic core. *Nat. Commun.*, **7**, 12194.
17. Francez-Charlot,A., Frunzke,J., Reichen,C., Ebneter,J.Z., Gourion,B. and Vorholt,J.A. (2009) Sigma factor mimicry involved in regulation of general stress response. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 3467–3472.
18. Wecke,T., Halang,P., Staroń,A., Dufour,Y.S., Donohue,T.J. and Mascher,T. (2012) Extracytoplasmic function σ factors of the widely distributed group ECF41 contain a fused regulatory domain. *Microbiologyopen*, **1**, 194–213.
19. Liu,Q., Pinto,D. and Mascher,T. (2018) Characterization of the widely distributed novel ECF42 group of extracytoplasmic function σ factors in streptomyces venezuelae. *J. Bacteriol.*, **200**, e00437-18.
20. Iyer,S.C., Casas-Pastor,D., Kraus,D., Mann,P., Schirner,K., Glatter,T., Fritz,G. and Ringgaard,S. (2020) Transcriptional regulation by σ factor phosphorylation in bacteria. *Nat. Microbiol.*, **5**, 395–406.
21. Nizan-Koren,R., Manulis,S., Mor,H., Iraki,N.M. and Barash,I. (2003) The regulatory cascade that activates the Hrp regulon in Erwinia herbicola pv. gypsophilae. *Mol. Plant-Microbe Interact.*, **16**, 249–260.
22. Paget,M.S., Leibovitz,E. and Buttner,M.J. (1999) A putative two-component signal transduction system regulates sigmaE, a sigma factor required for normal cell wall integrity in Streptomyces coelicolor A3(2). *Mol. Microbiol.*, **33**, 97–107.
23. 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 σ factors involved in the regulation of extracytoplasmic functions. *Proc. Natl. Acad. Sci. U.S.A.*, **91**, 7573–7577.
24. Wiegand,S., Jogler,M., Boedeker,C., Pinto,D., Vollmers,J., Rivas-Marín,E., Kohn,T., Peeters,S.H., Heuer,A., Rast,P. *et al.* (2020) Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. *Nat. Microbiol.*, **5**, 126–140.
25. Sievers,F., Wilm,A., Dineen,D., Gibson,T.J., Karplus,K., Li,W., Lopez,R., McWilliam,H., Remmert,M., Söding,J. *et al.* (2011) Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.*, **7**, 539.

26. Finn,R.D., Clements,J. and Eddy,S.R. (2011) HMMER web server: Interactive sequence similarity searching. *Nucleic Acids Res.*, **39**, W29–W37.
27. Nguyen,L.T., Schmidt,H.A., Von Haeseler,A. and Minh,B.Q. (2015) IQ-TREE: A fast and effective stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol. Biol. Evol.*, **32**, 268–274.
28. 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 Res.*, **44**, W242–W245.
29. Tsirigos,K.D., Peters,C., Shu,N., Käll,L. and Elofsson,A. (2015) The TOPCONS web server for consensus prediction of membrane protein topology and signal peptides. *Nucleic Acids Res.*, **43**, W401–W407.
30. Tsirigos,K.D., Elofsson,A. and Bagos,P.G. (2016) PRED-TMBB2: Improved topology prediction and detection of beta-barrel outer membrane proteins. *In Bioinformatics*, **32**, i665–i671.
31. Steinegger,M. and Söding,J. (2017) MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat. Biotechnol.*, **35**, 1026–1028.
32. Rhodius,V.A., Segall-Shapiro,T.H., Sharon,B.D., Ghodasara,A., Orlova,E., Tabakh,H., Burkhardt,D.H., Clancy,K., Peterson,T.C., Gross,C.A. *et al.* (2013) Design of orthogonal genetic switches based on a crosstalk map of σs, anti-σs, and promoters. *Mol. Syst. Biol.*, **9**, 702.
33. Grass,G., Fricke,B. and Nies,D.H. (2005) Control of expression of a periplasmic nickel efflux pump by periplasmic nickel concentrations. *In BioMetals. Biometals*, **18**, 437–448.
34. Marcos-Torres,F.J., Perez,J., Gomez-Santos,N., Moraleda-Munoz,A. and Munoz-Dorado,J. (2016) In depth analysis of the mechanism of action of metal-dependent sigma factors: characterization of CorE2 from Myxococcus xanthus. *Nucleic Acids Res.*, **44**, 5571–5584.
35. Chevalier,S., Bouffartigues,E., Bazire,A., Tahrioui,A., Duchesne,R., Tortuel,D., Maillot,O., Clamens,T., Orange,N., Feuilloley,M.G.J. *et al.* (2019) Extracytoplasmic function sigma factors in Pseudomonas aeruginosa. *Biochim. Biophys. Acta - Gene Regul. Mech.*, **1862**, 706–721.
36. Yoshimura,M., Asai,K., Sadaie,Y. and Yoshikawa,H. (2004) Interaction of Bacillus subtilis extracytoplasmic function (ECF) sigma factors with the N-terminal regions of their potential anti-sigma factors. *Microbiology*, **150**, 591–599.
37. Wu,H., Liu,Q., Casas-Pastor,D., Dürr,F., Mascher,T. and Fritz,G. (2019) The role of C-terminal extensions in controlling ECF σ factor activity in the widely conserved groups ECF41 and ECF42. *Mol. Microbiol.*, **112**, 498–514.
38. Pérez,J., Muñoz-Dorado,J. and Moraleda-Muñoz,A. (2018) The complex global response to copper in the multicellular bacterium Myxococcus xanthus. *Metallomics*, **10**, 876–886.
39. Luo,Y., Asai,K., Sadaie,Y. and Helmann,J.D. (2010) Transcriptomic and phenotypic characterization of a Bacillus subtilis strain without extracytoplasmic function σ factors. *J. Bacteriol.*, **192**, 5736–5745.
40. Bibb,M.J. and Buttner,M.J. (2003) The streptomyces coelicolor developmental transcription factor σBldN is synthesized as a proprotein. *J. Bacteriol.*, **185**, 2338–2345.
41. Thakur,K.G., Joshi,A.M. and Gopal,B. (2007) Structural and biophysical studies on two promoter recognition domains of the extra-cytoplasmic function σ factor σC from Mycobacterium tuberculosis. *J. Biol. Chem.*, **282**, 4711–4718.
42. Kim,M.S., Hahn,M.Y., Cho,Y., Cho,S.N. and Roe,J.H. (2009) Positive and negative feedback regulatory loops of thiol-oxidative stress response mediated by an unstable isoform of σR in actinomycetes. *Mol. Microbiol.*, **73**, 815–825.
43. Lan,L., Deng,X., Zhou,J. and Tang,X. (2006) Genome-wide gene expression analysis of Pseudomonas syringae pv. tomato DC3000 reveals overlapping and distinct pathways regulated by hrpL and hrpRS. *Mol. Plant-Microbe Interact.*, **19**, 976–987.
44. Merighi,M., Majerczak,D.R., Stover,E.H. and Coplin,D.L. (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. *Mol. Plant-Microbe Interact.*, **16**, 238–248.
45. Luo,S., Sun,D., Zhu,J., Chen,Z., Wen,Y. and Li,J. (2014) An extracytoplasmic function sigma factor, σ25, differentially regulates avermectin and oligomycin biosynthesis in Streptomyces avermitilis. *Appl. Microbiol. Biotechnol.*, **98**, 7097–7112.
46. Tran,N.T., Huang,X., Hong,H.J., Bush,M.J., Chandra,G., Pinto,D., Bibb,M.J., Hutchings,M.I., Mascher,T. and Buttner,M.J. (2019) Defining the regulon of genes controlled by σE, a key regulator of the cell envelope stress response in Streptomyces coelicolor. *Mol. Microbiol.*, **112**, 461–481.
47. Kadowaki,T., Yukitake,H., Naito,M., Sato,K., Kikuchi,Y., Kondo,Y., Shoji,M. and Nakayama,K. (2016) A two-component system regulates gene expression of the type IX secretion component proteins via an ECF sigma factor. *Sci. Rep.*, **6**, 23288.
48. Seipke,R.F., Patrick,E. and Hutchings,M.I. (2014) Regulation of antimycin biosynthesis by the orphan ECF RNA polymerase sigma factor σAntA. *PeerJ*, **2**, e253.
49. Yanamandra,S.S., Sarrafee,S.S., Anaya-Bergman,C., Jones,K. and Lewis,J.P. (2012) Role of the Porphyromonas gingivalis extracytoplasmic function sigma factor, SigH. *Mol. Oral Microbiol.*, **27**, 202–219.
50. Heruth,D.P., Pond,F.R., Dilts,J.A. and Quackenbush,R.L. (1994) Characterization of genetic determinants for R body synthesis and assembly in Caedibacter taeniospiralis 47 and 116. *J. Bacteriol.*, **176**, 3559–3567.
51. Müller,M., Reiß,S., Schlüter,R., Mäder,U., Beyer,A., Reiß,W., Marles-Wright,J., Lewis,R.J., Pförtner,H., Völker,U. *et al.* (2014) Deletion of membrane-associated Asp23 leads to upregulation of cell wall stress genes in Staphylococcus aureus. *Mol. Microbiol.*, **93**, 1259–1268.
52. Shukla,J., Gupta,R., Thakur,K.G., Gokhale,R. and Gopal,B. (2014) Structural basis for the redox sensitivity of the Mycobacterium tuberculosis SigK-RskA σ-anti-σ complex. *Acta Crystallogr. Sect. D Biol. Crystallogr.*, **70**, 1026–1036.
53. Han,K., Li,Z.F., Peng,R., Zhu,L.P., Zhou,T., Wang,L.G., Li,S.G., Zhang,X.B., Hu,W., Wu,Z.H. *et al.* (2013) Extraordinary expansion of a Sorangium cellulosum genome from an alkaline milieu. *Sci. Rep.*, **3**, 2101.
54. Nechaev,S. and Severinov,K. (2003) Bacteriophage-induced modifications of host RNA polymerase. *Annu. Rev. Microbiol.*, **57**, 301–322.
55. Peng,Q. and Yuan,Y. (2018) Characterization of a novel phage infecting the pathogenic multidrug-resistant Bacillus cereus and functional analysis of its endolysin. *Appl. Microbiol. Biotechnol.*, **102**, 7901–7912.
56. Brister,J.R., Ako-Adjei,D., Bao,Y. and Blinkova,O. (2015) NCBI viral genomes resource. *Nucleic. Acids. Res.*, **43**, D571–D577.
57. Ortiz de Ora,L., Lamed,R., Liu,Y.J., Xu,J., Cui,Q., Feng,Y., Shoham,Y., Bayer,E.A. and Muñoz-Gutiérrez,I. (2018) Regulation of biomass degradation by alternative σ factors in cellulolytic clostridia. *Sci. Rep.*, **8**, 11036.
58. MacLellan,S.R., Guariglia-Oropeza,V., Gaballa,A. and Helmann,J.D. (2009) A two-subunit bacterial sigma-factor activates transcription in Bacillus subtilis. *Proc. Natl. Acad. Sci. U.S.A.*, **106**, 21323–21328.
59. Xue,X., Davis,M.C., Steeves,T., Bishop,A., Breen,J., MacEacheron,A., Kesthely,C.A., Hsu,F. and MacLellan,S.R. (2016) Characterization of a protein-protein interaction within the SigO-RsoA two-subunit σ factor: the σ70 region 2.3-like segment of RsoA mediates interaction with SigO. *Microbiology*, **162**, 1857–1869.
60. Braun,V., Mahren,S. and Ogierman,M. (2003) Regulation of the Fecl-type ECF sigma factor by transmembrane signalling. *Curr. Opin. Microbiol.*, **6**, 173–180.