

This is the **accepted version** of the journal article:

Sánchez-Osuna, Miquel; Barbé García, Jordi; Erill, Ivan. «Comparative genomics of the DNA damage-inducible network in the *Patescibacteria*». *Environmental Microbiology*, Vol. 19, issue 9 (Sep. 2017), p. 3465-3474. DOI 10.1111/1462-2920.13826

This version is available at <https://ddd.uab.cat/record/288703>

under the terms of the  **IN**
COPYRIGHT license

Comparative genomics of the DNA-damage inducible network in the Patescibacteria

Miquel Sánchez-Osuna¹, Jordi Barbé¹ & Ivan Erill^{2*} 

¹ Departament de Genètica i de Microbiologia, Universitat Autònoma de Barcelona, Spain

² Department of Biological Sciences, University of Maryland Baltimore County, Baltimore, Maryland, U.S.A.

* Corresponding author: Ivan Erill, Department of Biological Sciences, University of Maryland Baltimore County, 1000 Hilltop Circle, Baltimore, MD 21250, USA. Phone: +1-410-455-2470. Fax: +1-410-455-3875.

Running title: Transcriptional responses in uncultivated bacteria

ORIGINALITY-SIGNIFICANCE STATEMENT

Metagenomics analyses have revealed that uncultivated candidate phyla constitute a remarkably large fraction of many environmental bacterial communities. Metabolic profiling provides insights into the lifestyle and interactions of bacteria from these novel phyla, but many dynamic properties of these organisms, such as their response to environmental stressors, cannot be obtained through conventional metagenomics pipelines. Here we demonstrate an innovative approach combining *in silico* and *in vitro* methods to characterize *de novo* transcriptional regulatory networks in uncultivated bacteria using metagenomic data. We showcase this approach through the characterization of the SOS response against DNA damage in the large candidate superphylum Patescibacteria, thought to account for ~10% of the bacterial diversity in environmental samples. Our results identify a novel binding motif for the SOS transcriptional regulator, outline differences among the SOS response of different Patescibacteria phyla and provide insights into the lifestyle and genomic evolution of this recently described superphylum. More generally, this work illustrates how metagenomic data can be leveraged to infer transcriptionally encoded responses to environmental stressors in uncultivated organisms, enhancing our understanding of their physiology and adaptation to different environments.

ABSTRACT

Metagenomics provide unprecedented insights into the genetic diversity of uncultivated bacteria inhabiting natural environments. Recent surveys have uncovered a major radiation of candidate phyla encompassing the Patescibacteria superphylum. Patescibacteria have small genomes and a presumed symbiotic or parasitic lifestyle, but the difficulty in culturing representative members constrains the characterization of behavioral and adaptive traits. Here we combine *in silico* and *in vitro* approaches to characterize the SOS transcriptional response to DNA damage in the Patescibacteria superphylum.

This article has been accepted for publication and undergone full peer review but has not been through the copyediting, typesetting, pagination and proofreading process which may lead to differences between this version and the Version of Record. Please cite this article as an 'Accepted Article', doi: 10.1111/1462-2920.13826

Leveraging comparative genomics methods, we identify and experimentally define a novel binding motif for the SOS transcriptional repressor LexA, and we use this motif to characterize the conserved elements of the SOS regulatory network in Patescibacteria. The Patescibacteria LexA-binding motif has unusual direct-repeat structure, and comparative analyses reveal sequence and structural similarities with the distant Acidobacteria LexA protein. Our results reveal a shared core SOS network, complemented by varying degrees of LexA regulation of other core SOS functions. This work illustrates how the combination of computational and experimental methods can leverage metagenomic data to characterize transcriptional responses in uncultivated bacteria. The report of an operational SOS response in presumed symbiotic and parasitic bacteria hints at an intermediate step in the process of genome reduction.

KEYWORDS

metagenomics, gene regulation, genome repair, comparative genomics, evolutionary processes

INTRODUCTION

In recent years, the rapid development of high-throughput technologies for metagenomic and single-cell genomic sequencing has enabled researchers to study the genetic repertoire of microbial communities with unprecedented resolution (Riesenfeld *et al.*, 2004; Stepanauskas, 2012). The culture-independent nature of these methods allows studying the role of uncultivated organisms in microbial communities (Handelsman, 2004; Stepanauskas, 2012), and has made possible the genomic characterization of uncultivated bacteria in host-associated and environmental samples (Di Rienzi *et al.*, 2013; Rinke *et al.*, 2013). Environmental metagenomics surveys have uncovered several large clades of uncultivated bacteria that could not be detected through conventional 16S rRNA sequencing (Di Rienzi *et al.*, 2013; Brown *et al.*, 2015; Eloë-Fadrosh *et al.*, 2016). The vast majority of these novel clades belong to the candidate phyla radiation (CPR), a large group of highly divergent bacterial species lacking isolated representatives that is estimated to encompass more than 15% of the Bacteria domain (Brown *et al.*, 2015). Three of the candidate phyla in the CPR (Parcubacteria, Microgenomates and Gracilibacteria) have been grouped into the candidate Patescibacteria superphylum (Rinke *et al.*, 2013; Hedlund *et al.*, 2014). Patescibacteria sequences were first reported in groundwater and sediments of anoxic aquatic environments (Elshahed *et al.*, 2005; Youssef *et al.*, 2011; Wrighton *et al.*, 2012), but prospective

metagenomics analyses have since established that this superphylum has a widespread environmental distribution encompassing mostly semiaquatic habitats such in oil refineries, freshwater beaches, hydrothermal sites, Andean maize chacras or the Alpine permafrost (Rinke *et al.*, 2013; Correa-Galeote *et al.*, 2016; Frey *et al.*, 2016; Luo *et al.*, 2016; Mohiuddin *et al.*, 2017). Metagenomics analyses indicate that Patescibacteria share several genomic traits. Like most CPR organisms, Patescibacteria have small genomes (< 1 Mbp) and an unusual ribosome composition characterized primarily by the loss of ribosomal protein L30 (Brown *et al.*, 2015). In addition, Patescibacteria lack a tricarboxylic acid (TCA) cycle and most elements of the electron transfer chain, suggesting that they are strict anaerobic fermenters (Wrighton *et al.*, 2012). Based on their limited biosynthetic capabilities and reduced repertoire of DNA repair genes, as well as on studies of cultivated isolates from other CPR clades, it has been postulated that members of the Patescibacteria may illustrate a genomic reduction process towards an ectosymbiotic lifestyle (He *et al.*, 2015, 7; Nelson and Stegen, 2015).

Metagenomics analyses are conducive to the study of communal genome properties, such as the presence and interspecies setup of metabolic pathways, through the mapping of identified genes onto well-established knowledge-bases (De Filippo *et al.*, 2012). These methodologies have been successfully leveraged to unravel complex interspecies metabolic networks and ecological interactions driving essential biogeochemical processes in aquatic environments (Anantharaman *et al.*, 2016; Speth *et al.*, 2016; Probst *et al.*, 2017), and complemented by functional metagenomics studies to biochemically characterize key components of such processes (Wrighton *et al.*, 2016; Sukul *et al.*, 2017; Yaniv *et al.*, 2017). However, the computational pipelines traditionally used in metagenomics analyses are not designed to characterize some dynamic aspects of the bacterial genome, such as the coordinated response to intra- and extracellular stimuli orchestrated by transcriptional regulatory networks, because the mere identification of putative network genes does not provide information on their regulation (Cornish *et al.*, 2014). The SOS response is a well-studied transcriptional network aimed at addressing DNA damage (Walker *et al.*, 2000). Present in most bacterial phyla, the SOS response coopts the universal recombination protein RecA to sense DNA damage and promote auto-cleavage of the SOS transcriptional repressor, LexA, inducing the expression of genes involved in DNA repair, cell-division and translesion synthesis (Erill *et al.*, 2007). In contrast with many other transcriptional regulators, the LexA repressor targets highly-divergent motifs in different bacterial clades, ranging from palindromic repeats in the Firmicutes and the Gammaproteobacteria, to direct repeats in the Alphaproteobacteria and

Acidobacteria (Fernandez de Henestrosa *et al.*, 1998; Erill *et al.*, 2003; Au *et al.*, 2005; Mazon *et al.*, 2006). In this work we combine *in silico* and *in vitro* methods to define the LexA-binding motif of the Patescibacteria, and we put forward a comparative genomics approach that leverages metagenomics data to characterize the SOS regulatory network in this superphylum. Our results reveal that LexA targets an unusual direct-repeat motif to regulate a small but varied network of genes involved in core and non-canonical SOS functions. The identification of an operational SOS response in a bacterial clade of presumed ectosymbionts can shed light into the process of genomic reduction associated with this lifestyle.

RESULTS

Definition of the Patescibacteria LexA-binding motif

Most bacterial transcriptional repressors regulate their own transcription, enabling the identification of their target DNA motifs through comparative analysis of their promoter regions (Sahota and Stormo, 2010; Cornish *et al.*, 2012). To define the Patescibacteria LexA-binding motif, we first compiled a panel of 48 experimentally-validated LexA proteins (Table S1) and used BLASTP to search the protein records associated with distinct Microgenomates, Parcubacteria and Gracilibacteria species records in the NCBI GenBank database (May 2016 release) (Altschul *et al.*, 1997; O'Leary *et al.*, 2016). Bona-fide LexA orthologs in species from these clades were identified as bidirectional BLAST hits with broad coverage (>85% of query) and containing the signature Ala-Gly cleavage site and Ser-Lys motif required for LexA self-proteolysis (Lin and Little, 1988). This resulted in the identification of 177 LexA orthologs in the Microgenomates and 40 LexA orthologs in the Parcubacteria (Table S2). No LexA orthologs were identified in the Gracilibacteria, but this is most likely due to the small number of currently available sequences for this candidate phylum. To analyze the effect of undersampling on the identification of LexA orthologs we mapped the distribution of identified LexA proteins on a previously reported phylogeny of the Parcubacteria (Hug *et al.*, 2016). The observed distribution of LexA across the Microgenomates and the Parcubacteria was fairly homogeneous (Fig. S1). Several genera in both phyla show evidence of LexA loss, but only three genera ("*Candidatus* Moranbacteria", "*Candidatus* Wolfebacteria" and "*Candidatus* Roizmanbacteria") have enough sequence data to support this hypothesis. We identified a weak but statistically significant correlation (Pearson $R=0.35$, $p\text{-value}=2.2\cdot 10^{-}$

¹⁶⁾ between the number of identified LexA proteins and the number of available protein sequences in any given clade, suggesting that the number of identified LexA orthologs in poorly represented clades is likely an underestimate. The upstream region (-250 to +50 bp relative to the predicted start codon) of the genes coding for these LexA orthologs was extracted and filtered to remove highly-similar (>90% identity) sequences. The resulting sets of putative *lexA* promoters (79 for Microgenomates and 20 for Parcubacteria) were then submitted to MEME for detection of overrepresented motifs (Bailey and Elkan, 1994). In both the Microgenomates and the Parcubacteria, the most significant motif reported by MEME displays a similar direct repeat structure with consensus sequence TTCGG-N6-TTCGG (Figure 1A; Table S3). This motif is present in 67 Microgenomates *lexA* promoters, typically as a single instance located immediately downstream of a putative -35 element. In contrast, the corresponding Parcubacteria motif was identified in 14 *lexA* promoters, but usually in a tandem organization with the two motif instances located in opposite strands and separated by 10-15 bp. In this tandem arrangement, one LexA-binding site partially overlaps the predicted -35 element, while the other site is located a few bases downstream of the -10 element. This organization is consistent with LexA self-repression and has been also reported in other phyla (Figure 1B; Fig. S2) (Sanchez-Alberola *et al.*, 2015). To validate that the two motifs identified by MEME correspond to a single LexA-binding motif conserved across the Patescibacteria, we cloned and overexpressed in *Escherichia coli* the gene coding for the “*Candidatus* Collierbacteria bacterium GW2011_GWB2_45_17” LexA ortholog and purified its product. Using the purified LexA protein from this Microgenomates species, we performed electromobility-shift assays (EMSA) on oligonucleotides containing wild-type and mutated versions of the putative LexA-binding site within the promoter region of the Parcubacteria group bacterium GW2011_GWA2_48_9 *lexA* gene.

The results shown in Figure 1C demonstrate that the Microgenomates LexA is capable of binding Parcubacteria LexA-binding sites, confirming that the LexA-binding motifs identified *in silico* correspond to a single conserved motif in the Patescibacteria. EMSA on mutated versions of the LexA-binding site establish that binds the Patescibacteria LexA-binding motif identified *in silico* by targeting its conserved TTCGG direct repeat element. Single and multiple transversions within the TTCGG element abolish binding, whereas transversions, but not insertions, are tolerated in the 6 bp spacer region between TTCGG elements. These results are consistent with those obtained for other LexA proteins, where LexA has been shown to make specific contacts with dyad elements and to be highly sensitive to changes in the length of the dyad spacer (Erill *et al.*, 2007, 2016; Zhang *et al.*, 2010). To date, the LexA protein has

been shown to bind motifs with a direct-repeat structure in only two clades (Acidobacteria and Alphaproteobacteria). In both clades, LexA targets a direct-repeat motif similar to the one reported here for the Patescibacteria (Fig. S3). This motif has consensus sequence GTTC-N7-GTTC and it has been proposed that this unusual motif structure arose in the Acidobacteria and was subsequently acquired by the Alphaproteobacteria through lateral transfer of the *lexA* gene (Mazon *et al.*, 2006). The α 3-helix of the LexA helix-turn-helix DNA binding motif is known to play a major role in specific recognition of LexA-binding motif dyads (Thliveris and Mount, 1992; Groban *et al.*, 2005; Zhang *et al.*, 2010). We compared the Patescibacteria LexA α 3-helix with a panel of previously reported LexA α 3-helix alignments using TomTom (Gupta *et al.*, 2007; Sanchez-Alberola *et al.*, 2015) (Fig. S3; Table S4). Our results indicate that the Patescibacteria LexA α 3-helix is most closely related to the Acidobacteria one, showing only minor changes in the N-terminal part of the helix known to determine direct readout sequence specificity (Figure 1D; Fig. S4). Furthermore, phylogenetic analysis using the LexA protein sequence places the Acidobacteria LexA within the Patescibacteria clade (Fig. S5), in contrast to their distant placement in multiple phylogenetic analyses (Rinke *et al.*, 2013; Hug *et al.*, 2016; Yeoh *et al.*, 2016). These results support to the notion that this unusual LexA-binding motif structure has been disseminated through lateral gene transfer.

Reconstruction of the Patescibacteria LexA regulon

The availability of a conserved transcription factor-binding motif for a bacterial clade can be leveraged to interrogate the regulatory network it controls through comparative genomics (Gelfand *et al.*, 2000; Cornish *et al.*, 2012). This approach is maximally informative when analyzing complete genomes, but it can be generalized to take into account missing information and provide reliable estimates of network composition on partial genome sequences (Erill *et al.*, 2016). To elucidate the composition of the Patescibacteria LexA regulon, we searched all sequences available in NCBI GenBank for Microgenomates and Parcubacteria species records with their respective LexA-binding motifs (Figure 1A; Table S3) using FIMO (Grant *et al.*, 2011). Statistically significant LexA-binding site instances ($q < 0.05$) were associated with predicted operons if they located to the putative promoter region (-250 to +50 of the predicted start codon) of the first gene in the operon. The relative frequency of regulation across genomes in each phylum was computed independently for each gene within the operon and EMSA with purified Microgenomates LexA protein were performed to validate several predicted LexA-binding sites in the Microgenomates and the Parcubacteria. The results shown in Figure 2 show both similarities and

remarkable differences in the composition of the LexA regulon for these two Patescibacteria phyla (Fig. S6; Table S5; Table S6). The core Patescibacteria LexA regulon is composed of *lexA*, *recA*, *dinD* and a *rhuM* homolog. Previous clade-wide analyses of the LexA regulon have reported small core regulons encompassing both *lexA* and *recA* (Erill *et al.*, 2007), and the role of *dinD* as a SOS-regulated modulator of RecA activity has been established in *E. coli* (Uranga *et al.*, 2011). The identified *rhuM* homologs map to COG3943, which contains members annotated as Fic domain-containing toxins that have been linked to virulence in *Salmonella typhimurium* and to growth arrest through interaction with DNA topoisomerases in *E. coli* (Amavisit *et al.*, 2003; Harms *et al.*, 2015). A *rhuM* homolog was also found to be LexA regulated in a metagenomic analysis of the Firmicutes SOS response (Cornish *et al.*, 2014). In this context, the putative LexA regulation of a *rhuM* homolog could operate as a functional analog of the SOS mediated inhibition of cell-division reported in *E. coli* and *Bacillus subtilis* (Bi and Lutkenhaus, 1993; Kawai *et al.*, 2003).

Beyond the core regulon, only one gene appears to be consistently regulated by LexA in the Parcubacteria (Figure 2B; Fig. S6). The *ppaC* gene product is an inorganic pyrophosphatase. These enzymes are known to play an important role in DNA polymerization, and their upregulation in the aftermath of DNA damage could contribute to enhance the processivity of DNA polymerases (Lapenta *et al.*, 2016). In contrast, the predicted Microgenomates LexA regulon is substantially larger and encompasses several canonical members of the SOS response (Figure 2A; Fig. S6). Besides *lexA*, *recA* and *dinD*, the Microgenomates LexA regulon includes the nucleotide excision repair *uvrABC* operon, transcribed divergently from *lexA*, as well as homologs of the translesion synthesis polymerase DinB. This suite of canonical SOS elements in the Microgenomates LexA regulon is complemented by several interesting additions. SOS regulation of the *dnaG-rpoD* operon, reported previously in *E. coli*, is probably geared to facilitate the reestablishment of DNA replication after replication fork stall (Lupski *et al.*, 1984). The presence of a LexA regulated operon encompassing *rarA* and a NuDiX hydrolase with homology to MutT is consistent with the role of RarA at stalled replication forks and suggests an involvement of the base excision repair (BER) system in the Microgenomates SOS response (Lestini and Michel, 2007). This hypothesis is further substantiated by the putative regulation of a large polycistronic unit containing two BER-related genes (*polA* and *mutM*). The Microgenomates LexA regulon also appears to encompass the *dcw/mra* gene cluster, responsible for peptidoglycan biosynthesis and cell division control. The gene product of this operon's lead gene (*mraZ*) is known to inhibit cell division (Eraso *et al.*, 2014) and LexA-

binding sites in its promoter region have been reported in *E. coli* and the Firmicutes (Vicente *et al.*, 1998; Cornish *et al.*, 2014), suggesting a role in SOS induced cell-division arrest.

The SOS response as an indicator of bacterial lifestyle

The absence of many DNA repair genes is a common signature of endosymbiont genomes (Dale *et al.*, 2003). Furthermore, the loss of DNA repair mechanism has been postulated as a driving mechanism for genomic reduction in both free-living and endosymbiotic bacteria, with an increase in the effective mutation rate promoting niche colonization and the rapid loss of non-essential genes (Blanc *et al.*, 2007; Marais *et al.*, 2008; Batut *et al.*, 2014). Given that the unregulated expression of genes involved in DNA repair can often be detrimental to the cell, it can be argued that de-regulation of DNA repair mechanisms should precede their inactivation and loss. A BLASTP analysis of DNA repair systems in 29 bacterial clades with substantial genomic reduction gives support to this notion (Table S7). With independence of the DNA repair systems they have lost, all the analyzed clades except the *Morganellaceae* and the here-reported Patescibacteria lack a LexA ortholog, suggesting that the absence of a functional SOS response is a defining trait of host-associated genomic reduction. In this context, the characterization of an operational SOS response in the Patescibacteria provides additional insight into their lifestyle and genomic organization. Although additional work is required to establish it, based on their limited biosynthetic repertoire it has been postulated that the Patescibacteria and several other members of the CPR are either ectosymbionts or epibiotic parasites (He *et al.*, 2015; Nelson and Stegen, 2015; Yeoh *et al.*, 2016). While consistent with an epibiotic lifestyle, the presence of an operational SOS response indicates a need to cope independently with changing environments and external assaults. This could be due to an interim free-living stage in the context of a loose host-symbiont association or to an epibiotic lifestyle restricted to unicellular hosts, which are unable to confer sufficient protection from environmental insults (He *et al.*, 2015; Yeoh *et al.*, 2016). The apparent loss of LexA in some Patescibacteria genera (e.g. “*Candidatus* Moranbacteria” and “*Candidatus* Roizmanbacteria”) could hence indicate transition to an endosymbiotic lifestyle or association with a multicellular host. At least one Parcubacteria species has been shown to have adopted an endosymbiotic lifestyle (Gong *et al.*, 2014), but currently there is not enough sequence data for this genus to ascertain whether such transition correlates with the absence of a functional SOS response.

CONCLUSIONS

Metagenomics and single-cell genomics provide the means to explore the genetic repertoire of microbial communities and characterize uncultivated clades and organisms. A substantial part of their biology, however, is encoded in regulatory circuits and cannot be elicited directly through conventional metagenomics *in silico* pipelines. Here we show how comparative genomics methods, in combination with *in vitro* experiments, can be used to characterize *de novo* the transcriptional response to DNA damage in the Patescibacteria, a bacterial superphylum within the candidate phyla radiation lacking isolated representatives. Our work shows that the Patescibacteria LexA controls a shared core set of SOS genes by binding to a novel LexA-binding motif with an unusual direct-repeat structure that appears to have been disseminated through lateral gene transfer. The analysis also reveals significant differences in the composition of the LexA regulon between the main Patescibacteria phyla (Microgenomates and Parcubacteria), and points towards the adoption of convergent mechanisms in the inhibition of cell division by the SOS response. Leveraging draft genomic sequences from metagenomics datasets, the combination of *in silico* and *in vitro* methods to reconstruct regulatory networks has the potential to shed light on the lifestyle and evolution of uncultivated organisms. The description of an operational SOS response in the Patescibacteria provides further support for an epibiotic lifestyle, most likely in association with unicellular hosts. Furthermore, if further research confirms that the Patescibacteria have undergone genomic reduction, as their limited biosynthetic capabilities and small genome size seem to indicate, the presence of a functional SOS response hints at an unconventional process of genomic reduction in this superphylum.

EXPERIMENTAL PROCEDURES

A full version of the computational and experimental methods and associated references is available in the Supporting Information online.

Bioinformatics techniques

LexA orthologs in Patescibacteria were identified through bidirectional BLAST hit using a panel of experimentally-validated LexA proteins and conservative e-value ($<1e-30$) and query coverage ($>85\%$) limits. Orthologs were further validated via identification of conserved motifs required for LexA self-proteolysis (Lin and Little, 1988). The LexA-binding motif was inferred with MEME on a collection of non-redundant ($<90\%$ identity) sequences upstream of identified LexA homologs using a 14-20 bp motif

size, the ANR sites per sequence model and otherwise default parameters (Bailey and Elkan, 1994). To reconstruct LexA regulons, instances of LexA-binding sites in nucleotide sequences belonging to Patescibacteria species with putatively regulated *lexA* genes were identified with FIMO using a cutoff *q*-value of 0.05 (Grant *et al.*, 2011). Genes in the same strand with a maximum intergenic distance of 150 bp were considered to constitute operons. LexA-binding sites mapping to the upstream region (-250 to +50 of the predicted start codon) of the first gene in a predicted operon were retained for analysis and associated with all operon members. Alignments of LexA α 3-helix sequences for Patescibacteria and previously reported LexA orthologs were compared using TomTom (Gupta *et al.*, 2007). Predicted proteins were mapped to non-supervised orthologous groups (NOGs) using the eggNOG v4.0 web service (Powell *et al.*, 2014). The scripts used for the analysis are available at the GitHub ErillLab repository.

DNA-binding assays

The “*Candidatus* Collierbacteria bacterium GW2011_GWB2_45_17” *lexA* gene [UX01_C0005G0039] was overexpressed in *E. coli* BL21-CodonPlus(DE3)-RIL (Stratagene) competent cells and purified following the protocol previously described for other LexA proteins (Cambray *et al.*, 2011). 100 bp-long DNA probes for electro-mobility shift assays (EMSA) were generated using two complementary synthetic oligonucleotides centered on predicted LexA-binding sites (Table S8), and performing PCR with M13 forward and reverse digoxigenin-labeled oligos, as described previously (Campoy *et al.*, 2003). EMSAs were performed, using 20 ng of each digoxigenin-marked DNA probe in the binding mixture and adding 40 nM of LexA protein to the mixture as described previously (Sanchez-Alberola *et al.*, 2012). Samples were loaded onto 6% non-denaturing Tris-glycine polyacrylamide gels and digoxigenin-labeled DNA-protein complexes were detected using the manufacturer's protocol (Roche NimbleGen). Mutants of the GW2011_GWA2_48_9 *lexA* gene [UY34_C0013G0015] promoter region were obtained by PCR mutagenesis using synthetic oligonucleotides carrying the desired substitutions (Table S8). The DNA sequence of all PCR-mutagenized fragments was determined on an ABI 3730XL sequencer (Macrogen).

CONFLICT OF INTEREST STATEMENT

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

ACKNOWLEDGEMENTS

This work was supported by Ministerio de Economía, Industria y Competitividad (BIO2016–77011-R) and Generalitat de Catalunya (2014SGR572) awards to JB and by a U.S. National Science Foundation (MCB-1158056) award to IE. MSO is recipient of a predoctoral fellowship from the Ministerio de Educación, Cultura y Deporte. The authors wish to thank Joan Ruiz and Susana Escribano for their technical support in some of the experimental procedures.

REFERENCES

- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25**: 3389–402.
- Amavisit, P., Lightfoot, D., Browning, G., and Markham, P. (2003) Variation between pathogenic serovars within *Salmonella* pathogenicity islands. *J. Bacteriol.* **185**: 3624–3635.
- Anantharaman, K., Brown, C.T., Hug, L.A., Sharon, I., Castelle, C.J., Probst, A.J., et al. (2016) Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. *Nat. Commun.* **7**: 13219.
- Anzaldi, L.J., Muñoz-Fernández, D., and Erill, I. (2012) BioWord: A sequence manipulation suite for Microsoft Word. *BMC Bioinformatics* **13**: 124.
- Au, N., Kuester-Schoeck, E., Mandava, V., Bothwell, L.E., Canny, S.P., Chachu, K., et al. (2005) Genetic composition of the *Bacillus subtilis* SOS system. *J Bacteriol* **187**: 7655–66.
- Bailey, T.L. and Elkan, C. (1994) Fitting a mixture model by expectation maximization to discover motifs in biopolymers. *Proc Int Conf Intell Syst Mol Biol* **2**: 28–36.
- Batut, B., Knibbe, C., Marais, G., and Daubin, V. (2014) Reductive genome evolution at both ends of the bacterial population size spectrum. *Nat. Rev. Microbiol.* **12**: 841–850.
- Bi, E. and Lutkenhaus, J. (1993) Cell division inhibitors SulA and MinCD prevent formation of the FtsZ ring. *J Bacteriol* **175**: 1118–25.
- Blanc, G., Ogata, H., Robert, C., Audic, S., Suhre, K., Vestris, G., et al. (2007) Reductive genome evolution from the mother of Rickettsia. *PLoS Genet* **3**: e14.
- Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., et al. (2015) Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* **523**: 208–211.
- Cambray, G., Sanchez-Alberola, N., Campoy, S., Guerin, E., Da Re, S., Gonzalez-Zorn, B., et al. (2011) Prevalence of SOS-mediated control of integron integrase expression as an adaptive trait of chromosomal and mobile integrons. *Mob DNA* **2**: 6.
- Campoy, S., Fontes, M., Padmanabhan, S., Cortes, P., Llagostera, M., and Barbe, J. (2003) LexA-independent DNA damage-mediated induction of gene expression in *Myxococcus xanthus*. *Mol Microbiol* **49**: 769–81.
- Colaert, N., Helsen, K., Martens, L., Vandekerckhove, J., and Gevaert, K. (2009) Improved visualization of protein consensus sequences by iceLogo. *Nat. Methods* **6**: 786–787.
- Cornish, J.P., Matthews, F., Thomas, J.R., and Erill, I. (2012) Inference of self-regulated transcriptional networks by comparative genomics. *Evol Bioinform Online* **8**: 449–61.
- Cornish, J.P., Sanchez-Alberola, N., O'Neill, P.K., O'Keefe, R., Gheba, J., and Erill, I. (2014) Characterization of the SOS meta-regulon in the human gut microbiome. *Bioinformatics* **30**: 1193–1197.
- Correa-Galeote, D., Bedmar, E.J., Fernández-González, A.J., Fernández-López, M., and Arone, G.J. (2016) Bacterial Communities in the Rhizosphere of Amilaceous Maize (*Zea mays* L.) as Assessed by Pyrosequencing. *Plant Sci.* 1016.
- Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004) WebLogo: a sequence logo generator. *Genome Res* **14**: 1188–90.

- Dale, C., Wang, B., Moran, N., and Ochman, H. (2003) Loss of DNA recombinational repair enzymes in the initial stages of genome degeneration. *Mol. Biol. Evol.* **20**: 1188–1194.
- De Filippo, C., Ramazzotti, M., Fontana, P., and Cavalieri, D. (2012) Bioinformatic approaches for functional annotation and pathway inference in metagenomics data. *Brief. Bioinform.* **13**: 696–710.
- Di Rienzi, S.C., Sharon, I., Wrighton, K.C., Koren, O., Hug, L.A., Thomas, B.C., et al. (2013) The human gut and groundwater harbor non-photosynthetic bacteria belonging to a new candidate phylum sibling to Cyanobacteria. *eLife* **2**: e01102.
- Eloe-Fadrosh, E.A., Ivanova, N.N., Woyke, T., and Kyrpides, N.C. (2016) Metagenomics uncovers gaps in amplicon-based detection of microbial diversity. *Nat. Microbiol.* **1**: 15032.
- Elshahed, M.S., Najar, F.Z., Aycock, M., Qu, C., Roe, B.A., and Krumholz, L.R. (2005) Metagenomic analysis of the microbial community at Zodletone Spring (Oklahoma): insights into the genome of a member of the novel candidate division OD1. *Appl. Environ. Microbiol.* **71**: 7598–7602.
- Eraso, J.M., Markillie, L.M., Mitchell, H.D., Taylor, R.C., Orr, G., and Margolin, W. (2014) The highly conserved MraZ protein is a transcriptional regulator in *Escherichia coli*. *J. Bacteriol.* **196**: 2053–2066.
- Erill, I., Campoy, S., and Barbe, J. (2007) Aeons of distress: an evolutionary perspective on the bacterial SOS response. *FEMS Microbiol Rev* **31**: 637–56.
- Erill, I., Campoy, S., Kılıç, S., and Barbé, J. (2016) The Verrucomicrobia LexA-Binding Motif: Insights into the Evolutionary Dynamics of the SOS Response. *Front. Mol. Biosci.* **3**.
- Erill, I., Escribano, M., Campoy, S., and Barbe, J. (2003) In silico analysis reveals substantial variability in the gene contents of the gamma proteobacteria LexA-regulon. *Bioinformatics* **19**: 2225–36.
- Fernandez de Henestrosa, A.R., Rivera, E., Tapias, A., and Barbe, J. (1998) Identification of the *Rhodobacter sphaeroides* SOS box. *Mol Microbiol* **28**: 991–1003.
- Frey, B., Rime, T., Phillips, M., Stierli, B., Hajdas, I., Widmer, F., and Hartmann, M. (2016) Microbial diversity in European alpine permafrost and active layers. *FEMS Microbiol. Ecol.* **92**.
- Gelfand, M.S., Novichkov, P.S., Novichkova, E.S., and Mironov, A.A. (2000) Comparative analysis of regulatory patterns in bacterial genomes. *Brief. Bioinform.* **1**: 357–371.
- Gong, J., Qing, Y., Guo, X., and Warren, A. (2014) “*Candidatus Sonnebornia yantaiensis*”, a member of candidate division OD1, as intracellular bacteria of the ciliated protist *Paramecium bursaria* (Ciliophora, Oligohymenophorea). *Syst. Appl. Microbiol.* **37**: 35–41.
- Grant, C.E., Bailey, T.L., and Noble, W.S. (2011) FIMO: scanning for occurrences of a given motif. *Bioinforma. Oxf. Engl.* **27**: 1017–1018.
- Groban, E.S., Johnson, M.B., Banky, P., Burnett, P.G., Calderon, G.L., Dwyer, E.C., et al. (2005) Binding of the *Bacillus subtilis* LexA protein to the SOS operator. *Nucleic Acids Res* **33**: 6287–95.
- Gupta, S., Stamatiyannopoulos, J.A., Bailey, T.L., and Noble, W.S. (2007) Quantifying similarity between motifs. *Genome Biol.* **8**: R24.
- Handelsman, J. (2004) Metagenomics: application of genomics to uncultured microorganisms. *Microbiol. Mol. Biol. Rev. MMBR* **68**: 669–685.
- Harms, A., Stanger, F.V., Scheu, P.D., de Jong, I.G., Goepfert, A., Glatter, T., et al. (2015) Adenylation of Gyrase and Topo IV by FicT Toxins Disrupts Bacterial DNA Topology. *Cell Rep.* **12**: 1497–1507.
- He, X., McLean, J.S., Edlund, A., Yooseph, S., Hall, A.P., Liu, S.-Y., et al. (2015) Cultivation of a human-associated TM7 phylotype reveals a reduced genome and epibiotic parasitic lifestyle. *Proc. Natl. Acad. Sci. U. S. A.* **112**: 244–249.
- Hedlund, B.P., Dodsworth, J.A., Murugapiran, S.K., Rinke, C., and Woyke, T. (2014) Impact of single-cell genomics and metagenomics on the emerging view of extremophile “microbial dark matter.” *Extrem. Life Extreme Cond.* **18**: 865–875.
- Hug, L.A., Baker, B.J., Anantharaman, K., Brown, C.T., Probst, A.J., Castelle, C.J., et al. (2016) A new view of the tree of life. *Nat. Microbiol.* **1**: 16048.
- Kawai, Y., Moriya, S., and Ogasawara, N. (2003) Identification of a protein, YneA, responsible for cell division suppression during the SOS response in *Bacillus subtilis*. *Mol Microbiol* **47**: 1113–22.
- Lapenta, F., Silva, A.M., Brandimarti, R., Lanzi, M., Gratani, F.L., Gonzalez, P.V., et al. (2016) *Escherichia coli* DnaE Polymerase Couples Pyrophosphatase Activity to DNA Replication. *PLOS ONE* **11**: e0152915.
- Lestini, R. and Michel, B. (2007) UvrD controls the access of recombination proteins to blocked replication forks. *EMBO J.* **26**: 3804–3814.

- Lin, L.L. and Little, J.W. (1988) Isolation and characterization of noncleavable (Ind-) mutants of the LexA repressor of *Escherichia coli* K-12. *J Bacteriol* **170**: 2163–73.
- Luo, F., Devine, C.E., and Edwards, E.A. (2016) Cultivating microbial dark matter in benzene-degrading methanogenic consortia. *Environ. Microbiol.* **18**: 2923–2936.
- Lupski, J.R., Ruiz, A.A., and Godson, G.N. (1984) Promotion, termination, and anti-termination in the rpsU-dnaG-rpoD macromolecular synthesis operon of *E. coli* K-12. *Mol. Gen. Genet. MGG* **195**: 391–401.
- Marais, G.A.B., Calteau, A., and Tenaillon, O. (2008) Mutation rate and genome reduction in endosymbiotic and free-living bacteria. *Genetica* **134**: 205–210.
- Mazon, G., Campoy, S., Erill, I., and Barbe, J. (2006) Identification of the *Acidobacterium capsulatum* LexA box reveals a lateral acquisition of the Alphaproteobacteria *lexA* gene. *Microbiology* **152**: 1109–18.
- Mohiuddin, M.M., Salama, Y., Schellhorn, H.E., and Golding, G.B. (2017) Shotgun metagenomic sequencing reveals freshwater beach sands as reservoir of bacterial pathogens. *Water Res.* **115**: 360–369.
- Nelson, W.C. and Stegen, J.C. (2015) The reduced genomes of Parcubacteria (OD1) contain signatures of a symbiotic lifestyle. *Front. Microbiol.* **6**: 713.
- O’Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., et al. (2016) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**: D733–745.
- Powell, S., Forslund, K., Szklarczyk, D., Trachana, K., Roth, A., Huerta-Cepas, J., et al. (2014) eggNOG v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Res.* **42**: D231–239.
- Probst, A.J., Castelle, C.J., Singh, A., Brown, C.T., Anantharaman, K., Sharon, I., et al. (2017) Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO₂ concentrations. *Environ. Microbiol.* **19**: 459–474.
- Riesenfeld, C.S., Schloss, P.D., and Handelsman, J. (2004) METAGENOMICS: Genomic Analysis of Microbial Communities. *Annu. Rev. Genet.* **38**: 525–552.
- Rinke, C., Schwientek, P., Sczyrba, A., Ivanova, N.N., Anderson, I.J., Cheng, J.-F., et al. (2013) Insights into the phylogeny and coding potential of microbial dark matter. *Nature* **499**: 431–437.
- Sahota, G. and Stormo, G.D. (2010) Novel sequence-based method for identifying transcription factor binding sites in prokaryotic genomes. *Bioinformatics* **26**: 501.
- Sanchez-Alberola, N., Campoy, S., Barbe, J., and Erill, I. (2012) Analysis of the SOS response of *Vibrio* and other bacteria with multiple chromosomes. *BMC Genomics* **13**: 58.
- Sanchez-Alberola, N., Campoy, S., Emerson, D., Barbe, J., and Erill, I. (2015) A SOS regulon under control of a non-canonical LexA-binding motif in the Betaproteobacteria. *J. Bacteriol.* (accepted for publication).
- Speth, D.R., Zandt, M.H., Guerrero-Cruz, S., Dutilh, B.E., and Jetten, M.S.M. (2016) Genome-based microbial ecology of anammox granules in a full-scale wastewater treatment system. *Nat. Commun.* **7**: 11172.
- Stepanauskas, R. (2012) Single cell genomics: an individual look at microbes. *Curr. Opin. Microbiol.* **15**: 613–620.
- Sukul, P., Schäfermann, S., Bandow, J.E., Kusnezowa, A., Nowrousian, M., and Leichert, L.I. (2017) Simple discovery of bacterial biocatalysts from environmental samples through functional metaproteomics. *Microbiome* **5**: 28.
- Thliveris, A.T. and Mount, D.W. (1992) Genetic identification of the DNA binding domain of *Escherichia coli* LexA protein. *Proc. Natl. Acad. Sci. U. S. A.* **89**: 4500–4504.
- Uranga, L.A., Balise, V.D., Benally, C.V., Grey, A., and Lusetti, S.L. (2011) The *Escherichia coli* DinD Protein Modulates RecA Activity by Inhibiting Postsynaptic RecA Filaments. *J. Biol. Chem.* **286**: 29480–29491.
- Vicente, M., Gomez, M.J., and Ayala, J.A. (1998) Regulation of transcription of cell division genes in the *Escherichia coli* cdx cluster. *Cell. Mol. Life Sci. CMLS* **54**: 317–324.
- Walker, G.C., Smith, B.T., and Sutton, M.D. (2000) The SOS response to DNA damage. In: Storz, G. and Hengge-Aronis, R. (eds), *Bacterial stress responses*. Washington, D.C., pp. 131–144.
- Wrighton, K.C., Castelle, C.J., Varaljay, V.A., Satagopan, S., Brown, C.T., Wilkins, M.J., et al. (2016) RubisCO of a nucleoside pathway known from Archaea is found in diverse uncultivated phyla in bacteria. *ISME J.* **10**: 2702–2714.
- Wrighton, K.C., Thomas, B.C., Sharon, I., Miller, C.S., Castelle, C.J., VerBerkmoes, N.C., et al. (2012) Fermentation, hydrogen, and sulfur metabolism in multiple uncultivated bacterial phyla. *Science* **337**: 1661–1665.

- Yaniv, K., Golberg, K., Kramarsky-Winter, E., Marks, R., Pushkarev, A., Béjà, O., and Kushmaro, A. (2017) Functional marine metagenomic screening for anti-quorum sensing and anti-biofilm activity. *Biofouling* **33**: 1–13.
- Yeoh, Y.K., Sekiguchi, Y., Parks, D.H., and Hugenholtz, P. (2016) Comparative Genomics of Candidate Phylum TM6 Suggests That Parasitism Is Widespread and Ancestral in This Lineage. *Mol. Biol. Evol.* **33**: 915–927.
- Youssef, N.H., Blainey, P.C., Quake, S.R., and Elshahed, M.S. (2011) Partial genome assembly for a candidate division OP11 single cell from an anoxic spring (Zodletone Spring, Oklahoma). *Appl. Environ. Microbiol.* **77**: 7804–7814.
- Zhang, A.P.P., Pigli, Y.Z., and Rice, P.A. (2010) Structure of the LexA-DNA complex and implications for SOS box measurement. *Nature* **466**: 883–886.

FIGURES LEGENDS

Figure 1 – (A) Sequence logos (Crooks *et al.*, 2004) for the LexA-binding motifs inferred by MEME on the promoter sequences of genes encoding LexA orthologs in the Microgenomates (top) and the Parcubacteria (bottom). (B) Representative promoter arrangement of LexA-binding motif instances in the Microgenomates (top) and the Parcubacteria (bottom). Predicted promoter elements are indicated by lines atop bold sequence. LexA-binding site sequences are underlined, with TTCGG repeat elements boxed. (C) EMSAs with purified “*Candidatus Collierbacteria* bacterium GW2011_GWB2_45_17” LexA protein on 100 bp-long oligonucleotides harboring the wild-type and site-directed mutagenesis variants of the promoter region of the gene encoding the Parcubacteria group bacterium GW2011_GWA2_48_9 LexA ortholog. In all panels, the “–” symbol denotes absence of protein and “+” the addition of 40 nM of the “*Candidatus Collierbacteria* bacterium GW2011_GWB2_45_17” LexA in the binding mixture. A black arrow indicates the retardation band created by LexA binding a LexA-binding site. A consensus logo (Anzaldi *et al.*, 2012) of the LexA-binding site indicates site conservation within the Patescibacteria LexA-binding motif, and arrows fanning from it denote site-directed mutagenesis changes introduced at each nucleotide. (D) IceLogo (Colaert *et al.*, 2009) depicting statistically significant positional differences between multiple sequence alignments of Patescibacteria and Acidobacteria LexA $\alpha 3$ helix sequences. Regular sequence logos for Patescibacteria (top) and Acidobacteria (bottom) $\alpha 3$ helix sequence alignments are superimposed on the iceLogo to facilitate the comparison. The upper part of the plot shows residues overrepresented in the Patescibacteria LexA $\alpha 3$ helix; the bottom part shows residues overrepresented in the Acidobacteria. Only differences with significant z-score under a confidence interval of 0.01 are shown.

Figure 2 – Representative instances of predicted operons with conserved LexA regulation in the Microgenomates (A) and the Parcubacteria (B). Each arrow depicts a gene within the operon. Numbers inside arrows indicate the fraction of orthologs showing LexA regulation versus the total number of orthologs detected in species where LexA shows evidence of regulation. Common gene names and the locus tag for the representative gene are shown, respectively, above and below each arrow. Gene names for genes in which binding of LexA to their promoter region has been established through EMSA (Fig. S7) are underlined. Triangles denote the position and strand of predicted LexA-binding sites. Dashed arrows illustrate low-frequency members of predicted operons. Arrow shading indicates NOG categories.

SUPPLEMENTARY MATERIAL LEGENDS

Fig. S1 – Mapping of the predicted presence/absence of LexA orthologs onto a previously reported phylogeny of the Patescibacteria (Hug *et al.*, 2016, *Nature Microbiol.*). For each clade, the figure shows the average number (\pm standard deviation) of protein sequences available for each species in the clade, the fraction of species in the clade in which LexA orthologs have been identified, the number of species with available sequences in the clade and the total number of protein sequences available for all the species in the clade.

Fig. S2 – Multiple sequence alignment of Parcubacteria promoters harboring tandem LexA-binding sites. Promoter elements (-35 and -10 sites) are shadowed and predicted LexA-binding sites are shown in bold and underlined. The predicted start of translation is boxed.

Fig. S3 – Sequence logos for aligned LexA-binding site collections and LexA $\alpha 3$ helix sequences in bacterial clades with an experimentally described LexA-binding motif. LexA-binding sites were obtained from CollecTF (Kiliç et al., 2014, *Nucl. Acids Res.*). LexA $\alpha 3$ helix sequences were obtained from (Sanchez-Alberola et al., 2015, *J. Bacteriol.*) and (Erill et al., 2016, *Front. Mol. Biosci.*).

Fig. S4 – iceLogos (Colaert et al., 2009, *Nature Methods*) comparing an alignment of Patescibacteria LexA $\alpha 3$ helices with previously reported alignments of LexA $\alpha 3$ helices in other clades (Sanchez-Alberola et al., 2015, *J. Bacteriol.*). Below each iceLogo the p-value of the TomTom comparison (Gupta et al., 2007) is reported.

Fig. S5 – Phylogeny of LexA protein sequences inferred through multiple-chain Monte Carlo sampling using MrBayes (Ronquist & Huelsenbeck, 2003, *Bioinformatics*). The figure depicts the consensus tree with estimated branch lengths and posterior probability values for each branching point. The tree was inferred from a multiple sequence alignment of LexA protein sequences obtained with T-COFFEE (Notredame et al., 2000, *J. Mol. Biol.*) and edited with Gblocks (Castresana, 2000, *Mol. Biol. Evol.*). The location of the *Acidobacterium capsulatum* LexA protein sequence is highlighted.

Fig. S6 – Heatmap plot summarizing LexA regulation across the Microgenomates and the Parcubacteria. Each row in the heatmap depicts a species with a LexA ortholog harboring the identified LexA-binding motif in its promoter sequence. Species are clustered based on a phylogenetic tree inferred with MrBayes (Ronquist & Huelsenbeck, 2003, *Bioinformatics*). Columns represent orthologous groups (NOGs). For each species and orthologous group, cell coloring indicates evidence of LexA regulation (red), lack thereof (pale red) or absence of a gene mapping to the ortholog group in that particular species (grey). Dark red denotes the presence of dual LexA-binding sites.

Fig. S7 – Electromobility shift assays with purified “*Candidatus Collierbacteria bacterium* GW2011_GWB2_45_17” LexA protein on 100 bp-long oligonucleotides harboring predicted LexA-binding sites on the promoter region of several canonical SOS genes from Microgenomates and Parcubacteria. In all cases (+) and (-) represent the presence or absence of purified “*Candidatus Collierbacteria bacterium* GW2011_GWB2_45_17” LexA protein (LexA_{Collier}), respectively. To determine the specificity of LexA binding, a 400-fold molar excess of the same unlabeled promoter was used as a specific competitor fragment in each case (C). The *E. coli recA* promoter was used as a negative control. Black arrows indicate the retardation band generated by the “*Candidatus Collierbacteria bacterium* GW2011_GWB2_45_17” LexA binding to the corresponding promoter.

Table S1 – Reference set of LexA protein sequences. The table columns designate the species, the protein identifier, the amino acid sequence and the PubMed identifier for the work reporting the experimental characterization of each LexA protein.

Table S2 – List of identified LexA homologs in the Patescibacteria identified through reciprocal BLAST using a set of experimentally-validated LexA proteins and e-value ($<1e-30$) and query coverage ($>85\%$) thresholds. The table columns designate the species, the protein identifier and the amino acid sequence of each LexA homolog.

Table S3 – LexA-binding sites for the Patescibacteria LexA-binding motif in the promoter region of *lexA* orthologs of Microgenomates and Parcubacteria species. The table shows the identified LexA-binding instance, the species name and its NCBI nucleotide accession number.

Table S4 – Multiple sequence alignment of LexA $\alpha 3$ helix sequences across bacterial clades with an experimentally described LexA-binding motif.

Table S5 – Predicted LexA-binding sites and regulated proteins in Parcubacteria and Microgenomates species with a putatively regulated LexA gene. For each putatively regulated protein, the table reports the candidate phylum and species, the genome and protein accession numbers, the predicted LexA-binding site and, when available, the matched NOG identifier and its description.

Table S6 – Summary of regulon composition for Parcubacteria and Microgenomates species with a putatively regulated LexA gene. For each putatively regulated NOG identifier, the table reports the candidate phylum, the total number of genes coding for the protein, the number of said genes with a predicted LexA-binding site and their relative fraction, as well as the NOG category and, when available, gene name for the representative gene of each NOG. Only NOGs with a proportion of putatively LexA-regulated genes larger than 10% are shown.

Table S7 – Presence or absence of DNA repair enzymes across different bacterial clades undergoing significant genome reduction. For each clade, the table provides the reported lifestyle of species in the genus, the PubMed identifier for the main manuscript reporting the lifestyle, and the total number of proteins available for that genus. The table indicates the presence or absence of canonical DNA repair enzymes as determined through a BLASTP search with minimum e-value of $1e-20$ and minimum query coverage of 80%, using *E. coli* and *B. subtilis* protein sequences as a query. The clades included in the table were selected based on the results of a composite search for small complete genomes <txid2[Organism:exp] AND "complete genome"[title] AND (bacteria[filter] AND biomol_genomic[PROP] AND refseq[filter] AND ("500000"[SLEN] : "1000000"[SLEN]))> in the NCBI nucleotide database.

Table S8 – List of oligonucleotides used in this work.

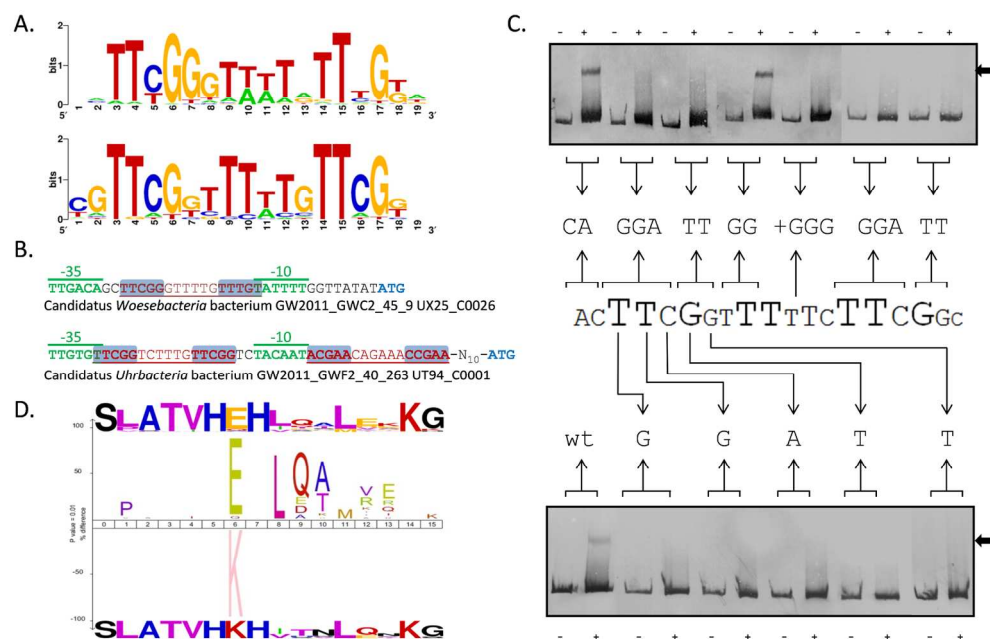


Figure 1 – (A) Sequence logos (Crooks et al., 2004) for the LexA-binding motifs inferred by MEME on the promoter sequences of genes encoding LexA orthologs in the Microgenomates (top) and the Parcubacteria (bottom). (B) Representative promoter arrangement of LexA-binding motif instances in the Microgenomates (top) and the Parcubacteria (bottom). Predicted promoter elements are indicated by lines atop bold sequence. LexA-binding site sequences are underlined, with TTCGG repeat elements boxed. (C) EMSAs with purified “*Candidatus Collierbacteria bacterium GW2011_GWB2_45_17*” LexA protein on 100 bp-long oligonucleotides harboring the wild-type and site-directed mutagenesis variants of the promoter region of the gene encoding the Parcubacteria group bacterium GW2011_GWA2_48_9 LexA ortholog. In all panels, the “–” symbol denotes absence of protein and “+” the addition of 40 nM of the “*Candidatus Collierbacteria bacterium GW2011_GWB2_45_17*” LexA in the binding mixture. A black arrow indicates the retardation band created by LexA binding a LexA-binding site. A consensus logo (Anzaldi et al., 2012) of the LexA-binding site indicates site conservation within the Patescibacteria LexA-binding motif, and arrows fanning from it denote site-directed mutagenesis changes introduced at each nucleotide. (D) IceLogo (Colaert et al., 2009) depicting statistically significant positional differences between multiple sequence alignments of Patescibacteria and Acidobacteria LexA $\alpha 3$ helix sequences. Regular sequence logos for Patescibacteria (top) and Acidobacteria (bottom) $\alpha 3$ helix sequence alignments are superimposed on the iceLogo to facilitate the comparison. The upper part of the plot shows residues overrepresented in the Patescibacteria LexA $\alpha 3$ helix; the bottom part shows residues overrepresented in the Acidobacteria. Only differences with significant z-score under a confidence interval of 0.01 are shown.

152x98mm (300 x 300 DPI)

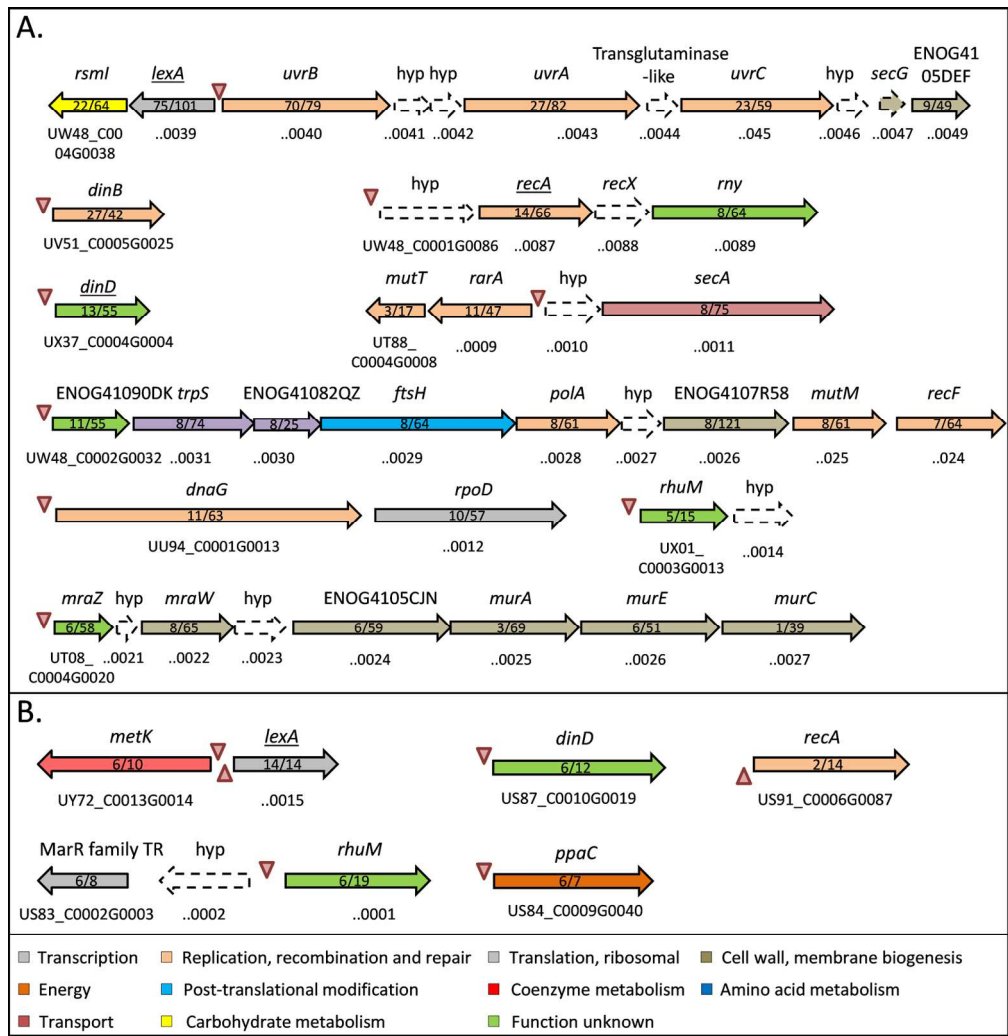


Figure 2 – Representative instances of predicted operons with conserved LexA regulation in the Microgenomates (A) and the Parcubacteria (B). Each arrow depicts a gene within the operon. Numbers inside arrows indicate the fraction of orthologs showing LexA regulation versus the total number of orthologs detected in species where LexA shows evidence of regulation. Common gene names and the locus tag for the representative gene are shown, respectively, above and below each arrow. Gene names for genes in which binding of LexA to their promoter region has been established through EMSA (Fig. S7) are underlined. Triangles denote the position and strand of predicted LexA-binding sites. Dashed arrows illustrate low-frequency members of predicted operons. Arrow shading indicates NOG categories.

169x174mm (300 x 300 DPI)