# Neo-formation of chromosomes in bacteria

Olivier B. Poirion[1,2†] & Bénédicte Lafay[1,2,3*]

[1] Université de Lyon, F-69134 Lyon, France

[2] CNRS (French National Center for Scientific Research) UMR5005, Laboratoire Ampère, École Centrale de Lyon, 36 avenue Guy de Collongue, 69134 Écully, France

[3] CNRS (French National Center for Scientific Research) UMR5558, Laboratoire de Biométrie et Biologie Évolutive, Université Claude Bernard – Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France

[†] Current address: Center for Epigenomics, Department of Cellular and Molecular Medicine, University of California, San Diego, School of Medicine, 9500 Gilman Drive, La Jolla, CA 92093, USA

* Author for correspondence: benedicte.lafay@univ-lyon1.fr

## ABSTRACT

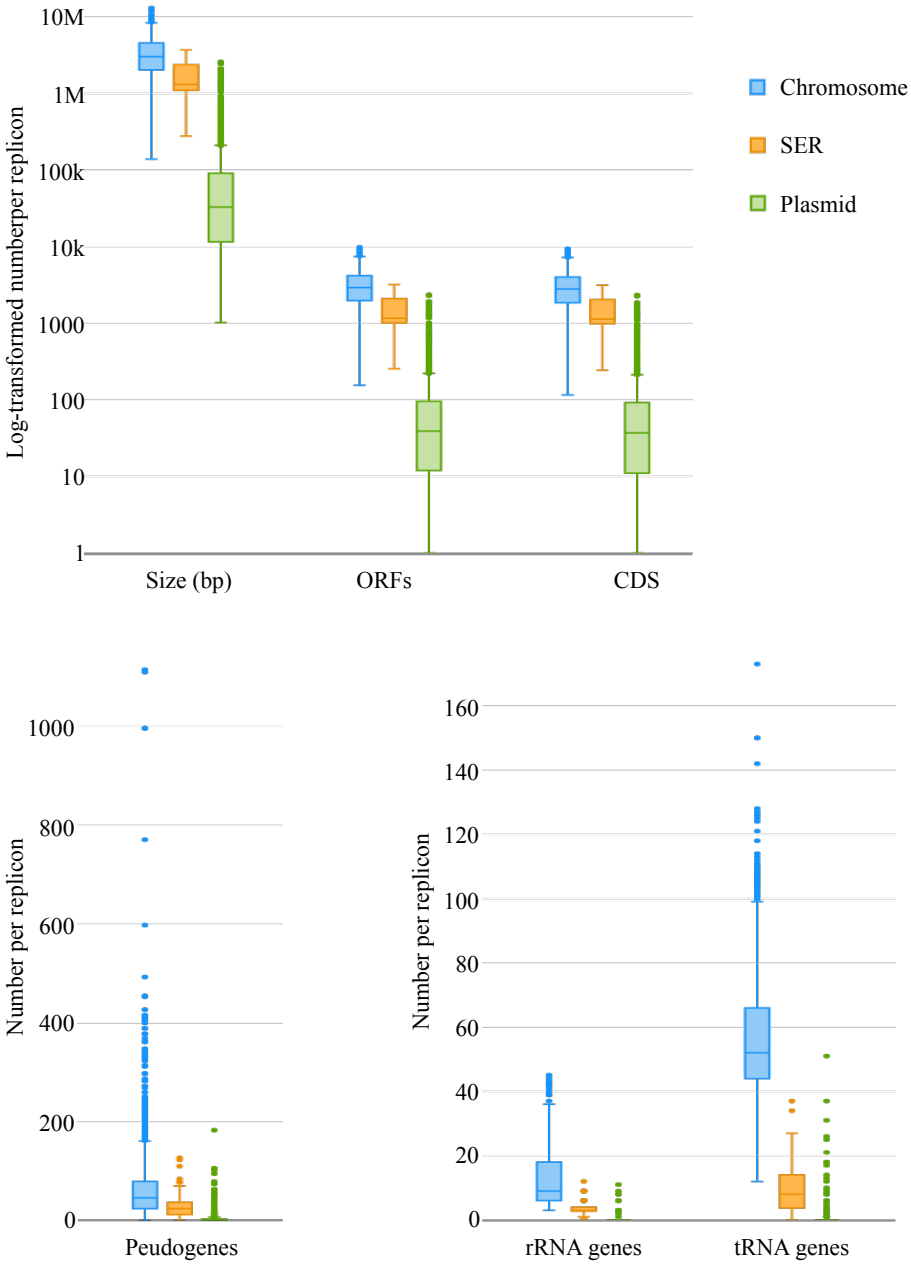
Although the bacterial secondary chromosomes/megaplasmids/chromids, first noticed about forty years ago, are commonly held to originate from stabilized plasmids, their true nature and definition are yet to be resolved. On the premise that the integration of a replicon within the cell cycle is key to deciphering its essential nature, we show that the content in genes involved in the replication, partition and segregation of the replicons and in the cell cycle discriminates the bacterial replicons into chromosomes, plasmids, and another class of essential genomic elements that function as chromosomes. These latter do not derive directly from plasmids. Rather, they arise from the fission of a multi-replicon molecule corresponding to the co-integrated and rearranged ancestral chromosome and plasmid. All essential replicons in a distributed genome are thus neo-chromosomes. Having a distributed genome appears to extend and accelerate the exploration of the bacterial genome evolutionary landscape, producing complex regulation and leading to novel eco-phenotypes and species diversification.

27 **INTRODUCTION**

28 Chromosomes are the only components of the genome that encode the necessary
29 information for replication and life of the cell/organism under normal growth conditions.
30 Their number varies across taxa, a single chromosome being the standard in bacteria
31 (Krawiec and Riley, 1990). Evidence accumulated over the past forty years is proving
32 otherwise: bacterial genomes can be distributed on more than one chromosome-like
33 autonomously replicating genomic element (replicon) (Casjens, 1998; diCenzo and
34 Finan, 2017; Mackenzie et al., 2004). The largest, primary, essential replicon (ER) in a
35 multipartite genome corresponds to a *bona fide* chromosome and the additional,
36 secondary, ERs (SERs) are expected to derive from accessory replicons (plasmids
37 (Lederberg, 1998)). The most popular model of SER formation posits that a plasmid
38 acquired by a mono-chromosome progenitor bacterium is stabilized in the genome
39 through the transfer from the chromosome of genes essential to the cell viability
40 (diCenzo and Finan, 2017; diCenzo et al., 2013; Slater et al., 2009). The existence in
41 SERs of plasmid-like replication and partition systems (Dubarry et al., 2006; Egan and
42 Waldor, 2003; Livny et al., 2007; MacLellan et al., 2004, 2006; Slater et al., 2009;
43 Yamaichi et al., 2007) as well as experimental results (diCenzo et al., 2014) support this
44 view. Yet, the duplication and maintenance processes of SERs contrast with the typical
45 behaviour of plasmids for which both the timing of replication initiation and the
46 centromere movement are random (Million-Weaver and Camps, 2014; Reyes-Lamothe
47 et al., 2014). Indeed, the SERs share many characteristic features with chromosomes:
48 enrichment in Dam methylation sites of the replication origin (Egan and Waldor, 2003;
49 Gerding et al., 2015), presence of initiator titration sites (Egan and Waldor, 2003;
50 Venkova-Canova and Chattoraj, 2011), synchronization of the replication with the cell
51 cycle (De Nisco et al., 2014; Deghelt et al., 2014; Egan and Waldor, 2003; Egan et al.,

52 2004; Fiebig et al., 2006; Frage et al., 2016; Kahng and Shapiro, 2003; Rasmussen et al.,
53 2007; Srivastava et al., 2006; Stokke et al., 2011), KOPS-guided FtsK translocation (Val
54 et al., 2008), FtsK-dependent dimer resolution system (Val et al., 2008), MatP/matS
55 macrodomain organisation system (Demarre et al., 2014), and similar fine-scale
56 segregation dynamics (Fiebig et al., 2006; Frage et al., 2016). Within a multipartite
57 genome, the replication of the chromosome and that of the SER(s) are initiated at
58 different time points (De Nisco et al., 2014; Deghelt et al., 2014; Fiebig et al., 2006;
59 Frage et al., 2016; Rasmussen et al., 2007; Srivastava et al., 2006; Stokke et al., 2011),
60 and use replicon-specific systems (Drevinek et al., 2008; Egan and Waldor, 2003;
61 Galardini et al., 2013; MacLellan et al., 2004, 2006; Slater et al., 2009). Yet, they are
62 coordinated, hence maintaining the genome stoichiometry (Deghelt et al., 2014; Egan et
63 al., 2004; Fiebig et al., 2006; Frage et al., 2016; Stokke et al., 2011). In the few species
64 where this was studied, the replication of the SER is initiated after that of the
65 chromosome (De Nisco et al., 2014; Deghelt et al., 2014; Fiebig et al., 2006; Frage et al.,
66 2016; Rasmussen et al., 2007; Srivastava, 2006; Stokke et al., 2011) under various
67 modalities. In the Vibrionaceae, the replication of a short region of the chromosome
68 licenses the SER duplication (Baek and Chattoraj, 2014; Kemter et al., 2018), and the
69 advancement of the SER replication and segregation triggers the divisome assembly
70 (Galli et al., 2016). In turn, the altering of the chromosome replication does not affect
71 the replication initiation control of the SER in α-proteobacterium *Ensifer/Sinorhizobium*
72 *meliloti* (Frage et al., 2016).
73 Beside the exploration of the replication/segregation mechanistic, studies of multipartite
74 genomes, targeting a single bacterial species or genus (diCenzo et al., 2013, 2014;
75 Dubarry et al., 2006; Mackenzie et al., 2004; Slater et al., 2009) or using a more
76 extensive set of taxa (diCenzo and Finan, 2017; Harrison et al., 2010), relied on

77 inadequate (replicon size, nucleotide composition, coding of core essential genes for
78 growth and survival (diCenzo and Finan, 2017; Harrison et al., 2010; Liu et al., 2015);
79 Figure 1) and/or oriented (presence of plasmid-type systems for genome maintenance
80 and replication initiation (Harrison et al., 2010)) criteria to characterize the SERs.




81

82 **Figure 1. Structural features of the replicons**

83 Boxplots of the lengths (base pairs) and numbers of genes (ORFs), protein-coding genes (CDS), pseudogenes,
84 ribosomal RNA genes and transfer RNA genes for the 2016 chromosomes (blue), 129 SERs (orange), and 2783
85 plasmids (green) included in the final dataset (4928 replicons).

86 While clarifying the functional and evolutionary contributions of each type of replicon
87 to a multipartite genome in given bacterial lineages (Galardini et al., 2013; Harrison et
88 al., 2010; MacLellan et al., 2004; Slater et al., 2009), these studies produced no absolute
89 definition of SERs (diCenzo and Finan, 2017; Harrison et al., 2010) or universal model
90 for their emergence (diCenzo and Finan, 2017; diCenzo et al., 2013, 2014; Galardini et
91 al., 2013; Harrison et al., 2010). We thus set out investigating the nature(s) and origin(s)
92 of these replicons using as few assumptions as possible.

93 **RESULTS**

94 **Replicon inheritance systems as diagnostic features**

95 We did not limit our study to a particular multipartite genome or a unique gene family.
96 Rather, we performed a global analysis encompassing all bacterial replicons whose
97 complete sequence was available in public sequence databases (Figure 2). We reasoned
98 that the key property discriminating the chromosomes from the plasmids is their
99 transmission from mother to daughter cells during the bacterial cell cycle. The functions
100 involved in the replication, partition and maintenance of a replicon, *i.e.*, its inheritance
101 systems (ISs), thence are expected to reflect the replicon degree of integration into the
102 host cycle.
103 We first faced the challenge of identifying all IS functional homologues. The inheritance
104 of genetic information requires functionally diverse and heterogeneous actuators
105 depending on the replicon type and the characteristics of the organism. Also, selecting
106 sequence orthologues whilst avoiding false positives (*e.g.*, sequence paralogues) can be
107 tricky since remote sequence homology most likely prevails among
108 chromosome/plasmid protein-homologue pairs.

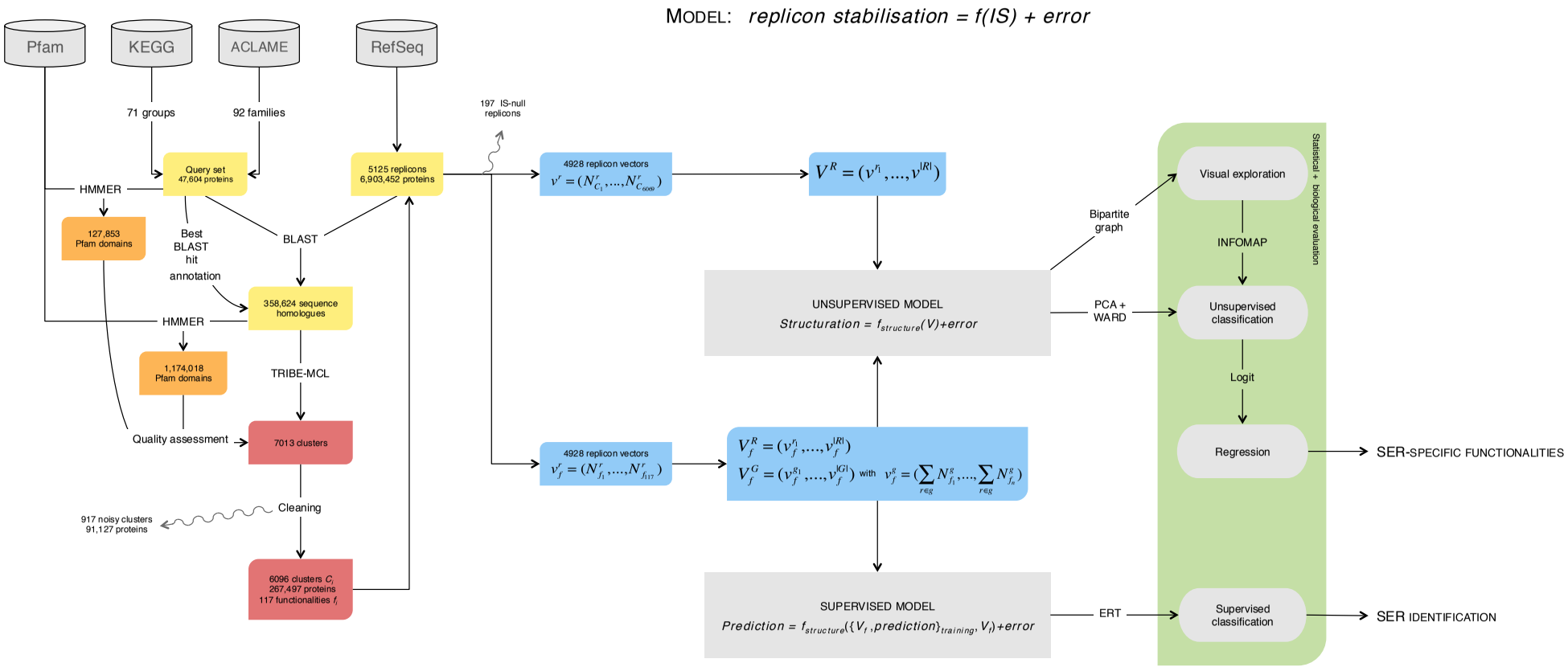



**Figure 2. Analytical procedure**

110 Starting from an initial dataset of 5125 replicons, we identified 358,624 putative IS
111 functional homologues, overall corresponding to 1711 Pfam functional domains (Figure
112 3a), using a query set of 47,604 chromosomal and plasmidic IS-related proteins selected
113 from the KEGG and ACLAME databases (Tables 1 and 2).

114 **Table 1. ACLAME families used in the building of the query set**

| PROCESS | FAMILY | PROTEIN DESCRIPTION |
|---|---|---|
| | *32* | RepB, pi, initiator protein, RepE, RepA |
| | *76* | Rep, RepB, Rep of rolling circle initiator, RepA, RepU, OrfB, Rep2 |
| | *107* | RepC, RepCa1, RepCa2, RepCd |
| | *114* | Helicase, UrvD rep helicase, helicase super family 1, Yga2F, helicase II |
| | *118* | CdsE, CdsJ |
| | *133* | RepA, W0005, RepA1/A2 |
| | *171* | RepA, RepB, putative theta replicative protein |
| | *207* | replicative DNA helicase, DnaB, pGP1 |
| | *208* | RepA, W0013, W0041, RepFIB |
| | *224* | long form TrfA, TrfA1, TrfA2, S-TrfA, plasmid initiation protein |
| | *237* | RepA, putative RepA, truncated RepA |
| | *244* | RepA, RepB, CopB, repA1/A2, w0004 |
| | *294* | Rop regulatory protein, RNAI modulator, RNA modulator, plasmid copy number control |
| | *297* | primase activity/DNA initiation, LtrC/LtrC-like hypothetical protein, PcfD |
| | *330* | DNA repair/ DNA helicase, type III restriction enzyme, res subunit, DEAD/DEAH box helicase |
| | *377* | replicase, replication initiation, RepC, RepJ, RepE, RepL |
| | *383* | RepA, Rb100 |
| | *404* | RepA,RepB,RepW |
| Replication | *412* | Rep, RepA |
| | *423* | truncated RCR replication, RepRC, RepB, OrfA |
| | *426* | cell division control protein 6 homolog |
| | *440* | Rep 14-4, rm protein, RepA hypothetical protein |
| | *451* | RepA, host type : *Corynebacterium* |
| | *477* | Rep, RepS, RepE, host type : *Bacillus*, RepS, RepR |
| | *612* | RepL, replication initiation |
| | *775* | DNA helicase activity, RepA, putative helicase |
| | *854* | DNA helicase activity, RepC, putative initiator protein |
| | *921* | RepA |
| | *931* | DNA replication initiation, putative protein, CdsD |
| | *1005* | helicase activity, putative protein, hypothetical helicase |
| | *1055* | RNA polymerase σ factor, σ 70 family, bacteriocin uviA, sigF/V/G, tetR, host type : *Clostridium* |
| | *1095* | DNA repair/helicase, RuvB, DNA pol III γ and τ subunits, DNA pol δ subunit |
| | *1099* | putative theta replicase, RepB, Rep2 |
| | *1187* | DNA replication, RepH, RepI |
| | *1288* | RepA |
| | *1345* | DNA primase activity, DNA primase , primase CHC2 family |
| | *1398* | helicase activity, GcrE, GcrC |

| | | |
|---|---|---|
| | *1652* | DNA repair/exonuclease activity, DNA exonuclease protein, SbcCD related protein |
| | *1837* | putative replication protein |
| | *2881* | RepC-like, Pif |
| Partition | *4* | plasmid partition protein, ParA, ParA IncC protein, ParA InC1/ IncC2, SopA, virC1 |
| | *14* | RepB, RepB partitionning, KorB repressor and partitionning, ParB-like domain, YefA, YdeB, ParB, ParB-like |
| | *102* | DNA binding, partitionning protein, control protein, ParB, VirB, partition protein B |
| | *128* | DNA segregation/DNA translocase activity, cell division FtsK/ SpoIIIE, SpoI, TraB |
| | *289* | ParM family, go : translocase, hypothetical protein, rode shape protein, putative ATPase of class HSP70 |
| | *316* | microfilament motor activity, ParM family, StbA protein, stable inheritance protein, ParA |
| | *318* | ATPase, regulation of cell division, chromosome patition, GumC, ExoP related protein, EpsB, MPA1 family |
| | *427* | ATPase family, ParR family, ParB, StbB, mediator of plasmid stability |
| | *875* | DNA binding, partitionning protein family ParB/Spo0J, YPMT1.28c |
| | *876* | DNA binding, partitionning protein family ParB/Spo0J, YPMT1.29c |
| | *983* | DNA binding, ParB, CopG |
| | *1227* | DNA plasmid copy number control, CopG |
| | *2158* | RepC |
| | *2894* | DNA binding |
| Dimer resolution | *5* | serine based recombinase activity, ylb, resolvase, second invertase, TniR, ParA |
| | *10* | tyrosine-based recombinase, integrase, putative integrase, Xer, recombinase-like SAM |
| | *101* | plasmid dimer resolution, tyrosine-based recombinase, yld, SAM-like protein |
| | *170* | tyrosine-based recombinase, OrfA, hypothetical protein |
| | *589* | tyrosine based protein, Fis protein |
| | *688* | tyrosine based protein, SAM like protein, XerD |
| Maintenance | *100* | Postsegregational killing system vapBC/vag |
| | *136* | Postsegregational killing system parDE |
| | *156* | Postsegregational killing system epsilon-zeta |
| | *201* | Postsegregational killing system higBA |
| | *212* | Postsegregational killing system parDE |
| | *293* | Postsegregational killing system mazEF |
| | *319* | Postsegregational killing system relBE |
| | *326* | Postsegregational killing system mazEF |
| | *335* | Postsegregational killing system HOK/SOK |
| | *338* | Postsegregational killing system parDE |
| | *356* | Postsegregational killing system parDE |
| | *366* | Postsegregational killing system vapBC/vag |
| | *380* | Postsegregational killing system phD-doc |
| | *428* | Postsegregational killing system ccd |
| | *470* | Postsegregational killing system yacA |
| | *474* | Postsegregational killing system relBE |
| | *515* | Postsegregational killing system relBE |
| | *556* | Postsegregational killing system higBA |
| | *563* | Postsegregational killing system ccd |
| | *588* | Postsegregational killing system higBA |
| | *677* | Postsegregational killing system higBA |
| | *798* | Postsegregational killing system mazEF |
| | *916* | Postsegregational killing system relBE |
| | *1031* | Postsegregational killing system HOK/SOK |
| | *1180* | Postsegregational killing system vapXD |
| | *1308* | Postsegregational killing system HicAB |
| | *1559* | Postsegregational killing system epsilon-zeta |
| | *1927* | Postsegregational killing system mazEF |

| | |
|---|---|
| *3357* | Postsegregational killing system, plasmid maintenance |
| *4776* | Postsegregational killing system, parC |
| *4777* | Postsegregational killing system parDE, parD |
| *16584* | Postsegregational killing system vapXD |

**Table 2. KEGG "Prokaryotic-type chromosome" orthology groups used in the building of the query set**

| BRITE HIERARCHY | | KEGG ENTRY | NAME | DEFINITION |
|---|---|---|---|---|
| Chromosome replication | Initiation factors (bacterial) | *K02313* | DnaA | chromosomal replication initiator protein |
| | | *K02314* | DnaB | replicative DNA helicase [EC:3.6.4.12] |
| | | *K03346* | DnaB2, DnaB | replication initiation and membrane attachment protein |
| | | *K02315* | DnaC | DNA replication factor, helicase loader |
| | | *K02316* | DnaG | DNA primase [EC:2.7.7.-] |
| | | *K11144* | DnaI | primosomal protein DnaI |
| | | *K05787* | HupA | DNA-binding protein HU-alpha |
| | | *K03530* | hupB | DNA-binding protein HU-beta |
| | | *K04764* | IhfA, HimA | integration host factor subunit alpha |
| | | *K05788* | IhfB, HimD | integration host factor subunit beta |
| | | *K03111* | ssb | single-strand DNA-binding protein |
| | Terminus site-binding protein | *K10748* | Tus, Tau | DNA replication terminus site-binding protein |
| | DNA methylation enzyme | *K06223* | Dam | DNA adenine methylase [EC:2.1.1.72] |
| | Prevention of re-replication factors | *K10763* | Hda | DnaA-homolog protein |
| | | *K03645* | SeqA | negative modulator of initiation of replication |
| Chromosome partition | MukBEF complex | *K03632* | MukB | chromosome partition protein MukB |
| | | *K03804* | MukE | chromosome partition protein MukE |
| | | *K03633* | MukF | chromosome partition protein MukF |
| | Condensin-like complex | *K03529* | Smc | chromosome segregation protein |
| | | *K05896* | ScpA | segregation and condensation protein A |
| | | *K06024* | ScpB | segregation and condensation protein B |
| | Divisome proteins | *K03585* | AcrA | membrane fusion protein |
| | | *K01448* | AmiA,AmiB, AmiC | N-acetylmuramoyl-L-alanine amidase [EC:3.5.1.28] |
| | | *K13052* | DivIC, DivA | cell division protein DivIC |
| | | *K03590* | FtsA | cell division protein FtsA |
| | | *K05589* | FtsB | cell division protein FtsB |
| | | *K09812* | FtsE | cell division transport system ATP-binding protein |
| | | *K03587* | FtsI | cell division protein FtsI [EC:2.4.1.129] |
| | | *K03466* | FtsK, SpoIIIE | DNA segregation ATPase FtsK/SpoIIIE, S-DNA-T family |
| | | *K03586* | FtsL | cell division protein FtsL |
| | | *K03591* | FtsN | cell division protein FtsN |
| | | *K03589* | FtsQ | cell division protein FtsQ |
| | | *K03588* | FtsW, SpoVE | cell division protein FtsW |
| | | *K09811* | FtsX | cell division transport system permease protein |
| | | *K03531* | FtsZ | cell division protein FtsZ |
| | | *K09888* | ZapA | cell division protein ZapA |
| | | *K03528* | ZipA | cell division protein ZipA |
| | Inhibitors of FtsZ assembly | *K04074* | DivIVA | cell division initiation protein |
| | | *K06286* | EzrA | septation ring formation regulator |

| | | | | |
|---|---|---|---|---|
| | | *K03610* | MinC | septum site-determining protein MinC |
| | | *K03609* | MinD | septum site-determining protein MinD |
| | | *K03608* | MinE | cell division topological specificity factor |
| | | *K05501* | SlmA, Ttk | TetR/AcrR family transcriptional regulator |
| | | *K09772* | SepF | cell division inhibitor SepF |
| | | *K13053* | SulA | cell division inhibitor, FtsZ assembly inhibitor |
| | Other chromosome partitioning proteins | *K04095* | Fic | cell filamentation protein |
| | | *K04094* | Gid, TrmFO | methylenetetrahydrofolate--tRNA-[uracil-5-)-methyltransferase [EC:2.1.1.74] |
| | | *K03495* | GidA, MnmG, MTO1 | tRNA uridine 5-carboxymethylaminomethyl modification enzyme |
| | | *K03501* | GidB, RsmG | 16S rRNA [guanine527-N7)-methyltransferase [EC:2.1.1.170] |
| | | *K03569* | MreB | rod shape-determining protein MreB and related proteins |
| | | *K03570* | MreC | rod shape-determining protein MreC |
| | | *K03571* | MreD | rod shape-determining protein MreD |
| | | *K03593* | Mrp | ATP-binding protein involved in chromosome partitioning |
| | | *K03496* | ParA, Soj | chromosome partitioning protein |
| | | *K03497* | ParB, Spo0J | chromosome partitioning protein, ParB family |
| | | *K02621* | ParC | topoisomerase IV subunit A [EC:5.99.1.-] |
| | | *K02622* | ParE | topoisomerase IV subunit B [EC:5.99.1.-] |
| | | *K11686* | RacA | chromosome-anchoring protein RacA |
| | | *K05837* | RodA, MrdB | rod shape determining protein RodA |
| | | *K03645* | SeqA | negative modulator of initiation of replication |
| | | *K03733* | XerC | integrase/recombinase XerC |
| | | *K04763* | XerD | integrase/recombinase XerD |
| Nucleoid | HNS (histone-like nucleoid structuring protein) | *K03746* | H-NS | DNA-binding protein H-NS |
| | | *K11685* | StpA | DNA-binding protein StpA |
| | HU (heat unstable protein) | *K05787* | HupA | DNA-binding protein HU-alpha |
| | | *K03530* | HupB | DNA-binding protein HU-beta |
| | IHF (integration host factor) | *K04764* | IhfA, HimA | integration host factor subunit alpha |
| | | *K05788* | IhfB, HimD | integration host factor subunit beta |
| | Other nucleoid associated proteins | *K05516* | CbpA | curved DNA-binding protein |
| | | *K12961* | DiaA | chromosomal replication initiator protein |
| | | *K02313* | DnaA | DnaA initiator-associating protein |
| | | *K04047* | Dps | starvation-inducible DNA-binding protein |
| | | *K03557* | Fis | Fis family transcriptional regulator, factor for inversion stimulation protein |
| | | *K03666* | Hfq | host factor-I protein |
| | | *K05596* | IciA | chromosome initiation inhibitor, LysR family transcriptional regulator |
| | | *K03719* | Lrp | leucine-responsive regulatory protein, Lrp/AsnC family transcriptional regulator |
| | | *K05804* | Rob | right origin-binding protein, AraC family transcriptional regulator |

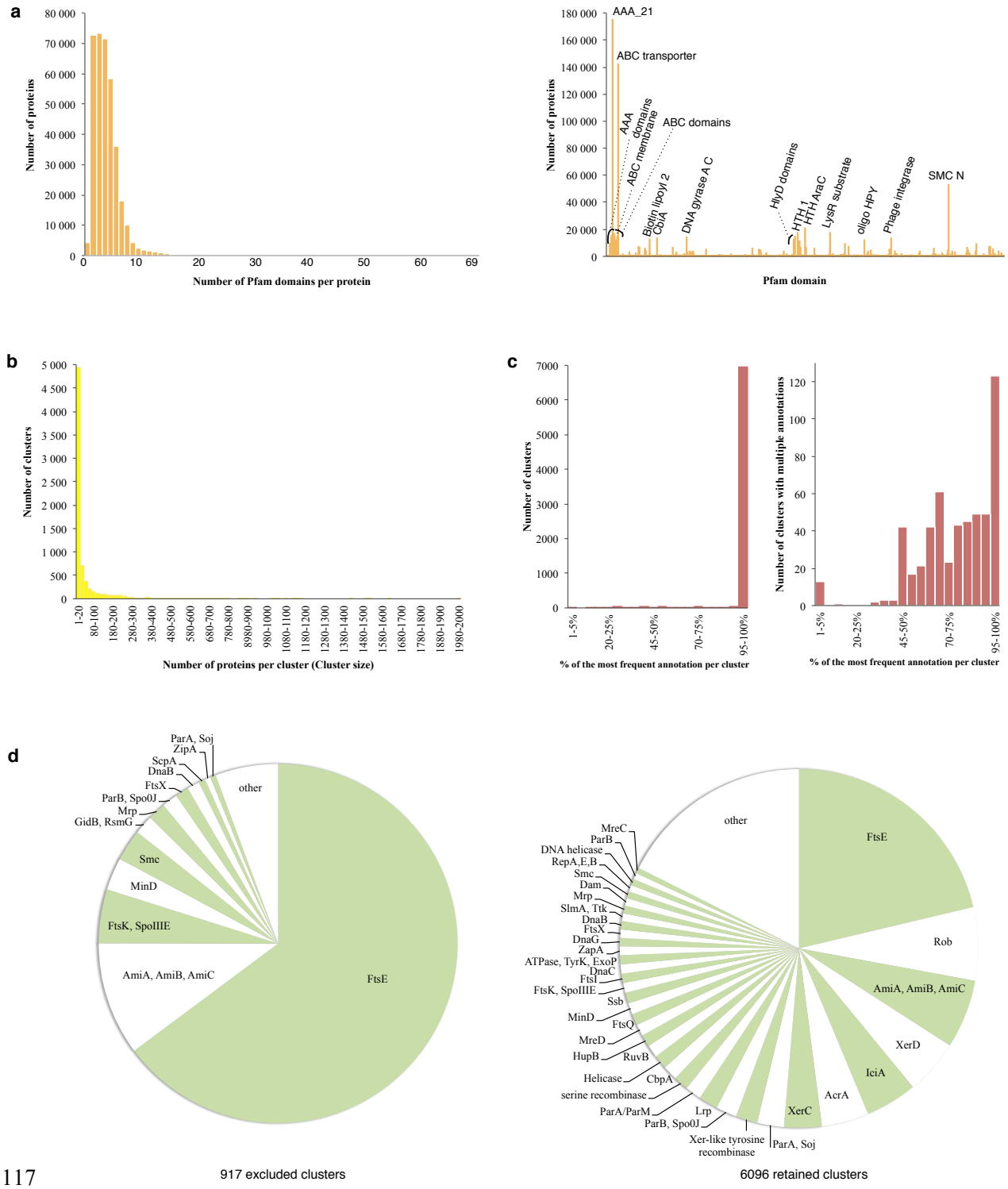



**Figure 3. Properties of the IS clustering**

(a) Frequency distribution of the 358,624 putative IS protein homologues according to their number of functional domains (0 to 69) per protein (left), and occurrences of the 1711 functional Pfam domains (right). The 20 top most frequently encountered functional domains are indicated. (b) Size distribution of the 7013 clusters, each comprising from a single to 1990 proteins. (c) Percentage distribution of the most frequent annotation per cluster among all clusters (left) and among clusters with multiple annotations (right). (d) Distribution of the most frequent annotation per cluster among the 917 excluded clusters (left) and the 6096 clusters retained for the analysis (right).

125 We then inferred 7013 homology groups using a clustering procedure and named the
126 clusters after the most frequent annotation found among their proteins (Figure 3b,c). Most
127 clusters were characterized by a single annotation whilst the remaining few (4.7%) each
128 harbored from 2 to 710 annotations, the most frequent annotation in a cluster generally
129 representing more than half of all annotations (Figure 3c). The removal of false positives
130 left 267,497 IS protein homologues distributed in 6096 clusters (Figure 3d) and coded by
131 4928 replicons out of the initial replicon dataset. Following the Genbank/RefSeq
132 annotations, our final dataset comprised 2016 complete genome sets corresponding to
133 3592 replicons (2016 chromosomes, 129 SERs, and 1447 plasmids) and 1336 plasmid
134 genomes (Supplementary table 1), irregularly distributed across the bacterial phylogeny
135 (Figure 4a). Multi-ER genomes are observed in 5.0% of all represented bacterial genera
136 and constitute 5.7% of the complete genomes (averaged over genera) available at the time
137 of study (Figure 4b). They are merely incidental (0.2% in Firmicutes) or reach up to
138 almost one third of the genomes (30.1% in β-Proteobacteria) depending on the lineage,
139 and are yet to be observed in most bacterial phyla, possibly because of the poor
140 representation of some lineages. Although found in ten phyla, they occur more than once
141 per genus in only three of them: Bacteroidetes, Proteobacteria and Spirochaetae.

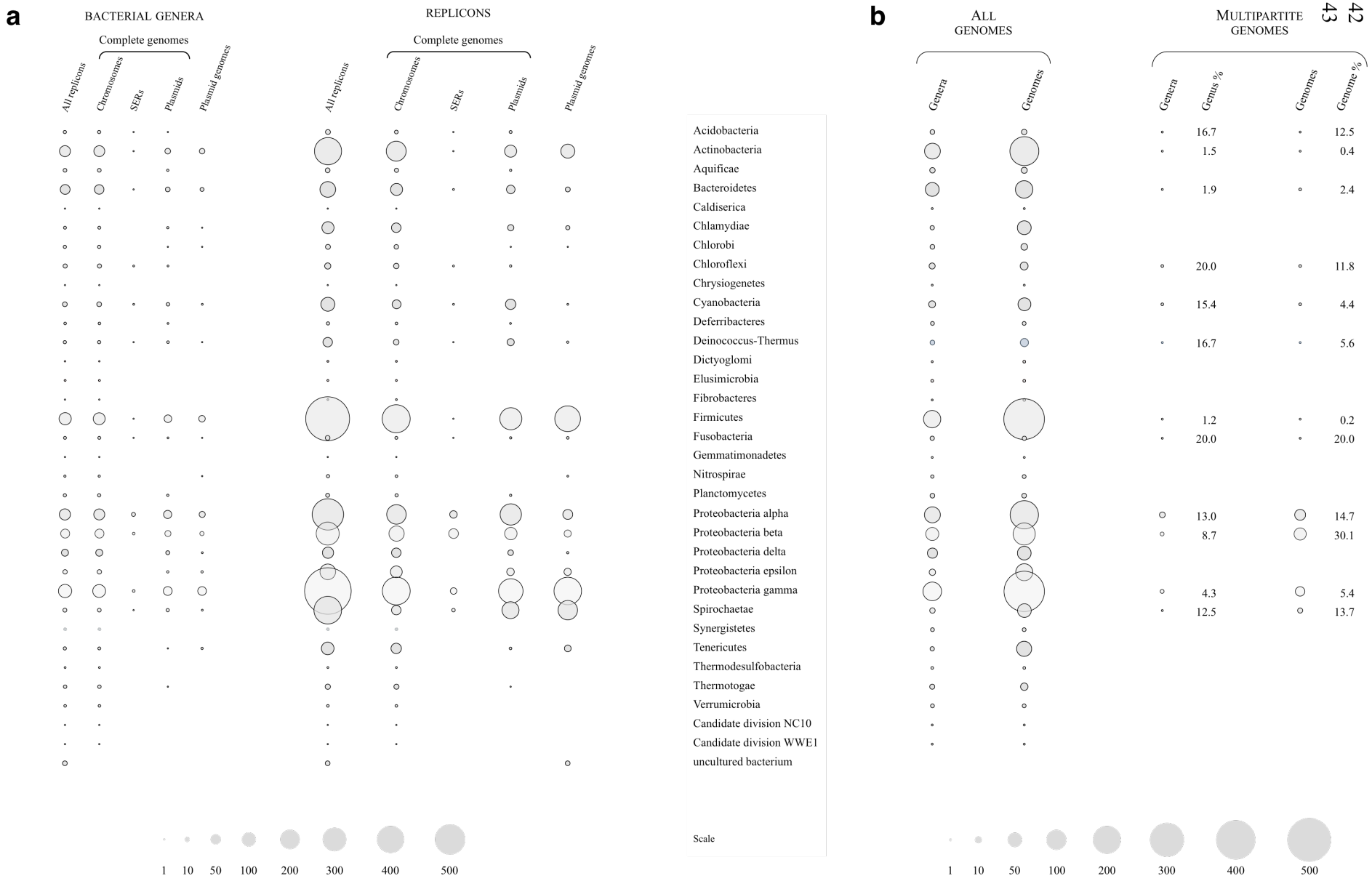



**Figure 4. Taxonomic structure of the replicon dataset**

Numbers of replicons (a) and complete genomes (b), and represented bacterial genera are shown according to datasets and host taxonomy. Surfaces represent the numbers of bacterial genera, replicons or genomes within each category. Percentages of multipartite genomes and corresponding bacterial genera are calculated for each host phylum or class (Proteobacteria).

144 **Exploration of the replicon diversity**

145 We explored the differences and similarities of the bacterial replicons with regard to
146 their IS usage using a data mining and machine learning approach (Methods). The 6096
147 retained IS clusters were used as distinct variables to ascribe each of the 4928 replicons
148 with a vector according to its IS usage profile. We transformed these data into bipartite
149 graphs depending on the number of proteins from the IS clusters coded by each replicon.
150 Bipartite graphs display both the vectors (replicons) and the variables (protein clusters)
151 together with their respective connections, and allow the interactive exploration of the
152 data. The majority of the replicons are interconnected (Figure 5) as testimony of the
153 shared evolutionary history of their IS sequences. Chromosomes and plasmids form
154 overall distinct groups and communities with varying degree of connectivity depending
155 on their functional specificities (Figure 5a) as well as on the bacterial taxonomy of their
156 hosts (Figure 5c). They nonetheless share many ISs, bearing witness to the continuity of
157 the genomic material and the extensive exchange of genetic material within bacterial
158 genomes. The occurrence of poorly IS cluster-connected plasmids within a group of
159 chromosomes did not consistently reflect a true relationship and rather resulted from
160 shared connections to a very small number (as low as one) of common ISs. While being
161 interconnected to both chromosomes and plasmids *via* numerous IS clusters, the SERs
162 generally stand apart from either these types of replicons and gather at the chromosome-
163 plasmid interface (Figure 5a,b). Their IS usage is neither chromosome-like nor plasmid-
164 like, suggesting that they may constitute a separate category of replicons. This is most
165 tangible in the case of the proteobacterial lineages where SERs occur most frequently
166 (top of Figure 5b).

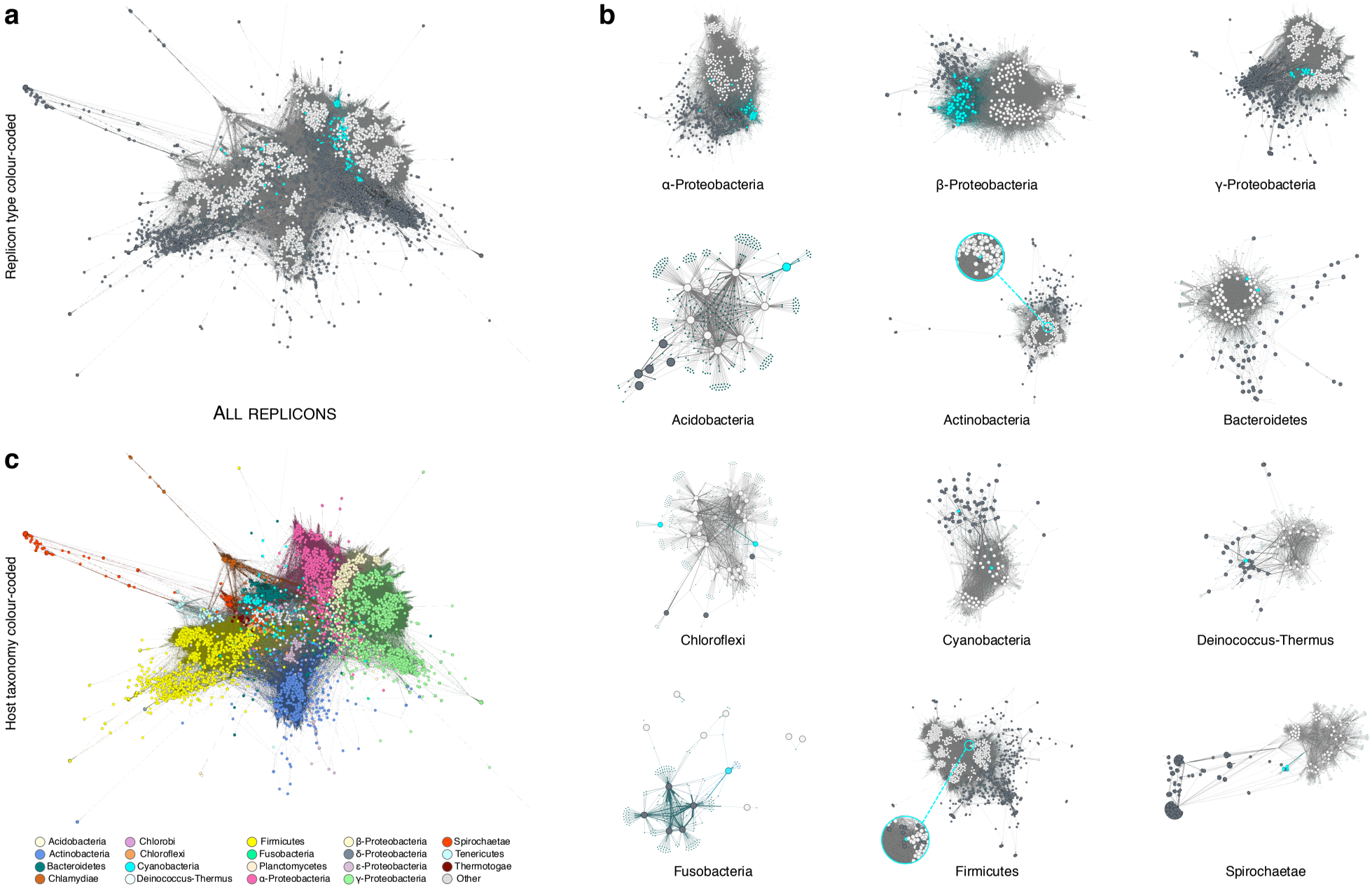



**Figure 5. Visualisation of the replicon IS-based relationshipss**

Gephi-generated bipartite-graphs for the whole dataset (a and b) or groups of replicons following the host taxonomy (c). Nodes correspond to the replicons (large dots) or the clusters of IS proteins (small dots). Edges linking replicons and protein clusters reflect the presence on a replicon of at least one protein of a protein cluster. Colouring according to replicon type (a and c): chromosomes (white), plasmids (grey), and SERs (blue), or according to host taxonomy (b).

169 All SERs in the β- and γ-Proteobacteria, and most in the α-Proteobacteria are linked to
170 remarkable chromosome-type IS clusters, such as AcrA, IciA, FtsE, HN-S and Lrp, as
171 well as to plasmid-like ParA/ParB, Rep and PSK IS clusters. A similar pattern is
172 observed for the SERs in actinobacterium *Nocardiopsis dassonvillei,* firmicute
173 *Butyrivibrio proteoclasticus*, and chloroflexi *Sphaerobacter thermophilus* and
174 *Thermobaculum terrenum* (Figure 5b). Interestingly, DNA primase DnaG-annotated
175 clusters connect the SERs present in all but one *Burkholderia* species (β-Proteobacteria)
176 as well as the chromosomes of all other bacteria. Since the sole exception,
177 *B. rhizoxinica*, possesses a SER-less reduced genome as an adaption to intracellular life,
178 the *Burkholderia* SERs likely originated from a single event prior to the diversification
179 of the genus, possibly in relation to the speciation event that gave rise to this lineage.
180 The second SERs harbored by only some *Burkholderia* species exhibit a higher level of
181 interconnection to plasmids, as do the SERs in α-proteobacterium *Sphingobium*,
182 cyanobacterium *Cyanothece* sp. ATCC 51142, *Deinococcus radiodurans* (Deinococcus-
183 Thermus) and fusobacterium *Ilyobacter polytropus*. This points to an incomplete
184 stabilization of the SERs into the genome that may reflect a recent, ongoing, event of
185 integration and/or differing selective pressures at play depending on the bacterial
186 lineages. At odds with these observations, some SERs group unambiguously with
187 chromosomes. The SERs in α-Proteobacteria *Asticcacaulis excentricus* and *Paracoccus*
188 *denitrificans*, Bacteroidetes *Prevotella intermedia* and *P. melaninogenica*,
189 acidobacterium *Chloracidobacterium thermophilum*, and cyanobacterium
190 *Anabaena* sp. 90 bear higher levels of interconnection to chromosomes than to plasmids
191 or other SERs. Indeed, the SERs in *Prevotella* spp. are hardly linked to plasmids, and
192 the few plasmid-like IS proteins that *C. thermophilum* SER codes (mostly Rep, Helicase
193 and PSK), albeit found in plasmids occurring in other phyla, are observed in none of the

194 Acidobacteria plasmids. An extreme situation is met in *Leptospira* spp. (Spirochaetae)
195 whose SERs are each linked to only three or four (out of a total of six) chromosome-like
196 IS clusters, always including ParA and ParB. Interestingly, the ParA cluster appears to
197 be specific to Spirochaetae chromosomes with the notable exception of one plasmid
198 found in Leptospiraceae *Turneriella parva*.

199 **IS-based relationships of the replicons**

200 We submitted the bipartite graph of the whole dataset to a community structure
201 detection algorithm (INFOMAP) that performs a random walk along the edges
202 connecting the graph vertices. We expected the replicon communities to be trapped in
203 high-density regions of the graph. We also performed a dimension reduction by
204 Principal Component Analysis followed by a hierarchical clustering procedure
205 (WARD). The clustering solutions (Supplementary tables 2 and 3) were meaningful
206 (high values reached by the stability criterion scores), and biologically relevant (efficient
207 separation of the chromosomes from the plasmids; high *homogeneity* values) using
208 either method (Table 3). In another experiment, we considered each genus as a unique
209 sample and averaged the variables over the replicons of the different species for each
210 replicon type. The aim was to control for the disparity in taxon representation of the
211 replicons. This dataset produced overall similar albeit slightly less stable clusters (lower
212 *homogeneity* values). Taxonomically homogeneous clusters of chromosomes were best
213 retrieved using the coupling of dimension reduction and hierarchical clustering with a
214 large enough number of clusters (*homogeneity* scores up to 0.93). In turn, the
215 community detection algorithm was more efficient in recovering the underlying
216 taxonomy of replicons (higher value of *completeness*), and was sole able to identify
217 small and scattered plasmid clusters (Supplementary tables 2 and 3).

218 **Table 3. Evaluation of the replicon IS-based clusterings**

| | | INDEX[a] | USING IS PROTEIN SEQUENCES | | | | USING IS FUNCTIONS | |
|---|---|---|---|---|---|---|---|---|
| | CLUSTERING | | INFOMAP | | PCA+ WARD[b] | | PCA+ WARD[b] | |
| PROCEDURE | Dataset[c] | | $V^R$ | $\bar{V}^R_{genus}$ | $V^R$ | $\bar{V}^R_{genus}$ | $V^R_f$ | $\bar{V}^R_{f,genus}$ |
| | Parameters | | 500 iterations | | $\begin{cases} k = 200 \\ pc = 30 \end{cases}$ | $\begin{cases} k = 200 \\ pc = 30 \end{cases}$ | $\begin{cases} k = 50 \\ pc = 4 \end{cases}$ | $\begin{cases} k = 20 \\ pc = 4 \end{cases}$ |
| | Number of clusters | | 223 | 77 | 175 | 75 | 49 | 19 |
| | PCA explained variance | | | | 57% | 58% | 87% | 85% |
| | Stability criterion ($\Delta^{Kl}$)[d] | | 0.82 | 0.76 | 0.85 | 0.74 | 0.80 | 0.71 |
| EVALUATED SEPARATION | Chromosomes *vs.* Plasmids | *homogeneity* | 0.82 | 0.63 | 0.93 | 0.83 | 0.85 | 0.68 |
| | | *completeness* | 0.15 | 0.15 | 0.25 | 0.20 | 0.30 | 0.23 |
| | | *V-measure* | 0.25 | 0.24 | 0.43 | 0.32 | 0.44 | 0.35 |
| | Chromosomes *per* host phylum | *homogeneity* | 0.93 | 0.69 | 0.93 | 0.80 | 0.50 | 0.44 |
| | | *completeness* | 0.60 | 0.61 | 0.35 | 0.40 | 0.27 | 0.33 |
| | | *V-measure* | 0.73 | 0.65 | 0.51 | 0.53 | 0.35 | 0.38 |
| | Chromosomes *per* host class | *homogeneity* | 0.85 | 0.64 | 0.93 | 0.80 | 0.47 | 0.37 |
| | | *completeness* | 0.80 | 0.82 | 0.16 | 0.58 | 0.36 | 0.41 |
| | | *V-measure* | 0.82 | 0.72 | 0.66 | 0.67 | 0.41 | 0.39 |
| | Plasmids *per* host phylum | *homogeneity* | 0.88 | 0.78 | 0.06 | 0.01 | 0.02 | 0.02 |
| | | *completeness* | 0.33 | 0.35 | 0.16 | 0.14 | 0.10 | 0.30 |
| | | *V-measure* | 0.48 | 0.48 | 0.08 | 0.02 | 0.03 | 0.03 |
| | Plasmids *per* host class | *homogeneity* | 0.84 | 0.74 | 0.07 | 0.02 | 0.03 | 0.02 |
| | | *completeness* | 0.43 | 0.51 | 0.28 | 0.36 | 0.25 | 0.28 |
| | | *V-measure* | 0.57 | 0.60 | 0.12 | 0.03 | 0.05 | 0.03 |

219 [a] *V-measure* according to Rosenberg and Hirschberg (2007)

220 [b] *k*: number of input clusters; *pc*: principal components used in WARD

221 [c] $V^R_f$: Ensemble of all IS function-based replicon vectors ($v^r_f$); $\bar{V}^R_{f,genus}$: Ensemble of IS function-based genus-

222 normalized replicon vectors ($v^r_{f,genus}$)

223 [d] Stability criterion according to Hennig (2007)

224 The plasmid clusters obtained using PCA+WARD lacked taxonomical patterning and,

225 although highly stable, only reflected the small Euclidian distances existing among the

226 plasmid replicons (*e.g.*, one cluster of 2656 plasmids had a stability score of 0.975). The

227 clusters obtained with INFOMAP mirrored the taxonomical structure of the data,
228 suggesting that the taxonomic signal, expected to be associated to the chromosomes, is
229 preserved among the IS protein families functionally specifying the plasmids. The
230 presence of a majority of the SERs amongst the chromosome clusters generated by
231 INFOMAP confirmed the affinities between these two genomic elements and the clear
232 individuation of the SERs from the plasmids. However, the larger number of
233 chromosomal ISs often caused the PCA+WARD approach to place SERs into plasmid
234 clusters. The SERs in *Butyrivibrio*, *Deinococcus*, *Leptospira* and *Rhodobacter* spp.
235 grouped consistently with plasmids while the SERs in *Vibrionaceae* and *Brucellaceae*
236 formed specific clusters (Table 4). Burkholderiales and *Agrobacterium* SERs, whose
237 homogenous clusters tended to be unstable, exhibited a higher affinity to plasmids
238 overall. The SERs of *Asticaccaulis*, *Paracoccus* and *Prevotella* spp. associated stably
239 with chromosomes using the two clustering methods (Table 4a,b) and possess IS profiles
240 that set them apart from both the plasmids and the other SERs.

241 **Table 4. IS protein cluster-based unsupervised classification of SERs**

242 a. INFOMAP clustering solution

| Bacterial genus | $C$ [a] | CHR% | SER% | PLD% | $wBHI$ [b] | $\overline{\Delta^C}$ [c] | $\overline{\Delta^r}$ [d] |
|---|---|---|---|---|---|---|---|
| *Agrobacterium* | 3 | 38 | 35 | 27 | 0.90 | 0.47 | 0.61 |
| *Aliivibrio* | 1 | 0 | 100 | 0 | 1.00 | 0.95 | 1.00 |
| *Anabaena* | 1 | 98 | 1 | 1 | 1.00 | 0.90 | 1.00 |
| *Asticcacaulis* | 1 | 96 | 1 | 3 | 1.00 | 0.97 | 1.00 |
| *Brucella* | 1 | 0 | 95 | 5 | 1.00 | 0.87 | 1.00 |
| *Burkholderia* | 2 | 64 | 17 | 19 | 0.99 | 0.77 | 0.99 |
| *Butyrivibrio* | 1 | 0 | 50 | 50 | 1.00 | 0.83 | 1.00 |
| *Chloracidobacterium* | 1 | 91 | <1 | 9 | 0.82 | 0.86 | 0.00 |
| *Cupriavidus* | 1 | 73 | 18 | 9 | 0.99 | 0.72 | 1.00 |
| *Cyanothece* | 1 | 0 | 6 | 94 | 0.89 | 0.61 | 0.33 |
| *Deinococcus* | 1 | 0 | 4 | 96 | 0.71 | 0.61 | 1.00 |
| *Ilyobacter* | 1 | 91 | <1 | 9 | 0.82 | 0.86 | 0.25 |
| *Leptospira* | 1 | 0 | 88 | 12 | 1.00 | 1.00 | 1.00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Nocardiopsis* | 1 | 91 | <1 | 9 | 0.97 | 0.97 | 1.00 |
| *Ochrobactrum* | 1 | 0 | 95 | 5 | 1.00 | 0.87 | *n.a.n.*[e] |
| *Paracoccus* | 1 | 96 | 1 | 3 | 1.00 | 0.97 | 1.00 |
| *Photobacterium* | 1 | 96 | 1 | 3 | 0.99 | 0.79 | 1.00 |
| *Prevotella* | 1 | 96 | 2 | 2 | 0.95 | 0.92 | 1.00 |
| *Pseudoalteromonas* | 1 | 96 | 1 | 3 | 0.99 | 0.79 | 0.56 |
| *Ralstonia* | 1 | 73 | 18 | 9 | 0.99 | 0.72 | 1.00 |
| *Rhodobacter* | 1 | 0 | 40 | 60 | 1.00 | 0.71 | 1.00 |
| *Ensifer* (*Sinorhizobium*) | 2 | 0 | 2 | 98 | 0.96 | 0.65 | 0.67 |
| *Sphaerobacter* | 1 | 0 | 50 | 50 | 1.00 | 1.00 | 1.00 |
| *Sphingobium* | 2 | 77 | 1 | 22 | 0.95 | 0.90 | 0.50 |
| *Thermobaculum* | 1 | 91 | <1 | 9 | 0.82 | 0.86 | 1.00 |
| *Variovorax* | 1 | 73 | 18 | 9 | 0.99 | 0.72 | 0.90 |
| *Vibrio* | 1 | 0 | 100 | 0 | 1.00 | 0.95 | 0.89 |

[a] number of clusters containing SERs of a given bacterial genus
[b] weighted biological homogeneity index value for the phylum of the replicons in the clusters
[c] mean value of the cluster stability estimator, weighted by the cluster sizes
[d] mean value of the SER stability estimator for a given bacterial genus
[e] "*n.a.n.*", standing for "not a number", indicates that the replicon appeared in none of the bootstrap replications performed in the clustering procedure

b. PCA+WARD clustering solution

| Bacterial genus | $C$[a] | CHR% | SER% | PLD% | *wBHI*[b] | $\overline{\Delta^C}$ [c] | $\overline{\Delta^r}$ [d] |
|---|---|---|---|---|---|---|---|
| *Agrobacterium* | 2 | 0 | 29 | 71 | 0.94 | 0.76 | 1.00 |
| *Aliivibrio* | 2 | 0 | 56 | 44 | 1.00 | 0.60 | 0.33 |
| *Anabaena* | 1 | 98 | 2 | 0 | 0.97 | 0.84 | 0.00 |
| *Asticcacaulis* | 1 | 88 | 8 | 4 | 1.00 | 0.88 | 1.00 |
| *Brucella* | 2 | 0 | 33 | 67 | 0.96 | 0.53 | 0.97 |
| *Burkholderia* | 7 | 0 | 79 | 21 | 0.97 | 0.69 | 0.84 |
| *Butyrivibrio* | 1 | <1 | 1 | 99 | 0.27 | 0.98 | 1.00 |
| *Chloracidobacterium* | 1 | <1 | 1 | 99 | 0.27 | 0.98 | 1.00 |
| *Cupriavidus* | 2 | 0 | 92 | 8 | 1.00 | 0.69 | 0.92 |
| *Cyanothece* | 1 | <1 | 1 | 99 | 0.27 | 0.98 | 1.00 |
| *Deinococcus* | 1 | <1 | 1 | 99 | 0.27 | 0.98 | 1.00 |
| *Ilyobacter* | 1 | <1 | 1 | 99 | 0.27 | 0.98 | 1.00 |
| *Leptospira* | 1 | <1 | 1 | 99 | 0.27 | 0.98 | 1.00 |
| *Nocardiopsis* | 1 | 0 | 2 | 98 | 0.58 | 0.40 | 1.00 |
| *Ochrobactrum* | 1 | 0 | 100 | 0 | 1.00 | 1.00 | 1.00 |
| *Paracoccus* | 1 | 88 | 8 | 4 | 1.00 | 0.88 | 1.00 |
| *Photobacterium* | 1 | 0 | 100 | 0 | 1.00 | 0.55 | 1.00 |
| *Prevotella* | 2 | 95 | 5 | 0 | 1.00 | 0.73 | 0.50 |
| *Pseudoalteromonas* | 2 | <1 | 1 | 99 | 0.28 | 0.82 | 0.83 |
| *Ralstonia* | 1 | 0 | 68 | 32 | 1.00 | 0.81 | 0.83 |
| *Rhodobacter* | 2 | 0 | 6 | 94 | 0.65 | 0.43 | 0.58 |
| *Ensifer* (*Sinorhizobium*) | 2 | 0 | 21 | 79 | 0.94 | 0.46 | 0.25 |
| *Sphaerobacter* | 1 | 0 | 20 | 80 | 0.93 | 0.52 | 0.50 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Sphingobium* | 1 | 0 | 39 | 61 | 1.00 | 0.66 | 0.83 |
| *Thermobaculum* | 1 | <1 | 1 | 99 | 0.27 | 0.98 | 1.00 |
| *Variovorax* | 1 | 0 | 67 | 33 | 1.00 | 0.48 | 0.00 |
| *Vibrio* | 2 | 0 | 56 | 44 | 1.00 | 0.60 | 0.79 |

250 [a] number of clusters containing SERs of a given bacterial genus
251 [b] weighted biological homogeneity index value for the phylum of the replicons in the clusters
252 [c] mean value of the cluster stability estimator, weighted by the cluster sizes
253 [d] mean value of the SER stability estimator for a given bacterial genus

254 We reached similar conclusions when performing a PCA+WARD clustering using the
255 117 functional annotations of the IS protein clusters rather than the IS clusters
256 themselves (Tables 3 and 5; Supplementary table 4).

257 **Table 5. Function-based unsupervised classification of SERs using PCA+WARD**

| Bacterial genus | $C^{a}$ | CHR% | SER% | PLD% | $wBHI^{b}$ | $\overline{\Delta^{C}}^{c}$ | $\overline{\Delta^{r}}^{d}$ |
|---|---|---|---|---|---|---|---|
| *Agrobacterium* | 3 | 64 | 21 | 15 | 0.86 | 0.60 | 0.81 |
| *Aliivibrio* | 1 | 0 | 70 | 30 | 1.00 | 0.70 | 0.67 |
| *Anabaena* | 1 | 77 | 4 | 19 | 0.21 | 0.64 | 1.00 |
| *Asticcacaulis* | 1 | 99 | 1 | 0 | 0.51 | 0.60 | 1.00 |
| *Brucella* | 1 | 43 | 32 | 25 | 0.75 | 0.80 | 1.00 |
| *Burkholderia* | 6 | 31 | 42 | 27 | 0.92 | 0.68 | 0.81 |
| *Butyrivibrio* | 1 | 77 | 4 | 19 | 0.21 | 0.64 | 1.00 |
| *Chloracidobacterium* | 1 | 1 | <1 | 99 | 0.29 | 0.98 | 0.00 |
| *Cupriavidus* | 1 | 5 | 95 | 0 | 1.00 | 0.66 | 0.66 |
| *Cyanothece* | 1 | 1 | <1 | 99 | 0.29 | 0.98 | 1.00 |
| *Deinococcus* | 1 | 1 | <1 | 99 | 0.29 | 0.98 | 1.00 |
| *Ilyobacter* | 1 | 77 | 4 | 19 | 0.21 | 0.64 | 1.00 |
| *Leptospira* | 1 | 1 | <1 | 99 | 0.29 | 0.98 | 1.00 |
| *Nocardiopsis* | 1 | 77 | 4 | 19 | 0.21 | 0.64 | 1.00 |
| *Ochrobactrum* | 1 | 90 | 8 | 2 | 1.00 | 0.40 | 0.73 |
| *Paracoccus* | 1 | 92 | 5 | 3 | 0.89 | 0.32 | 0.53 |
| *Photobacterium* | 1 | 0 | 70 | 30 | 1.00 | 0.70 | 0.36 |
| *Prevotella* | 2 | 86 | 3 | 11 | 0.34 | 0.62 | 0.50 |
| *Pseudoalteromonas* | 1 | 77 | 4 | 19 | 0.21 | 0.64 | 0.29 |
| *Ralstonia* | 1 | 0 | 70 | 30 | 1.00 | 0.70 | 0.22 |
| *Rhodobacter* | 2 | 27 | 32 | 41 | 0.84 | 0.73 | 0.83 |
| *Ensifer* (*Sinorhizobium*) | 2 | 25 | 48 | 27 | 0.86 | 0.76 | 0.63 |
| *Sphaerobacter* | 1 | 100 | <1 | 0 | 0.35 | 0.60 | 1.00 |
| *Sphingobium* | 2 | 62 | 17 | 21 | 0.34 | 0.64 | 0.86 |
| *Thermobaculum* | 1 | 77 | 4 | 19 | 0.21 | 0.64 | 1.00 |
| *Variovorax* | 1 | 0 | 70 | 30 | 1.00 | 0.70 | 1.00 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| *Vibrio* | 4 | 31 | 37 | 32 | 0.97 | 0.57 | 0.92 |

258 [a] number of clusters containing SERs of a given bacterial genus
259 [b] weighted biological homogeneity index value for the phylum of the replicons in the clusters
260 [c] mean value of the cluster stability estimator, weighted by the sizes of the clusters
261 [d] mean value of the SER stability estimator for a given bacterial genus

262 Remarkably, in this latter analysis, the chromosomes in the multipartite genomes of
263 *Prevotella intermedia* and *P. melaninogenica* were more similar to plasmids than to
264 other groups of chromosomes and to single chromosomes in other *Prevotella* species
265 (*P. denticola* and *P. ruminicola*).

266 **SER-specifying IS functions**

267 Next, we searched which of the IS functions are specific to the SERs. We performed
268 several logistic regression analyses to identify over- or under-represented ISs and to
269 assess their respective relevance to each class of replicons. Because of their
270 comparatively small number, all SERs were assembled into a single group despite their
271 disparity. A hundred and one IS functionalities (96% of KEGG-annotated chromosome-
272 like functions and 72% of ACLAME-annotated plasmid-like functions) were
273 significantly enriched in one replicon category over the other (Table 6). The large
274 majority of the IS functions differentiates the chromosomes from the plasmids. The
275 latter are only determined by ISs corresponding to ACLAME annotations Rep, Rop and
276 TrfA, involved in initiation of plasmid replication, and ParA and ParB, dedicated to
277 plasmid partition. Some KEGG-annotated functions, *e.g.,* DnaA, DnaB or FtsZ, appear
278 to be more highly specific to chromosomes (higher *OR* values) than others such as
279 DnaC, FtsE or H-NS (lower *OR* values). Strikingly, very few functions distinguish
280 significantly the chromosomes from the SERs, by contrast with plasmids.

## Table 6. IS usage comparison between replicon categories

Between classes of replicons logistic regressions for each IS function. Model significance: $0 < P_value < 0.01$: significant; $0.01 < P_value < 0.05$: poorly significant; $0.05 < P_value$: non significant (not shown). Odd-ratio (*OR*) favouring the first class: $10^0 \leq OR$, or the second class: $OR < 10^{\bar{0}}$. IS functions biased to the same order of magnitude in chromosomes and SERs when compared to plasmids are highlighted (blue).

| | IS function | Chromosomes *vs.* Plasmids | | Chromosomes *vs.* SERs | | SERs *vs.* Plasmids | |
|---|---|---|---|---|---|---|---|
| | | *P_value* | *OR* | *P_value* | *OR* | *P_value* | *OR* |
| REPLICATION | CbpA | $8.20 \times 10^{-27}$ | 2,608.4 | $9.90 \times 10^{-13}$ | 22.8 | $5.60 \times 10^{-07}$ | 36.1 |
| | Dam | $6.90 \times 10^{-16}$ | 16.7 | $3.60 \times 10^{-02}$ | 2.0 | $2.40 \times 10^{-02}$ | 4.3 |
| | DiaA | $1.50 \times 10^{-15}$ | 81.9 | $1.20 \times 10^{-03}$ | 38.4 | | |
| | DnaA | $3.00 \times 10^{-44}$ | 2,118.9 | $1.10 \times 10^{-19}$ | 239.6 | $3.50 \times 10^{-03}$ | 8.3 |
| | DnaB | $1.10 \times 10^{-43}$ | 1,992.9 | $5.10 \times 10^{-19}$ | 429.4 | $8.20 \times 10^{-03}$ | 3.7 |
| | DnaB2 | $6.70 \times 10^{-03}$ | 12.6 | | | | |
| | DnaC | $6.00 \times 10^{-12}$ | 2.6 | | | $4.60 \times 10^{-02}$ | 1.5 |
| | DnaG | $2.10 \times 10^{-50}$ | 1,861.5 | $1.90 \times 10^{-21}$ | 205.3 | $2.50 \times 10^{-03}$ | 4.5 |
| | DnaI | $5.20 \times 10^{-03}$ | 18.0 | | | | |
| | Dps | $9.10 \times 10^{-21}$ | 65.3 | $3.50 \times 10^{-05}$ | 8.4 | $8.70 \times 10^{-03}$ | 6.7 |
| | Fis | $5.80 \times 10^{-07}$ | 180.9 | $3.30 \times 10^{-03}$ | 7.9 | $1.40 \times 10^{-02}$ | 25.1 |
| | Hda | $7.30 \times 10^{-07}$ | 149.1 | $5.30 \times 10^{-03}$ | 7.9 | $1.90 \times 10^{-02}$ | 18.0 |
| | Hfq | $1.40 \times 10^{-12}$ | 121.7 | $3.00 \times 10^{-04}$ | 6.9 | $8.10 \times 10^{-04}$ | 19.3 |
| | H-NS | $1.10 \times 10^{-05}$ | 2.8 | | | $3.80 \times 10^{-04}$ | 2.8 |
| | HupA | $2.70 \times 10^{-04}$ | 15.1 | | | | |
| | HupB | $1.20 \times 10^{-53}$ | 97.6 | $2.30 \times 10^{-08}$ | 6.7 | $2.40 \times 10^{-07}$ | 11.6 |
| | IciA | $7.10 \times 10^{-20}$ | 3.2 | | | $4.50 \times 10^{-07}$ | 1.8 |
| | IhfA, HimA | $1.70 \times 10^{-12}$ | 63.8 | $1.40 \times 10^{-03}$ | 10.5 | $4.90 \times 10^{-02}$ | 6.9 |
| | IhfB, HimD | $1.20 \times 10^{-14}$ | 68.4 | $4.90 \times 10^{-04}$ | 8.4 | $8.40 \times 10^{-03}$ | 9.9 |
| | Lrp | $1.60 \times 10^{-19}$ | 8.4 | | | $5.40 \times 10^{-11}$ | 8.1 |
| | Rob | $6.30 \times 10^{-19}$ | 5.3 | | | $3.40 \times 10^{-08}$ | 4.2 |
| | SeqA | $1.60 \times 10^{-03}$ | 25.9 | | | | |
| | ssb | $5.90 \times 10^{-41}$ | 298.3 | $5.00 \times 10^{-18}$ | 160.6 | | |
| PARTITION | Fic | $3.10 \times 10^{-09}$ | 10.3 | | | $8.60 \times 10^{-03}$ | 7.2 |
| | GidA, MnmG, | $5.20 \times 10^{-13}$ | 1,477.2 | $2.90 \times 10^{-08}$ | 110.6 | $4.30 \times 10^{-02}$ | 18.2 |
| | GidB, RsmG | $6.70 \times 10^{-17}$ | 6,059.9 | $2.20 \times 10^{-15}$ | 252.5 | $9.00 \times 10^{-03}$ | 32.2 |
| | MreB | $1.30 \times 10^{-21}$ | 1,598.2 | $3.90 \times 10^{-12}$ | 40.1 | $1.40 \times 10^{-05}$ | 24.1 |
| | MreC | $2.90 \times 10^{-11}$ | 1,311.2 | $1.30 \times 10^{-08}$ | 46.3 | $8.90 \times 10^{-03}$ | 32.8 |
| | MreD | $1.80 \times 10^{-08}$ | 459.2 | $6.80 \times 10^{-05}$ | 19.8 | $1.90 \times 10^{-02}$ | 18.2 |
| | Mrp | $6.60 \times 10^{-17}$ | 2,599.3 | $1.30 \times 10^{-14}$ | 35.0 | $2.50 \times 10^{-05}$ | 86.2 |
| | MukB | $2.30 \times 10^{-03}$ | 27.4 | | | $1.90 \times 10^{-02}$ | 18.2 |
| | MukE | $3.10 \times 10^{-03}$ | 21.0 | | | $1.90 \times 10^{-02}$ | 18.2 |
| | MukF | $3.70 \times 10^{-03}$ | 19.6 | | | $1.90 \times 10^{-02}$ | 18.2 |
| | ParA, Soj | $2.70 \times 10^{-38}$ | 9.9 | $9.00 \times 10^{-06}$ | 2.6 | $8.40 \times 10^{-06}$ | 3.8 |
| | ParB, Spo0J | $2.50 \times 10^{-44}$ | 13.7 | $3.00 \times 10^{-03}$ | 2.1 | $2.30 \times 10^{-06}$ | 4.1 |
| | ParC | $3.00 \times 10^{-27}$ | 4,149.3 | $3.00 \times 10^{-16}$ | 134.0 | $4.60 \times 10^{-04}$ | 12.3 |
| | ParE | $7.30 \times 10^{-26}$ | 5,842.4 | $5.70 \times 10^{-15}$ | 350.1 | $2.40 \times 10^{-04}$ | 15.8 |
| | RodA, MrdB | $2.80 \times 10^{-12}$ | 1,233.1 | $9.70 \times 10^{-10}$ | 33.0 | $2.60 \times 10^{-03}$ | 55.3 |
| | TrmFO, Gid | $1.50 \times 10^{-06}$ | 182.5 | $4.40 \times 10^{-03}$ | 8.3 | $1.90 \times 10^{-02}$ | 18.0 |
| | XerC | $1.70 \times 10^{-43}$ | 55.0 | $3.10 \times 10^{-08}$ | 8.8 | $1.80 \times 10^{-04}$ | 6.7 |
| | XerD | $1.30 \times 10^{-38}$ | 26.6 | $4.10 \times 10^{-08}$ | 3.4 | $2.50 \times 10^{-06}$ | 6.2 |
| SEGREGATION | ScpA | $1.40 \times 10^{-11}$ | 789.4 | $5.70 \times 10^{-07}$ | 42.9 | $2.10 \times 10^{-02}$ | 16.6 |
| | ScpB | $7.50 \times 10^{-32}$ | 102.5 | $1.80 \times 10^{-07}$ | 25.8 | | |
| | SepF | $1.80 \times 10^{-07}$ | 68.8 | $1.40 \times 10^{-02}$ | 12.3 | | |
| | SlmA, Ttk | $3.80 \times 10^{-09}$ | 52.3 | $1.20 \times 10^{-02}$ | 4.6 | $1.50 \times 10^{-02}$ | 7.5 |
| | Smc | $1.60 \times 10^{-08}$ | 3,090.5 | $1.40 \times 10^{-05}$ | 131.9 | | |
| | SulA | $3.30 \times 10^{-06}$ | 17.5 | | | $1.50 \times 10^{-02}$ | 10.7 |
| | AcrA | $6.60 \times 10^{-19}$ | 2.8 | $1.70 \times 10^{-02}$ | 1.1 | $5.30 \times 10^{-10}$ | 2.7 |
| CELL DIVISION | AmiA, AmiB, | $6.40 \times 10^{-36}$ | 46.4 | $2.90 \times 10^{-10}$ | 8.9 | $4.60 \times 10^{-03}$ | 3.0 |
| | DivIC, DivA | $4.90 \times 10^{-05}$ | 90.5 | $4.70 \times 10^{-02}$ | 8.1 | | |
| | DivIVA | $4.10 \times 10^{-06}$ | 128.0 | $1.10 \times 10^{-02}$ | 13.4 | | |
| | EzrA | $1.00 \times 10^{-02}$ | 13.7 | | | | |
| | FtsA | $9.50 \times 10^{-12}$ | 742.7 | $2.20 \times 10^{-08}$ | 24.6 | $2.50 \times 10^{-03}$ | 41.7 |
| | FtsB | $1.10 \times 10^{-06}$ | 167.2 | $5.40 \times 10^{-03}$ | 16.1 | | |
| | FtsE | $4.20 \times 10^{-24}$ | 2.3 | $1.30 \times 10^{-06}$ | 1.1 | $4.00 \times 10^{-11}$ | 1.9 |
| | FtsI | $9.80 \times 10^{-09}$ | 47.0 | $7.00 \times 10^{-16}$ | 3.9 | $2.20 \times 10^{-07}$ | 76.7 |
| | FtsK, SpoIIIE | $2.80 \times 10^{-37}$ | 76.9 | $2.70 \times 10^{-08}$ | 15.8 | $1.40 \times 10^{-02}$ | 4.2 |
| | FtsL | $1.20 \times 10^{-05}$ | 91.5 | $2.70 \times 10^{-02}$ | 9.8 | | |
| | FtsN | $1.60 \times 10^{-04}$ | 53.0 | | | | |
| | FtsQ | $1.70 \times 10^{-15}$ | 2,135.0 | $1.30 \times 10^{-11}$ | 99.3 | $9.00 \times 10^{-03}$ | 28.8 |
| | FtsW, SpoVE | $5.70 \times 10^{-16}$ | 4,266.4 | $4.40 \times 10^{-16}$ | 87.7 | $8.20 \times 10^{-04}$ | 55.0 |
| | FtsX | $9.30 \times 10^{-12}$ | 972.9 | $1.30 \times 10^{-08}$ | 13.8 | $4.80 \times 10^{-04}$ | 146.2 |
| | FtsZ | $3.10 \times 10^{-31}$ | 2,747.0 | $1.20 \times 10^{-19}$ | 101.6 | $9.70 \times 10^{-04}$ | 16.5 |
| | MinC | $4.40 \times 10^{-09}$ | 172.3 | $1.20 \times 10^{-02}$ | 3.0 | $5.80 \times 10^{-05}$ | 76.8 |
| | MinD | $3.10 \times 10^{-19}$ | 42.8 | $1.60 \times 10^{-04}$ | 2.3 | $5.40 \times 10^{-11}$ | 81.5 |
| | MinE | $9.00 \times 10^{-09}$ | 152.9 | $3.10 \times 10^{-02}$ | 2.6 | $5.90 \times 10^{-05}$ | 75.2 |
| | ZapA | $8.20 \times 10^{-09}$ | 602.8 | $7.40 \times 10^{-06}$ | 17.3 | $7.30 \times 10^{-03}$ | 56.1 |
| | ZipA | $7.90 \times 10^{-05}$ | 66.0 | | | | |

KEGG entries (chromosome-like)

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| ACLAME Families (plasmid-like) | | | | | | | |
| | CdsD | | | $4.40 \times 10^{-02}$ | 0.1 | | |
| REPLICATION | DNA helicase | $5.80 \times 10^{-21}$ | 33.6 | $2.70 \times 10^{-04}$ | 4.1 | $1.30 \times 10^{-04}$ | 9.8 |
| | Helicase-1 | $1.60 \times 10^{-27}$ | 71.1 | $1.90 \times 10^{-13}$ | 20.0 | $1.10 \times 10^{-04}$ | 4.6 |
| | DNA repair | $2.20 \times 10^{-04}$ | 34.0 | | | $5.70 \times 10^{-04}$ | 43.6 |
| | primase, LtrC | $3.10 \times 10^{-02}$ | 1.8 | | | | |
| | RepA | $5.90 \times 10^{-03}$ | 0.7 | | | | |
| | RepAEB | $1.70 \times 10^{-16}$ | 0.0 | $1.90 \times 10^{-04}$ | 0.1 | | |
| | RepC | | | | | $9.60 \times 10^{-03}$ | 2.7 |
| | RepCJE | | | $4.40 \times 10^{-02}$ | 0.1 | | |
| | RepRSE | $1.30 \times 10^{-02}$ | 0.0 | $4.90 \times 10^{-02}$ | 0.1 | | |
| | RNA polymerase | $3.20 \times 10^{-02}$ | 6.3 | | | | |
| | Rop | $3.20 \times 10^{-02}$ | 0.0 | $4.40 \times 10^{-02}$ | 0.1 | | |
| | RuvB | $1.20 \times 10^{-08}$ | 433.0 | $5.70 \times 10^{-08}$ | 17.7 | $1.40 \times 10^{-05}$ | 37.8 |
| | TrfA | $1.40 \times 10^{-02}$ | 0.3 | | | | |
| PARTITION | ATPase, TyrK, | $2.20 \times 10^{-20}$ | 19.4 | | | $8.50 \times 10^{-07}$ | 9.3 |
| | CopG | | | $2.70 \times 10^{-02}$ | 0.2 | $4.60 \times 10^{-03}$ | 23.1 |
| | DNA-binding protein | | | $4.40 \times 10^{-02}$ | 0.1 | | |
| | FtsK, SpoIIIE | $1.90 \times 10^{-07}$ | 6.0 | | | $9.90 \times 10^{-05}$ | 9.8 |
| | ParA, ParM | $1.50 \times 10^{-10}$ | 0.4 | $4.00 \times 10^{-06}$ | 0.3 | | |
| | ParB | $5.70 \times 10^{-12}$ | 0.1 | $1.40 \times 10^{-05}$ | 0.2 | | |
| | serine recombinase | $2.50 \times 10^{-06}$ | 1.4 | $1.50 \times 10^{-03}$ | 2.9 | $1.80 \times 10^{-02}$ | 0.4 |
| | tyrosine recombinase | $3.40 \times 10^{-04}$ | 3.3 | | | $7.40 \times 10^{-04}$ | 8.7 |
| | Xer-like Tyrosine | $7.60 \times 10^{-11}$ | 2.0 | | | $6.30 \times 10^{-03}$ | 1.6 |
| MAINTENANCE | Ccd (PSK) | $4.60 \times 10^{-02}$ | 3.9 | | | | |
| | HicAB (PSK) | $4.30 \times 10^{-05}$ | 25.2 | | | $4.80 \times 10^{-03}$ | 15.1 |
| | HigBA (PSK) | $3.30 \times 10^{-15}$ | 3.4 | $2.40 \times 10^{-02}$ | 1.5 | $1.20 \times 10^{-03}$ | 2.5 |
| | MazEF (PSK) | $1.20 \times 10^{-11}$ | 5.2 | $2.90 \times 10^{-02}$ | 2.6 | | |
| | ParC (PSK) | | | $4.40 \times 10^{-02}$ | 0.1 | | |
| | ParDE (PSK) | $5.50 \times 10^{-08}$ | 2.3 | | | $7.80 \times 10^{-05}$ | 3.4 |
| | PhD, Doc (PSK) | $3.20 \times 10^{-07}$ | 11.9 | | | $2.90 \times 10^{-03}$ | 8.8 |
| | plasmid maintenance | | | $4.40 \times 10^{-02}$ | 0.1 | | |
| | RelBE (PSK) | $2.70 \times 10^{-08}$ | 3.5 | | | $6.10 \times 10^{-04}$ | 4.2 |
| | VapBC/Vag (PSK) | $1.20 \times 10^{-09}$ | 3.9 | | | $1.40 \times 10^{-05}$ | 5.8 |

286 Chromosome-signature ISs are also present on the SERs, and some of them are enriched
287 to the same order of magnitude in both classes but not in plasmids (highlighted in
288 Table 6). Among these latter, helicase loader DnaC participates to the replication
289 initiation of the chromosome (Chodavarapu et al., 2016) whilst Walker-type ATPase
290 ParA/Soj interacts with ParB/Spo0J in the *parABS* chromosomal partinioning system,
291 and is required for proper separation of sister origins and synchronous DNA replication
292 (Murray and Errington, 2008). The other ISs have a regulatory role, either locally or
293 globally. Nucleoid-associated proteins (NAPs; Dillon and Dorman, 2010) contribute to
294 the replication regulation: H-NS (histone-like nucleoid structuring protein), IciA
295 (chromosome initiator inhibitor, LysR family transcriptional regulator), MukBEF
296 (condensin), and Rob/ClpB (right arm of the replication origin binding protein/curved
297 DNA-binding protein B, AraC family transcriptional regulator) influence both the
298 conformation and the functions of chromosomal DNA, replication, recombination and

299 repair. The NAPs also have pleiotropic regulatory roles in global regulation of gene
300 transcription depending on cell growth conditions (H-NS, IciA, Lrp (leucine-responsive
301 regulatory protein, Lrp/AsnC family transcriptional regulator), and Rob/ClpB).
302 Similarly, the membrane fusion protein AcrA is a growth-dependent regulator, mostly
303 known for its role as a peripheral scaffold mediating the interaction between AcrB and
304 TolC in the AcrA-AcrB-TolC Resistance-Nodule-cell Division-type efflux pump that
305 extrudes from the cell compounds that are toxic or have a signaling role (Du et al.,
306 2018). It is central to the regulation of cell homeostasis and proper development (Anes
307 et al., 2015; Du et al., 2018; Webber et al., 2009) as well as biofilm formation (Alav et
308 al., 2018). Fic (cell filamentation protein) targets the DNA gyrase B (GyrB) to regulate
309 the cell division and cell morphology (Lu et al., 2018) whereas SulA inhibits FtsZ
310 assembly, hence causing incomplete cell division and filamentation (Chen et al, 2012).
311 FtsE is involved in the Z-ring assembly and the initiation of constriction, and in late
312 stage cell separation (Meier et al, 2017).

313 The main divergence between SERs and chromosomes lies in the distribution patterns of
314 the ACLAME-annotated ISs ($OR < 10^0$ in the chromosomes *vs.* SERs comparison). Their
315 higher abundance on the SERs suggests a stronger link of SERs to plasmids. This
316 pattern may also arise from the unbalanced taxon representation in our SER dataset due
317 to a single bacterial lineage. For example, the presence of RepC is likely to be specific to
318 Rhizobiales SERs (Pinto et al., 2012).

319 **Identification of candidate SERs**

320 Since the IS profiles constitute replicon-type signatures, we searched for new putative
321 SERs or chromosomes among the extra-chromosomal replicons. We used the IS

322 functions as features to perform supervised classification analyses with various training
323 sets (Table 7).

324 **Table 7. Performance of the ERT classification procedures**

| TRAINING SET [a] | $CV_{score}$ [b] | $\sigma_{CV_{score}}$ [c] | $OOB_{score}$ [d] | $\sigma_{OOB_{score}}$ [e] |
|---|---|---|---|---|
| $\{E_{SER}, E'_{plasmid}\}$ | 0.96 | - | 0.96 | - |
| $\{E_{SER}, E'_{plasmid}\}^{it}$ | 0.92 | 0.02 | 0.93 | 0.02 |
| $\{E_{chromosome}, E'_{plasmid}\}$ | 1.00 | - | 1.00 | - |
| $\{E_{SER}, E_{chromosome}\}^{it}$ | 0.98 | 0.00 | 0.98 | 0.01 |

325 [a] $E_{chromosome}$ and $E_{SER}$ are host genus-normalized sets of chromosomes or SERs, respectively (*cf.*
326 Table 13). $E'_{plasmid}$ is derived from the INFOMAP clustering solution, by discarding plasmids belonging to
327 clusters also harbouring SERs or chromosomes, and normalized according to host genus. "*it*" designates the
328 iterative procedure.
329 [b] Cross-validation score or mean of iteration cross-validation scores.
330 [c] Standard deviation of iteration cross-validation scores.
331 [d] Out-of bag estimate or mean of iteration Out-of bag estimates.
332 [e] Standard deviation of iteration Out-of bag estimates.

333 The coherence of the SER class (overall high values of the probability for a SER to be
334 assigned to its own class in Tables 7 and 8) confirmed that the ISs are robust genomic
335 markers for replicon characterization. The low SER probability scores presented by a
336 few SERs (Table 8) likely result from a low number of carried ISs (*e.g., Leptospira*), or
337 from the absence in the data of lineage-specific ISs (*e.g.*, SER idiosyncratic replication
338 initiator RtcB of Vibrionaceae).

339 **Table 8. SER probability to belong to the *SER* class**

| Genus | $\bar{P}_{SER}(SER)$ [a] |
|---|---|
| *Agrobacterium* | 0.90 |
| *Aliivibrio* | 0.87 |
| *Anabaena* | 0.94 |
| *Asticcacaulis* | 0.95 |

| | |
|---|---|
| *Brucella* | 0.92 |
| *Burkholderia* | 0.89 |
| *Butyrivibrio* | 0.83 |
| *Chloracidobacterium* | 0.88 |
| *Cupriavidus* | 0.94 |
| *Cyanothece* | 0.86 |
| *Deinococcus* | 0.78 |
| *Ensifer/Sinorhizobium* | 0.90 |
| *Ilyobacter* | 0.88 |
| *Leptospira* | 0.54 |
| *Nocardiopsis* | 0.90 |
| *Ochrobactrum* | 0.96 |
| *Paracoccus* | 0.96 |
| *Photobacterium* | 0.95 |
| *Prevotella* | 0.92 |
| *Pseudoalteromonas* | 0.91 |
| *Ralstonia* | 0.95 |
| *Rhodobacter* | 0.69 |
| *Sphaerobacter* | 0.88 |
| *Sphingobium* | 0.73 |
| *Thermobaculum* | 0.78 |
| *Variovorax* | 0.83 |
| *Vibrio* | 0.76 |

340 [a] SER probability, averaged per host genus, to belong to the *SER* class in the supervised classification
341 using $\{E_{SER}, E_{plasmid}\}^{it}$ as training set.

342 We detected a number of candidate SERs among the plasmids (Table 9a), most of which
343 are essential to the cell functioning and/or the fitness of the organism (*cf.* Box 1).
344 Whereas most belong to bacterial lineages known to harbour multipartite genomes,
345 novel taxa emerge as putative hosts to complex genomes (Rhodospirillales and, to a
346 lesser extent, Actinomycetales). In contrast, our analyses confirmed only one putative
347 SER (*Ruegeria* sp. TM1040) within the *Roseobacter* clade (Petersen et al., 2013).
348 Remarkably, we identified eight candidate chromosomes corresponding to two plasmids,
349 also identified as candidate SERs, that encode ISs hardly found in extra-chromosomal

350 elements (*e.g.*, DnaG, DnaB, ParC and ParE), and six SERs that part of, or all, our
351 analyses associate to standard chromosomes (Table 9b). Notably, *Prevotella intermedia*
352 SER (CP003503) shows a very high probability (> 0.98) to be a chromosome while its
353 annotated chromosome (CP003502), unique of its kind, falls within the plasmid class.
354 This approach can thus be extended to test the type of replicon for (re)annotation
355 purposes.

356 **Table 9. Identification of ERs among the extra-chromosomal replicons**

357 a. Candidate-SERs identified among plasmids

| REPLICON | PROBABILITY[a] |
|---|---|
| *Acaryochloris marina* MBIC11017 plasmid pREB1 [CYANOBACTERIA : Chroococcales] (CP000838) | 0.578 |
| *Acaryochloris marina* MBIC11017 plasmid pREB2 [CYANOBACTERIA : Chroococcales] (CP000839) | 0.582 |
| *Agrobacterium* sp. H13-3 plasmid pAspH13-3a [α-PROTEOBACTERIA : Rhizobiales] (CP0022) | 0.565 |
| *Arthrobacter chlorophenolicus* A6 plasmid pACHL01 [ACTINOBACTERIA : Actinomycetales] (CP001342) | 0.648 |
| *Azospirillum brasilense* Sp245 plasmid AZOBR_p1 [α-PROTEOBACTERIA : Rhodospirillales] (HE577328) | 0.878 |
| *Azospirillum brasilense* Sp245 plasmid AZOBR_p2 [α-PROTEOBACTERIA : Rhodospirillales] (HE577329) | 0.591 |
| *Azospirillum brasilense* Sp245 plasmid AZOBR_p3 [α-PROTEOBACTERIA : Rhodospirillales] (HE577330) | 0.603 |
| *Azospirillum lipoferum* 4B plasmid AZO_p1e [α-PROTEOBACTERIA : Rhodospirillales] (FQ311869) | 0.722 |
| *Azospirillum lipoferum* 4B plasmid AZO_p2 [α-PROTEOBACTERIA : Rhodospirillales] (FQ311870) | 0.609 |
| *Azospirillum lipoferum* 4B plasmid AZO_p4 [α-PROTEOBACTERIA : Rhodospirillales] (FQ311872) | 0.645 |
| *Azospirillum* sp. B510 plasmid pAB510a [α-PROTEOBACTERIA : Rhodospirillales] (AP010947) | 0.732 |
| *Azospirillum* sp. B510 plasmid pAB510c [α-PROTEOBACTERIA : Rhodospirillales] (AP010949) | 0.545 |
| *Azospirillum* sp. B510 plasmid pAB510d [α-PROTEOBACTERIA : Rhodospirillales] (AP010950) | 0.530 |
| *Burkholderia phenoliruptrix* BR3459a plasmid pSYMBR3459 [β-PROTEOBACTERIA : Burkholderiales] (CP003865) | 0.663 |
| *Burkholderia phymatum* STM815 plasmid pBPHY01 [β-PROTEOBACTERIA : Burkholderiales] (CP001045) | 0.733 |
| *Burkholderia* sp. YI23 plasmid byi_1p [β-PROTEOBACTERIA : Burkholderiales] (CP003090) | 0.846 |
| *Clostridium botulinum* A3 str. Loch Maree plasmid pCLK [FIRMICUTES : Clostridiales] (CP000963) | 0.531 |
| *Clostridium botulinum* Ba4 str. 657 plasmid pCLJ [FIRMICUTES : Clostridiales] (CP001081) | 0.531 |
| *Cupriavidus metallidurans* CH34 megaplasmid [β-PROTEOBACTERIA : Burkholderiales] (CP000353) | 0.883 |
| *Cupriavidus necator* N-1 plasmid BB1p [β-PROTEOBACTERIA : Burkholderiales] (CP002879) | 0.500 |
| *Cupriavidus pinatubonensis* JMP134 megaplasmid [β-PROTEOBACTERIA : Burkholderiales] (CP000092) | 0.513 |
| *Deinococcus geothermalis* DSM 11300 plasmid1 [DEINOCOCCUS-THERMUS : Deinococcales] (CP000358) | 0.622 |
| *Deinococcus gobiensis* I-0 plasmid P1 [DEINOCOCCUS-THERMUS : Deinococcales] (CP002192) | 0.812 |
| *Ensifer/Sinorhizobium fredii* HH103 plasmid pSfHH103e [α-PROTEOBACTERIA : Rhizobiales] (HE616899) | 0.915 |
| *Ensifer/Sinorhizobium fredii* NGR234 plasmid pNGR234b [α-PROTEOBACTERIA : Rhizobiales] (CP000874) | 0.894 |
| *Ensifer/Sinorhizobium medicae* WSM419 plasmid pSMED01 [α-PROTEOBACTERIA : Rhizobiales] (CP000739) | 0.942 |
| *Ensifer/Sinorhizobium medicae* WSM419 plasmid pSMED02 [α-PROTEOBACTERIA : Rhizobiales] (CP000740) | 0.836 |
| *Ensifer/Sinorhizobium meliloti* 1021 plasmid pSymA [α-PROTEOBACTERIA : Rhizobiales] (AE006469) | 0.818 |
| *Ensifer/Sinorhizobium meliloti* 1021 plasmid pSymB [α-PROTEOBACTERIA : Rhizobiales] (AL591985) | 0.949 |
| *Ensifer/Sinorhizobium meliloti* BL2C plasmid pSINMEB01 [α-PROTEOBACTERIA : Rhizobiales] (CP002741) | 0.800 |
| *Ensifer/Sinorhizobium meliloti* BL2C plasmid pSINMEB02 [α-PROTEOBACTERIA : Rhizobiales] (CP002742) | 0.961 |
| *Ensifer/Sinorhizobium meliloti* Rm41 plasmid pSYMA [α-PROTEOBACTERIA : Rhizobiales] (HE995407) | 0.922 |
| *Ensifer/Sinorhizobium meliloti* Rm41 plasmid pSYMB [α-PROTEOBACTERIA : Rhizobiales] (HE995408) | 0.960 |
| *Ensifer/Sinorhizobium meliloti* SM11 plasmid pSmeSM11c [α-PROTEOBACTERIA : Rhizobiales] (CP001831) | 0.877 |
| *Ensifer/Sinorhizobium meliloti* SM11 plasmid pSmeSM11d [α-PROTEOBACTERIA : Rhizobiales] (CP001832) | 0.947 |
| *Methylobacterium extorquens* AM1 megaplasmid [α-PROTEOBACTERIA : Rhizobiales] (CP001511) | 0.538 |
| *Novosphingobium* sp. PP1Y plasmid Mpl [α-PROTEOBACTERIA : Sphingomonadales] (FR856861) | 0.523 |
| *Pantoea* sp. At-9b plasmid pPAT9B01 [γ-PROTEOBACTERIA : Enterobacteriales] (CP002434) | 0.527 |

| | |
|---|---|
| *Paracoccus denitrificans* PD1222 plasmid1 [α-PROTEOBACTERIA : Rhodobacterales] (CP000491) | 0.769 |
| *Ralstonia solanacearum* GMI0 plasmid pGMI0MP [β-PROTEOBACTERIA : Burkholderiales] (AL646053) | 0.861 |
| *Ralstonia solanacearum* Po82 megaplasmid [β-PROTEOBACTERIA : Burkholderiales] (CP002820) | 0.865 |
| *Ralstonia solanacearum* PSI07 megaplasmid [β-PROTEOBACTERIA : Burkholderiales] (FP885891) | 0.827 |
| *Rhizobium etli* CFN 42 plasmid p42e [α-PROTEOBACTERIA : Rhizobiales] (CP000137) | 0.700 |
| *Rhizobium etli* CFN 42 plasmid p42f [α-PROTEOBACTERIA : Rhizobiales] (CP000138) | 0.555 |
| *Rhizobium etli* CIAT 652 plasmid pA [α-PROTEOBACTERIA : Rhizobiales] (CP0010) | 0.701 |
| *Rhizobium etli* CIAT 652 plasmid pC [α-PROTEOBACTERIA : Rhizobiales] (CP001077) | 0.792 |
| *Rhizobium leguminosarum* bv. trifolii WSM1325 plasmid pR132501 [α-PROTEOBACTERIA : Rhizobiales] (CP001623) | 0.711 |
| *Rhizobium leguminosarum* bv. trifolii WSM1325 plasmid pR132502 [α-PROTEOBACTERIA : Rhizobiales] (CP001624) | 0.741 |
| *Rhizobium leguminosarum* bv. *trifolii* WSM2304 plasmid pRLG201 [α-PROTEOBACTERIA : Rhizobiales] (CP001192) | 0.777 |
| *Rhizobium leguminosarum* bv. *trifolii* WSM2304 plasmid pRLG202 [α-PROTEOBACTERIA : Rhizobiales] (CP001193) | 0.630 |
| *Rhizobium leguminosarum* bv. *viciae* 3841 plasmid pRL11 [α-PROTEOBACTERIA : Rhizobiales] (AM236085) | 0.731 |
| *Rhizobium leguminosarum* bv. *viciae* 3841 plasmid pRL12 [α-PROTEOBACTERIA : Rhizobiales] (AM236086) | 0.718 |
| *Ruegeria* sp. TM1040 megaplasmid [α-PROTEOBACTERIA : Rhodobacterales] (CP000376) | 0.667 |
| *Streptomyces cattleya* NRRL 8057 plasmid pSCA [ACTINOBACTERIA : Actinomycetales] (FQ859184) | 0.727 |
| *Streptomyces cattleya* NRRL 8057 plasmid pSCATT [ACTINOBACTERIA : Actinomycetales] (CP003229) | 0.702 |
| *Streptomyces clavuligerus* ATCC 27064 plasmid pSCL4 [ACTINOBACTERIA : Actinomycetales] (CM000914) | 0.642 |
| *Streptomyces clavuligerus* ATCC 27064 plasmid pSCL4 [ACTINOBACTERIA : Actinomycetales] (CM001019) | 0.642 |
| *Thermus thermophilus* HB8 plasmid pTT27 [DEINOCOCCUS-THERMUS : Thermales] (AP008227) | 0.500 |
| *Thermus thermophilus* JL-18 plasmid pTTJL1801 [DEINOCOCCUS-THERMUS : Thermales] (CP0033) | 0.557 |
| *Tistrella mobilis* KA081020-065 plasmid pTM2 [α-PROTEOBACTERIA : Rhodospirillales] (CP003238) | 0.578 |
| *Tistrella mobilis* KA081020-065 plasmid pTM3 [α-PROTEOBACTERIA : Rhodospirillales] (CP003239) | 0.797 |

358 b. Candidate chromosomes identified among extra-chromosomal replicons

| REPLICON | PROBABILITY [a] |
|---|---|
| *Anaeba* sp. 90 chromosome chANA02 [CYANOBACTERIA : Chroococcales] (CP003285) | 0.638 |
| *Asticcacaulis excentricus* CB 48 chromosome 2 [α-PROTEOBACTERIA : Caulobacterales] (CP002396) | 0.637 |
| *Azospirillum brasilense* Sp245 plasmid AZOBR_p1 [α-PROTEOBACTERIA : Rhodospirillales] (HE577328) | 0.774 |
| *Methylobacterium extorquens* AM1 megaplasmid [α-PROTEOBACTERIA : Rhizobiales] (CP001511) | 0.669 |
| *Nocardioides dassonvillei* DSM 43111 chromosome 2 [ACTINOBACTERIA : Actinomycetales] (CP002041) | 0.539 |
| *Paracoccus denitrificans* PD1222 chromosome2 [α-PROTEOBACTERIA : Rhodobacterales] (CP000490) | 0.778 |
| *Prevotella intermedia* 17 chromosome II [BACTEROIDETES : Bacteroidales] (CP0033) | 0.984 |
| *Prevotella melaninogenica* ATCC 845 chromosome II [BACTEROIDETES : Bacteroidales] (CP002123) | 0.698 |

359 [a] Probability for an extra-chromosomal replicon, *i.e.*, plasmid or SER, to belong to the SER (a) or Chromosome
360 (b) class according to the supervised classification procedures.

361 **BOX 1. CHARACTERISTICS OF CANDIDATE SERS**

362 According to the literature, most candidate SERs that we detected among plasmids (Table 9a) were
363 expected to be essential to the cell functioning and/or to the fitness of the organism.

364 • *Azospirillum* genomes are constituted of multiple replicons, at least one of which is expected to be
365 essential. The largest extra-chromosomal replicon in *A. brasilense* was proposed to be essential for
366 bacterial life (Wisniewski-Dyé et al., 2011) since it encodes well-conserved housekeeping genes involved
367 in DNA replication, RNA metabolism and biosynthesis of nucleotides and cofactors, as well as in
368 transport and protein post-translational modifications. This replicon is unambiguously identified as a SER
369 by our analyses, as are additional replicons found in *A. lipoferum* and *A.* sp. B510, expected homologues

370 to *A. brasilense* SER (Acosta-Cruz et al*.,* 2012). In contrast, other extra-chromosomal replicons classified
371 as chromids by Wisniewski-Dyé et al*.* (2012) are unlikely to be true essential replicons. They were not
372 retrieved among our candidate SERs.

373 • In *Rhizobium etli* CFN42, functional interactions among sequences scattered in the different
374 extrachromosomal replicons are required for successful completion of life in symbiotic association
375 with plant roots or saprophytic growth (Brom et al., 2000). p42e (CP000137) is the only replicon other
376 than the chromosome that contains genes involved in the primary metabolism (Landeta et al., 2011;
377 Villaseñor *et al.* 2011) and evades its elimination by co-integration with other replicons including the
378 chromosome (Landeta et al., 2011). Furthermore, homologues to this replicon were identified in the
379 genomes of other *R. etli* strains as well as other *Rhizobium* species: *R. etli* CIAT652 pA,
380 *R. leguminosarum* bv. *viciae* 3841 pRL11, *R. leguminosarum* bv. *trifolii* WSM2304 pRLG202 and
381 *R. leguminosarum* bv. *trifolii* WSM1325 pR132502 (CP001075, AM236085, CP001193, and CP001624,
382 respectively) (Landeta et al., 2011; Villaseñor et al., 2011). These replicons were thus proposed to be
383 secondary chromosomes (Landeta et al., 2011).

384 • The genome of *Ensifer*/*Sinorhizobium meliloti* AK83 was the single multipartite-annotated
385 *Ensifer*/*Sinorhizobium* genomes present in our dataset. This bacterium carries two large extra-
386 chromosomal replicons that are involved in the establishment of the nitrogen fixation symbiosis with
387 legume plants. pSymA contains most of the genes involved in the nodulation and nitrogen fixation
388 whereas pSymB carries exopolysaccharide biosynthetic genes, also required for the establishment of the
389 symbiosis. Our analyses identifies candidate SERs similar to *S. meliloti* AK83 pSymA and pSymB in
390 other *S. meliloti* strains as well as in *S. fredii* and *S. medicae.* pSymB has been referred to as second
391 chromosome for carrying genes encoding essential house-keeping functions (Blanca-Ordóñez et al., 2010 ;
392 Galardini et al., 2011). It shows a higher level of conservation across strains and species than pSymA
393 (Galardini et al., 2013). pSymA, generally thought to be as stable as pSymB, greatly contribute to the
394 bacterial fitness in the rhizosphere (Blanca-Ordóñez et al., 2010; Galardini et al., 2013).

395 • The identification of *Methylobacterium extorquens* AM1 1.2 Mb megaplasmid as a SER is supported by
396 its presence in the genome in a predicted one copy number, by its coding a truncated *luxI* gene essential
397 for the operation of two chromosomally-located *luxI* genes, as well as the single *umuDC* cluster involved
398 in SOS DNA repair, and by the presence of a 130 kb region syntenic to a region of similar length in the

399 chromosome of *Methylobacterium extorquens* strain DM4 (Vuilleumier et al., 2009).

400 • The megaplasmid (821 kb) in *Ruegeria* sp. TM1040 carries more rRNA operons (3) than the
401 chromosome (1) and several unique genes (Moran et al., 2007). *Ruegeria* sp. TM1040 is the only species
402 in the *Roseobacter* group that possesses a SER. None of the plasmids in the other species included in our
403 datasets qualified as SERs according to our results in contrast to the commonly held view (Petersen et al.,
404 2013).

405 • In *Burkholderia* genus, additional ERs possess a centromere whose sequence is distinct from, but
406 strongly resembles that of the chromosome centromere (Dubarry et al., 2009). However, these
407 centromeres have a common origin and a plasmid ancestry (Passot et al., 2012). The evolution of these
408 replicons into SERs is best accounted for by the high level of plasticity observed in the *Burkholderia*
409 genomes, with extra-chromosomal replicons going through extensive exchange of genetic material among
410 them as well as with the chromosomes (Maida et al., 2014).

411 • *Acaryochloris marina* MBIC11017 pREB1 (CP000838) and pREB2 (CP000839) plasmids were
412 identified as candidate SERs. Both these megaplasmids code for metabolic key-proteins, and are thus
413 likely to contribute to the bacterium fitness (Swingley et al., 2008).

414 • The genomes of *Streptomyces cattleya* NRRL8057 and *S. clavuligerus* ATCC27064 harbour a linear
415 megaplasmid (1.8 Mb) that shows a high probability ($P \approx 0.7$) to be a SER. The megaplasmid of
416 *S. cattleya* NRRL8057 encodes genes involved in the synthesis of various antibiotics and secondary
417 metabolites and is expected to be important to the life of the bacterium in its usual habitat (Barbe et al.,
418 2011; O'Rourke et al., 2009). In *S. clavuligerus* ATCC27064, none of the megaplasmid-encoded genes are
419 expected to belong to the core genome (Medema et al., 2010). However, the megaplasmid is likely to
420 contribute to the bacterium firness. It represents more than 20% of the coding genome and constitutes a
421 large reservoir of genes involved in bioactive compound production and cross-regulation with the
422 chromosome (Medema et al., 2010). Furthermore, *S. clavuligerus* chromosome requires the SER-encoded
423 *tap* gene involved in the telomere replication.

424 • *Butyrivibrio proteoclasticus* B316 harbours two plasmid, one of which, pCY186 plasmid (CP001813),
425 was identified as a candidate SERs by our analysis, albeit with a low probability (0.56). In support to this,
426 it carries numerous genes coding for proteins involved in replication of the chromosome (Yeoman et al.,
427 2011). The second plasmid in that strain, pCY360 (CP001812), also proposed to be an essential replicon

428 in that bacterium (Yeoman et al., 2011), presents too low a probability (P = 0.32) in our analysis to qualify
429 as a SER.

430 **DISCUSSION**

431 The SERs clearly stand apart from plasmids, including those that occur consistently in a
432 bacterial species, *e.g.*, *Lactobacillus salivarius* pMP118-like plasmids (Li et al., 2007).
433 The replicon size proposed as a primary classification criterion to separate the SERs
434 from the plasmids (diCenzo and Finan, 2017; Harrison et al., 2010) proves to be
435 inoperative. The IS profiles accurately identify the SERs of *Leptospira* and *Butyrivibrio*
436 despite their plasmid-like size, and unambiguously ascribe the chromosomes in the
437 reduced genomes of endosymbionts (sizes down to 139 kb) to the chromosome class.
438 Conversely, they assign *Rhodococcus jostii* RHA1 1.12 Mb-long pRHL1 replicon to the
439 plasmid class, and do not discriminate the megaplasmids (>350 kb (diCenzo and Finan,
440 2017)) from smaller plasmids. Plasmids may be stabilized in a bacterial population by
441 rapid compensatory adaptation that alleviates the fitness cost incurred by their presence
442 in the cell (San Millan et al., 2014; Hall et al., 2017; Stalder et al., 2017). This
443 phenomenon involves mutations either on the chromosome only, on the plasmid only, or
444 both, and does not preclude the segregational loss of the plasmid. On the contrary, SERs
445 code for chromosome-type IS proteins that integrate them constitutively in the species
446 genome and the cell cycle. The SERs thence qualify as essential replicons regardless of
447 their size and of the phenotypical/ecological, possibly essential, functions that they
448 encode and which vary across host taxa.

449 Yet, SERs also carry plasmid-like ISs, suggesting a role for plasmids in their formation.
450 The prevailing opinion assumes that SERs derive from the amelioration of
451 megaplasmids (diCenzo and Finan, 2017; diCenzo et al., 2013; Harrison et al., 2010;

452 MacLellan et al., 2004; Slater et al., 2009): a plasmid bringing novel functions for the
453 adaptation of its host to a new environment is stabilized into the bacterial species
454 genome through the transfer from the chromosome of essential genes (diCenzo and
455 Finan, 2017; Slater et al., 2009). However, the generalized presence of chromosome-like
456 ISs in the SERs of the various taxa with multipartite genomes is unlikely to derive from
457 the action of environment-specific and lineage-specific selective forces. In reverse, all
458 bacteria with similar lifestyle and exhibiting some phylogenetic relatedness may not
459 harbor multiple ERs (*e.g.*, α-proteobacterial nitrogen-fixing legume symbionts). Also,
460 the gene shuttling from chromosome to plasmid proposition fails to account for the
461 situation met in the multipartite genomes of *Asticaccaulis excentricus, Paracoccus*
462 *denitrificans* and *Prevotella* species. Their chromosome-type ISs are evenly distributed
463 between the chromosome and the SER whereas their homologues in the mono- or
464 multipartite genomes of most closely related species are primarily chromosome-coded
465 (see Table 10 for an example). This pattern, mirrored in their whole gene content (Naito
466 et al., 2016; Poirion, 2014), hints at the stemming of the two essential replicons from a
467 single chromosome by either a splitting event or a duplication followed by massive gene
468 loss. Neither mechanism informs on the presence of plasmid-type maintenance
469 machinery on one of the replicons. The severing of a chromosome generates a single
470 true replicon carrying the chromosome replication origin and an origin-less remnant,
471 whilst the duplication of the chromosome produces two chromosomal replicons with
472 identical maintenance systems. Whereas multiple copies of the chromosome are known
473 to cohabit constitutively in polyploid bacteria (Ohtani et al., 2010), the co-occurrence of
474 dissimilar chromosomes bearing identical replication initiation and partition systems is
475 yet to be described in bacteria.

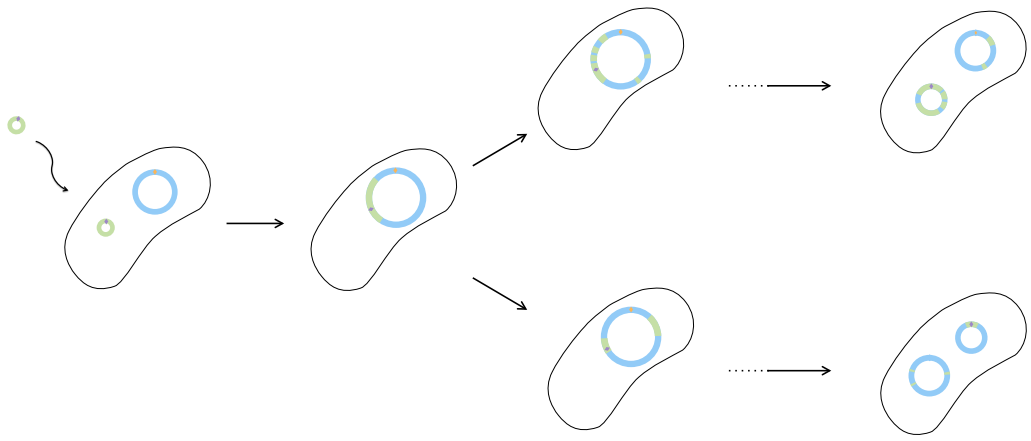
**Table 10. IS profiles of *Paracoccus denitrificans vs. Rhodobacter sphaeroides* (Rhodobacterales)**

Chromosome-like IS functions coded only by the SER in *P. denitrificans* or *R. sphaeroides* whilst by the chromosome in other Rhodobacterales are indicated by an asterisk. Numbers corresponds to the number of homologues (*P. denitrificans* PD1222) or the pourcentage of function-coding replicons (*R. sphaeroides* and Rhodobacterales genomes).

| | IS FUNCTION GROUP | | *Paracoccus denitrificans* PD1222 | | | *Rhodobacter sphaeroides* | | | Other Rhodobacterales | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | | | chromosome1 (CP000489) | chromosome2 (CP000490) | plasmid1 (CP000491) | chromosome % (n=12) | SER % (n=12) | plasmid % (n=14) | chromosome % (n=12) | plasmid % (n=33) | |
| KEGG entry (chromosome-like) | REPLICATION | CbpA | | 1 | | | 100 | | | 75 | |
| | | Dam | | 1 | | | 50 | 75 | | 25 | |
| | | DnaA | * | | 1 | | 100 | | | 92 | |
| | | DnaB | * | | 1 | | 100 | | | 100 | |
| | | DnaC | | | | | 25 | | | | 3 |
| | | DnaG | * | | 1 | | 100 | | | 100 | |
| | | Dps | | 1 | | | | | | 50 | 15 |
| | | Fis | * | | 1 | | 100 | | | 75 | |
| | | Hda | | | | | | | | 42 | |
| | | Hfq | * | | 1 | | 100 | | | 100 | |
| | | H-NS | | 2 | 2 | | 100 | | 29 | 67 | 9 |
| | | HupA | | | | | | | | 17 | |
| | | HupB | | 1 | | | 100 | | | 100 | |
| | | IciA | | 1 | | | | | 36 | 67 | |
| | | IhfA, HimA | | 1 | | | 100 | | | 100 | |
| | | IhfB, HimD | * | | 1 | * | | 100 | | 100 | |
| | | Lrp | | 4 | 1 | 3 | 100 | 25 | 50 | 100 | 9 |
| | | Rob | | | | | 25 | | | 8 | |
| | | ssb | | 1 | 3 | | 100 | 100 | | 100 | 3 |
| | PARTITION | Fic | | | | | | | 7 | 8 | 6 |
| | | GidA, MnmG, MTO1 | | 1 | | | 100 | | | 100 | |
| | | GidB, RsmG | | 1 | | | 100 | | | 100 | |
| | | MreB | * | | 1 | | 100 | | | 92 | |
| | | MreC | * | | 1 | | 100 | | | 92 | |
| | | MreD | * | | 1 | | 100 | | | 92 | |
| | | Mrp | | 1 | | | 100 | | | 100 | |
| | | ParA, Soj | | 4 | 1 | | 100 | | | 100 | 9 |
| | | ParB, Spo0J | | 3 | 1 | 1 | 100 | 75 | | 100 | 42 |
| | | ParC | | 1 | 1 | | 100 | | | 100 | |
| | | ParE | | 1 | | | 100 | | | 100 | |
| | | RodA, MrdB | * | | 1 | | 100 | | | 92 | |
| | | TrmFO, Gid | * | | 1 | | 100 | | | 100 | |
| | | XerC | | 1 | | | 100 | | | 100 | |
| | | XerD | | 1 | | | 100 | 25 | 7 | 100 | 6 |
| | SEGREGATION | ScpA | | 1 | | | 100 | | | 100 | |
| | | ScpB | | 1 | | | 100 | | | 100 | |
| | | Smc | | 1 | | | 100 | | | 100 | |
| | CELL DIVISION | AcrA | | 2 | 2 | * | | 100 | 7 | 58 | |
| | | AmiA, AmiB, AmiC | | 1 | | | 100 | | | 100 | |
| | | FtsA | * | | 1 | | 100 | | | 83 | |
| | | FtsE | | 1 | | | 100 | | 7 | 92 | |
| | | FtsI | | 1 | | | 100 | | | 92 | |
| | | FtsK, SpoIIIE | | 1 | | | 100 | | | 100 | |
| | | FtsQ | * | | 1 | | 100 | | | 92 | |
| | | FtsW, SpoVE | | | 1 | | 100 | | | 100 | |
| | | FtsX | | 1 | | | 100 | | 7 | 92 | |
| | | FtsZ | * | | 1 | | 100 | | | 100 | |
| | | MinC | | | | 1 | | | | | |
| | | MinD | | | | 1 | 75 | | | 75 | |
| | | MinE | | | | 1 | | | | | |
| | | ZapA | | 1 | | | 100 | | | 100 | |
| ACLAME family (plasmid-like) | REPLICATION | DNA helicase | | | | | 25 | 50 | 7 | 25 | |
| | | Helicase-1 | | 1 | 1 | | 100 | 50 | | 92 | 3 |
| | | Helicase-2 | | | | | | | | 8 | 3 |
| | | DNA repair | | | | | | | | 8 | |
| | | RepA | | 2 | 2 | | | | 43 | | 42 |
| | | RepAEB | | | | 1 | | 100 | 7 | | 15 |
| | | RepC | | 1 | | | 50 | | 43 | 8 | 24 |
| | | RuvB | | 1 | | | 100 | | 7 | 92 | |
| | PARTITION | ATPase, TyrK, ExoP | | 2 | | | 100 | 50 | 14 | 75 | 15 |
| | | ParA, ParM | | | 1 | 1 | | 100 | 100 | 8 | 88 |
| | | ParB | | | | | 75 | 100 | 93 | 25 | 64 |
| | | plasmid dimer resolution | | | | | | | 14 | | 15 |
| | | Serine recombinase | | 3 | | | 25 | | 50 | 50 | 30 |
| | | Tyrosine recombinase | | | | | 25 | | 7 | 33 | 6 |
| | | Xer-like tyrosine recombinase | | | 1 | | 25 | 50 | | 33 | 6 |
| | | XerD | | | | | 25 | | | | |
| | MAINTENANCE | Ccd (PSK) | | | | | | | 7 | | |
| | | HicAB (PSK) | | | | | | | 7 | | |
| | | HigBA (PSK) | | 2 | | | | | 14 | 17 | 6 |
| | | MazEF (PSK) | | | | | | | 7 | | 6 |
| | | ParDE (PSK) | | 2 | 6 | | 25 | 50 | 7 | 8 | 15 |
| | | PhD, Doc (PSK) | | | | | | | 7 | | |
| | | RelBE (PSK) | | 1 | | | | | | | 3 |
| | | VapBC/Vag (PSK) | | 3 | 2 | | | | 14 | 8 | 9 |

479 We propose that the requirement for maintenance system compatibility between co-
480 occurring replicons is the driving force behind the presence of plasmid-type replication
481 initiation and maintenance systems in bacterial SERs. Indeed, genes encoding
482 chromosome-like replication initiators (DnaA) are hardly found on SERs. When they
483 are, in *Paracoccus denitrificans*, *Prevotella intermedia* and *P. melaninogenica*, the
484 annotated chromosome in the corresponding genome does not carry one. Similarly,
485 chromosomal centromeres (*parS*) are found on a single replicon within a multipartite
486 genome, which is the chromosome in all genomes but one. In *P. intermedia*
487 (GCA_000261025.1), both replication initiation and partition systems define the SER as
488 the *bona fide* chromosome and the annotated chromosome as an extra-chromosomal
489 replicon. The harmonious coexistence of different replicons in a cell requires that they
490 use divergent enough maintenance systems. In the advent of a chromosome fission or
491 duplication, the involvement of an autonomously self-replicating element different from
492 the chromosome is mandatory to provide one of the generated DNA molecules with a
493 (non-chromosomal) maintenance machinery.
494 'Plasmid-first' and 'chromosome-first' hypotheses can be reconciled into a unified,
495 general Fusion-Shuffling-Scission model of SER emergence where a chromosome and a
496 plasmid combine into a cointegrate (Fig. 6). Plasmids are known to merge or to integrate
497 chromosomes in both experimental settings (Brom et al., 2000 ; Guo et al., 2003;
498 Iordănescu, 1975 ; Sýkora, 1992) and the natural environment (Cervantes et al., 2011;
499 Naito et al., 2016; Sýkora, 1992), as are the SER and chromosome of a multipartite
500 genome (Val et al., 2014; Xie et al., 2017; Yamamoto et al., 2018). When integrated, the
501 plasmids/SERs can thus replicate with the chromosome and persist in the bacterial
502 lineage through several generations (Cervantes et al., 2011; Val et al., 2014; Xie et al.,
503 2017). The co-integrate may resolve into its original components (Guo et al., 2003; Val

504 et al., 2014) or give rise to novel genomic architectures (Guo et al., 2003; Cervantes et
505 al., 2011; Val et al., 2014). The co-integration state likely facilitates inter-replicon gene
506 exchanges and genome rearrangements that may lead to the translocation of large
507 chromosome fragments to the resolved plasmid (Guo et al., 2003; Val et al., 2014).
508 Multiple cell divisions, and possibly several merging-resolution rounds, could provide
509 time and opportunity for the plasmid-chromosome re-assortment to take place, and for
510 multiple essential replicons and a viable distributed genome to form ultimately. In the
511 novel genome, one ER retains the chromosome-like origin of replication and
512 centrosome, and the other the plasmidic counterparts. The novel ERs differ from the
513 chromosome and plasmid that gathered in the progenitor host at the onset. They thus
514 constitute neo-chromosomes that carry divergent maintenance machineries and can
515 cohabit and function in the same cell. Depending on the number of cell cycles spent as
516 co-integrate, the level of genome reorganization, the acquisition of genetic material and
517 the environmental selective pressure acting upon the host, the final essential replicons
518 may exhibit diverse modalities of genome integration (Figure 6).



519

520 **Figure 6. Fusion-Shuffling-Scission model of distributed genome evolution**
521 Origins of replication are represented by diamonds.

522

523 The Fusion-Shuffling-Scission model of genome evolution that we propose accounts for
524 the extreme plasticity met in distributed genomes and the eco-phenotypic flexibility of
525 their hosts. Indeed, having a distributed genome appears to extend and accelerate the
526 exploration of the genome evolutionary landscape, producing complex regulation
527 (diCenzo et al., 2018; Galardini et al., 2015; Jiao et al., 2018) and leading to novel eco-
528 phenotypes and species diversification (*e.g.*, Burkholderiaceae and Vibrionaceae).
529 Furthermore, this model may explain the observed separation of the replicons according
530 to taxonomy. Chromosomes and plasmids thus appear as extremes on a continuum of a
531 lineage-specific genetic material.

## 532 MATERIALS AND METHODS

533 To understand the relationships between the chromosomal and plasmidic replicons, we
534 focused on the distribution of Inheritance System (IS) genes for each replicon and built
535 networks linking the replicons given their IS functional orthologues (Fig. 2).

### 536 Retrieval of IS functional homologues

537 A sample of proteins involved in the replication and segregation of bacterial replicons
538 and of the bacterial cell cycle was constructed using datasets available from the
539 ACLAME (Leplae et al., 2010) and KEGG (Kanehisa et al., 2012) databases. Gene
540 ontologies for “replication”, “partition”, “dimer resolution”, and “genome maintenance”
541 (Table 11) were used to select related ACLAME plasmid protein families (Table 1)
542 using a semi-automated procedure.

543 **Table 11. Gene ontologies related to plasmid ISs used to select groups of orthologous**
544 **proteins from the ACLAME database**

| PROCESS | ONTOLOGY | DESCRIPTION |
|---|---|---|
| Replication | *go:0006270* | DNA replication initiation |
| | *phi:0000268* | plasmid vegetative DNA replication |
| | *go:0003896* | DNA primase activity |
| | *go:0003887* | DNA-directed DNA polymerase activity |
| | *go:0045020* | error-prone DNA repair |
| | *go:0006260* | DNA replication |
| | *phi:0000114* | DNA helicase activity |
| | *go:0006281* | DNA repair |
| | *phi:0000196* | plasmid copy number control |
| Partition | *go:0003677* | DNA binding |
| | *575* | plasmid partitioning protein family ParB/Spo0J |
| | *go:0015616* | DNA translocase activity |
| | *576* | plasmid partitioning protein family ParM |
| | *go:0000146* | microfilament motor activity |
| | *go:0007059* | chromosome segregation |
| | *go:0015616* | DNA translocase activity |
| | *go:0007059* | chromosome segregation |
| | *go:0016887* | ATPase activity |
| | *go:0030541* | plasmid partitioning |
| | *go:0051302* | regulation of cell division |
| | *phi:0000196* | plasmid copy number control |
| Dimer resolution | *phi:0000134* | site specific DNA excision |
| | *phi:0000144* | serine based recombinase activity |
| | *phi : 0000131* | site specific DNA recombinaison |
| | *phi : 0000143* | Tyrosine-based recombinase activity |
| | *phi : 0000304* | plasmid dimer resolution |
| | *go : 0015616* | DNA translocase activity |
| | *phi:0000136* | transpositional recombination |
| Maintenance | *go : 0016740* | transferase activity |
| | *phi : 0000262* | toxin |
| | *phi:0000322* | PSK |
| | *547* | TA family parDE |
| | *544* | TA family epsilon zeta |
| | *go:0009008* | DNA methyltransferase activity |
| | *phi : 0000264* | nucleoid associated protein |
| | *go : 0006276* | plasmid maintenance |

545 KEGG orthology groups were selected following the KEGG BRITE hierarchical
546 classification (Table 2). Then, the proteins belonging to the relevant 92 ACLAME
547 protein families and 71 KEGG orthology groups (3,847 and 43,757 proteins,
548 respectively) were retrieved and pooled. Using this query set amounting to a total of

549 47,604 proteins, we performed a *blastp* search of the 6,903,452 protein sequences
550 available from the 5,125 complete sequences of bacterial replicons downloaded from
551 NCBI Reference Sequence database (RefSeq) (Pruitt et al., 2007) on 30/11/2012. We
552 identified 358,624 putative homologues using BLAST default parameters (Camacho et
553 al., 2009) and a $10^{-5}$ significance cut-off value. We chose this *E*-value threshold to
554 enable the capture of similarities between chromosome and plasmid proteins whilst
555 minimizing the production of false positives, *i.e.*, proteins in a given cluster exhibiting
556 small *E*-values despite not being functionally homologous. Using RefSeq ensured the
557 annotation consistency of the genomes included in our dataset.

558 **Clustering of IS functional homologues**

559 Using this dataset, we inferred clusters of IS functional homologues by coupling of an
560 *all-versus-all blastp* search using a $10^{-2}$ *E*-value threshold and a TRIBE-MCL (Enright et
561 al., 2002) clustering procedure. As input to TRIBE-MCL, we used the matrix of log
562 transformed *E*-value, $d(p_i, p_j) = -\log_{10}\left(e_{value}(p_i, p_j)\right)$, obtained from the comparisons
563 of all possible protein pairs. Using a granularity value of 4.0 (see below), we organized
564 the 358,624 IS homologues into 7013 clusters, each comprising from a single to 1990
565 proteins (Figure 3). We annotated IS homologues according to their best match (BLAST
566 hit with the lowest *E*-value) among the proteins of the query set, *i.e.*, according to one of
567 the 117 functions of the query set (71 from KEGG and 46 from ACLAME). Then, we
568 named the clusters of functional homologues using the most frequent annotation among
569 the proteins in the cluster. We used the number of protein annotations in a cluster to
570 determine the cluster quality, a single annotation being optimal. To select the best
571 granularity and to estimate the consistency of the clusters in terms of functional
572 homologues, we computed the weighted Biological Homogeneity Index (*wBHI*,

573 modified from the *BHI* (Datta and Datta, 2006), each cluster being weighted by its size)
574 and the Conservation Consistency Measure (*CCM*, similar to the *BHI* but using the
575 functional domains of the proteins to define the reference classes), which both take into
576 account the size distribution of the clusters (See next paragraph for details on index
577 calculation). The former gives an estimation of the overall consistency of clusters
578 annotations according to the protein annotations whereas the latter gives an estimation of
579 cluster homogeneity according to the protein domains identified beforehand. To build
580 the sets of functional domains, we performed an *hmmscan* (Finn et al., 2011) procedure
581 against the Pfam database (Finn et al., 2016) of each of the 358,624 putative IS
582 homologues. We annotated each protein according to the domain match(es) with *E*-value
583 $< 10^{-5}$ (individual *E*-value of the domain) and *c-E*-value $< 10^{-5}$ (conditional *E*- value that
584 measures the statistical significance of each domain). If two domains overlapped, we
585 only considered the domain exhibiting the smallest *E*-value. We estimated *wBHI* and
586 *CCM* indices for the clustering of the IS homologues and compared with values obtained
587 for random clusters simulated according to the cluster size distribution of the IS proteins,
588 irrespective of their length or function. For each of the clustering obtained for different
589 granularities, we constructed a random clustering following the original cluster size
590 distribution (assessed with a $\chi^2$ test) and composed with simulated proteins according to
591 the distributions of the type and number of functional domains of the data collected from
592 the 358,624 IS homologues. Overall, the clusters obtained using a granularity of 4.0 with
593 the TRIBE-MCL algorithm appeared to be homogenous in terms of proteins similarities
594 toward their best BLAST hits and their functional domain distributions (see below).

595 **Evaluation of the clustering procedures**

596 In order to select the best granularity and to estimate the consistency of the clusters in

terms of functional homologs, we computed the *weighted Biological Homogeneity Index* (*wBHI*) and the *Conservation Consistency Measure* (*CCM*). The former gives an estimate of the overall consistency of clusters annotations according to the protein annotations whereas the latter gives an estimate of cluster homogeneity according to protein domains identified beforehand. Although close to the *Biological Homogeneity Index* (*BHI*) introduced by Datta and Datta (2006), both these indices take into account the size distribution of the clusters.

The *BHI* was originally introduced to measure the biological homogeneity of clusters according to reference classes to evaluate clusters obtained with microarray data (Datta and Datta, 2006). Given a clustering $C=\{C_1,...,C_k\}$ of $k$ clusters with $n_i$ the size of the cluster $C_i$, a set of $m$ proteins $P=\{P_1,...,P_m\}$ and a set $r$ of reference classes $R$ where each class $R_i$ could be linked to the $m$ proteins, the BHI is defined as:

$$BHI = \frac{1}{k}\sum_{i=0}^{i<k} c_i$$

where $c_i$ is defined as:

$$c_i = \frac{1}{\left(n_i(n_i-1)\right)}\sum_{P_i,P_j \in C_i} d\left(P_i,P_j\right)$$

where $d(P_i,P_j)=1$ if $P_i$ and $P_j$ share at least one common reference class, and $d(P_i,P_j)=0$ otherwise. The reference classes here are the annotations defined according to the protein best BLAST hit. The *BHI* is thus an easy-to-interpret measure, which value is maximal when, for all clusters, all the proteins in a cluster share at least one annotation.

The *wBHI* is a modification of the *BHI*, where each cluster is weighted by its size $m$. Following the previous notation scheme, the *wBHI* is defined as:

$$wBHI = \frac{1}{m}\sum_{i=0}^{i<k} 2.c_i.n_i$$

616 The *CCM* is similar to the *BHI* but the functional domains of the proteins are used to
617 define the reference classes. The distance between the proteins is here computed as the
618 Jaccard distance between the functional domain sets of the proteins. Every protein $P_i$ can
619 be described as a vector of functional domains, $D_{Pi} = \{d_1,...,d_x\}$. The Jaccard distance
620 between the two sets of domains $d_2(P_i,P_j)$ can be defined as:

$$d_2(P_1, P_2) = 1 - \frac{|D_{P_1} \cap D_{P_2}|}{|D_{P_1} \cup D_{P_2}|}$$

621 where $D_{P_1}$ and $D_{P_2}$ are the clans or domains (when no clan could be assigned) identified
622 in $P_1$ and $P_2$ respectively. For a given cluster $C_i$, the *CCM* is calculated as:

$$CCM = \frac{1}{m}\sum_{i=0}^{i<k} 2.c'_i.n_i$$

623 where $c'_i$ is defined as:

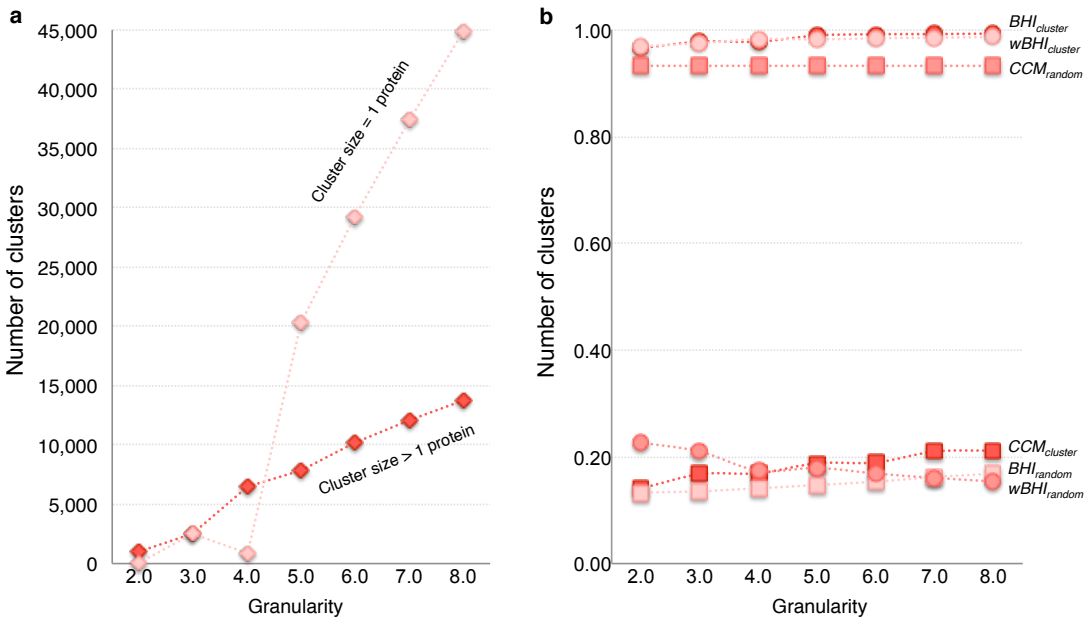
$$c'_i = \frac{1}{(n_i(n_i-1))}\sum_{P_i,P_j \in C_i} d_2(P_i, P_j)$$

624 Clusters which proteins have similar domains result in a *CCM* value close to 0, whereas
625 a *CCM* value close to 1 indicates that the clusters hold proteins with little domain
626 overlap.

627 **Choice of the clustering granularity**

628 We tested several levels of granularity to optimize the TRIBE-MCL clustering and
629 obtain the most informative IS clustering in terms of functional linkage. Too low a
630 granularity would produce large clusters containing multiple functional families. In turn,
631 increasing the granularity results in the tightening of the cluster. A high granularity tends
632 to split clusters harboring different protein subfamilies (*e.g.,* a cluster composed of
633 proteins from the tyrosine recombinase superfamily) and to produce multiple clusters of
634 proteins belonging to a single function family according to their level of sequence

635 dissimilarity. Furthermore, too high a granularity would result in the formation of
636 numerous single protein clusters, and would dramatically increase the computation times
637 of the following analyses. A granularity level of 4.0 constituted a good compromise
638 (Figure 8). Values of *CCM* and *BHI* are slightly improved compared to granularities of
639 2.0 and 3.0, and the high but still workable number of clusters is expected to prevent the
640 formation of clusters mingling distinct protein subfamilies.




641 **Figure 8. Influence of granularity on the clustering**

642 (a) Number of clusters with more than one protein (dark diamonds) or clusters holding a single protein (pale
643 diamonds). (b) *BHI* (dark), *wBHI* (pale) and *CCM* (medium) scores obtained with random clusters (squares) and
644 normal clusters (circles), respectively.

645 **Assessment of the homogeneity of IS functional homologues**

646 The homogeneity towards the functions of the proteins in the query set relied on the
647 assumption that the first BLAST cut-off ($10^{-5}$ *E*-value) was stringent enough to capture
648 only functional homologues to the query proteins. Potential bias might nevertheless arise
649 from query proteins possessing a supplementary functional domain unrelated to the IS
650 role, or from the selection of proteins belonging to the same superfamily but differing in

651 function. To address these issues, we calculated the functional vectors associated to each
652 KEGG group or ACLAME family of the query set, as well as those for all obtained
653 clusters. For a protein $P_i$, we defined the associated functional vector with respect to its
654 set of identified domains $D_{P_i}$ and to the set of all identified domains $D=\{d_1,...,d_X\}$ as:

$$v_{P_i} = \left(n_{d_1}^{P_i}, \dots, n_{d_x}^{P_i}\right)$$

655 where $n_{d_i}^{P_i}$ is the number of time $d_i$ is found in $D_{P_i}$. The functional vector associated to a given
656 cluster of proteins $C_i$ could then be defined as:

$$v_{C_i} = \left(n_{d_1}^{C_i}, \dots, n_{d_x}^{C_i}\right)$$

657 where $n_{d_i}^{C_i}$ is defined as:

$$n_{d_i}^{C_i} = \frac{1}{|C_i|} \sum_{P_i \in C_i} n_{d_x}^{P_j}$$

658 For each cluster $C_0$, the cosine distance between its associated vector $v_{C_0}$ and the associated
659 vector $v_{C_a}$ of the corresponding KEGG group or ACLAME family annotations $C_a$ was then
660 computed as:

$$d_{cosine}\left(v_{C_a}, v_{C_0}\right) = 1 - \frac{\sum_{i=1}^{X} n_{d_i}^{C_0} . n_{d_i}^{C_a}}{\sqrt{\sum_{i=1}^{X} n_{d_i}^{C_0}{}^2} \cdot \sqrt{\sum_{i=1}^{X} . n_{d_i}^{C_a}{}^2}}$$

661 For each cluster $C_0$, the cosine distance between its associated vector $v_{C_0}$ and the
662 associated vector $v_{C_a}$ of the corresponding KEGG group or ACLAME family
663 annotations $C_a$ was then computed as:

$$d_{cosine}\left(v_{C_a}, v_{C_0}\right) = 1 - \frac{\sum_{i=1}^{X} n_{d_i}^{C_0} . n_{d_i}^{C_a}}{\sqrt{\sum_{i=1}^{X} n_{d_i}^{C_0}{}^2} \cdot \sqrt{\sum_{i=1}^{X} . n_{d_i}^{C_a}{}^2}}$$

664 The $d_{cosine}\left(v_{C_a}, v_{C_0}\right)$ values were compared with those obtained using random clusters
665 $C_r$ of the same size than $C_0$. For each $C_0$ and its corresponding $C_a$, 200 random clusters
666 and their associated distances $d_{cosine}\left(v_{C_a}, v_{C_r}\right)$, from which the corresponding empirical

667 distribution $D_e$ was constructed, were computed. $C_0$ is then considered as noise if

668 $d_{cosine}(v_{C_a}, v_{C_0}) \notin Q_{10\%}^{D_e}$ where $Q_{10\%}^{D_e}$ is the 0.1-quantile of $D_e$.

669 **Unsupervised analyses of the replicon space**

670 We represented the bacterial replicons (Supplementary Table 1) as vectors according to

671 their content in IS genes. The number of IS protein clusters retained for the analysis

672 determined the vector dimension and the number of proteins in a replicon assigned to

673 each cluster gave the value of each vector component. We built matrices

674 $P = \begin{bmatrix} p_{1,1} & \cdots & p_{1,m} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,m} \end{bmatrix}$, where $n$ is the number of replicons, $m$ the number of protein

675 clusters, and $p_{i,j}$ the number of proteins of the $j^{th}$ cluster encoded by a gene present on the

676 $i^{th}$ replicon. We constructed several datasets to explore both the replicon type and the

677 host taxonomy effects on the separation of the replicons in the analyses (Table 12).

678 **Table 12. Reference classes used in the evaluation of the replicon IS protein-based**
679 **unsupervised clustering solutions**

| EVALUATED SEPARATION | ENSEMBLE | NORMALIZED ENSEMBLE [a] |
|---|---|---|
| Chromosomes *vs.* Plasmids | $\{R^{\{chromosome\}}, R^{\{plasmid\}}\}$ | $\{\overline{Kl}_{genus}^{R^{\{chromosome\}}}, \overline{Kl}_{genus}^{R^{\{plasmid\}}}\}$ |
| Chromosomes *per* host phylum | $Kl_{phylum}^{chromosome}$ | $\overline{Kl}_{genus}^{K} \mid K \in Kl_{phylum}^{chromosome}$ |
| Chromosomes *per* host class | $Kl_{class}^{chromosome}$ | $\overline{Kl}_{genus}^{K} \mid K \in Kl_{class}^{chromosome}$ |
| Plasmids *per* host phylum | $Kl_{phylum}^{plasmid}$ | $\overline{Kl}_{genus}^{K} \mid K \in Kl_{phylum}^{plasmid}$ |
| Plasmids *per* host class | $Kl_{class}^{plasmid}$ | $\overline{Kl}_{genus}^{K} \mid K \in Kl_{class}^{plasmid}$ |

680 [a] Normalisation according to host genus

681 The taxonomic representation bias was taken into account by normalizing the data with

682 regard to the host genus: a consensus vector was built for each bacterial genus present in
683 the datasets. The value of each vector attribute was calculated as the mean of the
684 attribute values in the vectors of the replicons that belong to the same bacterial genus.
685 As a first approach, we transformed data into bipartite graphs whose vertices are the
686 replicons and the proteins clusters. The graphs were spatialized using the force-directed
687 layout algorithm ForceAtlas2 (Jacomy et al., 2014) implemented in Gephi (Bastian et al.,
688 2009). Bipartite graphs are a powerful way of representing the data by naturally drawing
689 the links between the replicons while enabling the detailed analysis of the IS cluster-
690 based connections of each replicon by applying forces to each node with regard to its
691 connecting edges. To investigate further the IS-based relationships of the replicons, we
692 applied the community structure detection algorithm INFOMAP (Rosvall and
693 Bergstrom, 2008) using the *igraph* python library (Csardi and Nepusz, 2006). We also
694 performed a WARD hierarchical clustering (Johnson, 1967) after a dimension reduction
695 of the data using a Principal Component Analysis (Hotelling, 1933). To select an
696 optimal number of principal components, we relied on the measurements of the cluster
697 stabilities using a *stability criterion* (Hennig, 2007) and retained the first 30 principal
698 components (57% of the total variance). For consistency purpose, the number of clusters
699 in the WARD analysis was chosen to match that obtained with the INFOMAP
700 procedure. The number of clusters used was assessed by the stability index by Fang and
701 Wang (2012) (Table 3). The quality of the projection and clustering results were
702 confirmed using the V-measure indices (Rosenberg and Hirschberg, 2007) (*homogeneity*,
703 *completeness*, *V-measure*) as external cluster evaluation measures (Table 3). The
704 *homogeneity* indicates how uniform clusters are towards a class of reference. The
705 *completeness* indicates whether reference classes are embedded within clusters. The *V-*
706 *measure* is the harmonic mean between these two indices and indicates the quality of a

707 clustering solution relative to the classes of reference. These three indices vary between
708 0 and 1, with values closest to 1 reflecting the good quality of the clustering solution.
709 The type of replicons (*i.e.,* plasmid or chromosome) and the taxonomic affiliation
710 (phylum or class) for chromosomes or plasmids were used as references classes (Table
711 12). Additionally, the *stability criterion* (Hennig, 2007) of individual clusters, weighted
712 by their size, for a given clustering result was evaluated using the bootstrapping of the
713 original dataset as re-sampling scheme. Individual Jaccard coefficient for each replicon
714 were computed as the number of times that a given replicon of a cluster in a clustering
715 solution is also present in the closest cluster in the resampled datasets.

716 **Functional characterization of the replicons and genomes**

717 In order to characterize the functional bias of the replicons, 117 IS functionalities (46
718 from ACLAME and 71 from KEGG) were considered. When equivalent in plasmids and
719 chromosomes, functions from ACLAME and KEGG databases were considered to be
720 distinct. A $n*m$ matrix $F = \begin{bmatrix} f_{1,1} & \cdots & f_{1,m} \\ \vdots & \ddots & \vdots \\ f_{n,1} & \cdots & f_{n,m} \end{bmatrix}$ with $n$ the number of replicons and $m$ the
721 number of IS functionalities, was used as input to the projection algorithms. $f_{i,j}$
722 represents the number of times that genes coding for proteins annotated with the $j^{th}$
723 function are present on the $i^{th}$ replicon. Several datasets were analysed using PCA
724 dimension reduction of the data followed by WARD hierarchical clustering (Table 3).

725 **Logistic regression analyses**

726 Several reference classes of replicons and complete genomes were considered for
727 comparison (Table 13). Ambiguous, *i.e.*, potentially adapted, plasmids belonging to
728 INFOMAP clusters of plasmid replicons partially composed of SERs and/or
729 chromosomes were removed from the plasmid class. When appropriate, the taxonomic

730 representation bias was taken into account by normalizing the data with regard to the
731 host genus as before. Logistic regressions (McCullagh and Nelder, 1989) were
732 performed for the 117 IS functions using the R glm package coupled to the python
733 binder rpy2. The computed $P_{value}$ measured the probability of a functionality to be
734 predictive of a given group of replicons/genomes and the *Odd-Ratio* estimated how the
735 functionality occurrence influenced the belonging of a replicon/genome to a given
736 group.

737 **Table 13. Datasets used in the logistic regression analyses**

| ENSEMBLE OF REPLICONS OR GENOMES | NOTATION | DATASET | DIMENSION[a] |
|---|---|---|---|
| Genus-normalized SERs | $E_{SER}$ | $\bar{V}_{f,genus}^{R\{SER\}}$ | (28, 117) |
| Genus-normalized plasmids | $E_{plasmid}$ | $\bar{V}_{f,genus}^{R\{plasmid\}}$ | (262, 117) |
| Genus-normalized chromosomes | $E_{chromosome}$ | $\bar{V}_{f,genus}^{R\{chromosome\}}$ | (560, 117) |

738 [a] (Number of replicons, number of functions)

739 **Supervised classification of replicons and genomes**

740 In order to identify putative ill-defined SERs and chromosomes amongst plasmids, we
741 performed supervised classification analyses using random forest procedures (Geurts et
742 al., 2006). We used the IS functionalities as the set of features and the whole sets of
743 chromosomes, plasmids and SER as sets of samples to build four classification studies
744 (Table 7) and detect SER candidates (plasmids *vs.* SERs) and chromosome candidates
745 (chromosomes *vs.* SERs or chromosomes *vs.* plasmids). Because of the unbalanced sizes
746 of the training classes (SERs *vs.* chromosomes and plasmids), iterative sampling
747 procedures were performed using 1000 random subsets of the largest class, with a size

748 similar to that of the smallest class. The ensuing results were averaged to build the class
749 probabilities and relative importance of the variables. We also used the whole set of
750 plasmids when compared to SERs, to identify more robust SER candidates. The
751 discarding of plasmids in the iterative procedure increases the classifier sensitivity while
752 reducing the rate of false negatives by including more plasmid-annotated putative true
753 SERs, whereas it decreases the classifier precision while increasing the rate of false
754 positives. The ExtraTreeClassifier (a classifier similar to Random Forest) class from the
755 Scikit-learn python library (Pedregosa et al., 2011) was used to perform the
756 classifications, with *K=1000*, *max_feat=sqrt(number of variables)* and *min_split=1*. For
757 each run, the *feature_importances* and *estimate_proba* functions were used to compute,
758 respectively, the relative contribution of the input variables and the class probabilities of
759 replicons/genomes. The statistical probability of a replicon/genome belonging to a class
760 was calculated as the average predicted class of the trees in the forest. The relative
761 contribution of the input variables was estimated according to Breiman (2001). The
762 choices of the number of trees in the forest *K*, the number of variables selected for each
763 split *max_feat*, and the minimum number of samples required to split an internal node
764 *min_split* were cross-validated using a *Leave-One-Out* scheme. The performance of the
765 *Extremely-randomized-trees* classification procedures was assessed using a stratified 10-
766 fold cross-validation procedure following Han *et al.* (2012), and the out-of-bag estimate
767 (OOB score) (Izzenman, 2008; Pedregosa et al., 2011) computed using the *oob_score*
768 function of Scikit-learn python library.

769 **Data availability**

770 The data supporting the findings of this study are available within the Article and
771 its Supplementary Information or are available from the authors.

772 **REFERENCES**


773 Acosta-Cruz E, Wisniewski-Dyé F, Rouy Z, Barbe V, Valdés M, Mavingui P. 2012.
774 Insights into the 1.59-Mbp largest plasmid of *Azospirillum brasilense* CBG497.
775 *Archives in Microbiology*. **194**:725-736. doi: 10.1007/s00203-012-0805-2.

776 Alav I, Sutton JM, Rahman KM. 2018. Role of bacterial efflux pumps in biofilm
777 formation. *Journal of Antimicrobial Chemotherapy*. **73**:2003-2020. doi:
778 10.1093/jac/dky042.

779 Anes J, McCusker MP, Fanning S, Martins M. 2015. The ins and outs of RND efflux
780 pumps in Escherichia coli. *Frontiers in Microbiology*. **6**:587. doi:
781 10.3389/fmicb.2015.00587.

782 Baek JH, Chattoraj DK. 2014. Chromosome I controls chromosome II replication in
783 *Vibrio cholerae*. *PLOS Genetics*. **10**:e1004184. doi:
784 10.1371/journal.pgen.1004184.

785 Barbe V, Bouzon M, Mangenot S, Badet B, Poulain J, Segurens B, Vallenet D, Marlière
786 P, Weissenbach J. 2011. Complete genome sequence of *Streptomyces cattleya*
787 NRRL 8057, a producer of antibiotics and fluorometabolites. *Journal of*
788 *Bacteriology*. **193**:5055-5056. doi: 10.1128/JB.05583-11.

789 Bastian M, Heymann S, Jacomy M. 2009. Gephi: An open source software for exploring
790 and manipulating networks. *Third International AAAI Conference on Weblogs*
791 *and Social Media, San Jose Mc Enery Convention Center, May 17, 2009 - May*
792 *20, 2009: AAAI Publications*.

793 Blanco-Ordóñez H, Oliva-García JJ, Pérez-Mendoza D, Soto MJ, Olivares J, Sanjúan J,
794 Nogales J. 2010. pSymA-dependent mobilization of the *Sinorhizobium meliloti*
795 pSymB megaplasmid. *Journal of Bacteriology*. **192**:6309-6312. doi:

796 10.1128/JB.00549-10.
797 Breiman L. 2001. Random forests. *Machine Learning*. **45**:5-32. doi:
798 10.1023/A:1010933404324.
799 Brom S, García-de los Santos A, Cervantes L, Palacios R, Romero D. 2000. In
800 *Rhizobium etli* symbiotic plasmid transfer, nodulation competitivity and cellular
801 growth require interaction among different replicons. *Plasmid*. **44**:34-43. doi:
802 10.1006/plas.2000.1469
803 Camacho CJ, Coulouris G, Avagyan V, Ma N, Papdopoulos J, Bealer K, Madden TL.
804 2009. BLAST+: architecture and applications. *BMC Bioinformatics*. **10**:421. doi:
805 10.1186/1471-2105-10-421.
806 Casjens SR. 1998. The diverse and dynamic structure of bacterial genomes. *Annual*
807 *review of genetics*. **32**:339-377. doi: 10.1146/annurev.genet.32.1.339.
808 Cervantes L, Bustos P, Girard L, Santamaría RI, Dávila G, Vinuesa P, Romero D, Brom
809 S. 2011. The conjugative plasmid of a bean-nodulating *Sinorhizobium fredii*
810 strain is assembled from sequences of two *Rhizobium* plasmids and the
811 chromosome of a *Sinorhizobium* strain. *BMC microbiology*. **11**:149. doi:
812 10.1186/1471-2180-11-149.
813 Chen Y, Milam SL, Erickson HP. 2012. SulA inhibits assembly of FtsZ by a simple
814 sequestration mechanism. *Biochemistry.* **51**:3100–3109. doi: 10.1021/bi201669d.
815 Chodavarapu S, Jones AD, Feig M, Kaguni JM. 2016. DnaC traps DnaB as an open ring
816 and remodels the domain that binds primase. *Nucleic Acids Research*. **44**:210-
817 220. doi: 10.1093/nar/gkv961.
818 Csárdi G, Nepusz T. 2006. The igraph software package for complex network research.
819 *InterJournal, Complex Systems*.1695.
820 Datta S, Datta S. 2006. Methods for evaluating clustering algorithms for gene expression

821 data using a reference set of functional classes. *BMC Bioinformatics*. **7**:397. doi:
822 10.1186/1471-2105-7-397.
823 De Nisco NJ, Abo RP, Wu CM, Penterman J, Walker GC. 2014. Global analysis of cell
824 cycle gene expression of the legume symbiont *Sinorhizobium meliloti*.
825 *Proceedings of the National Academy of Sciences of the United States of*
826 *America*. **111**:3217-3224. doi: 10.1073/pnas.1400421111.
827 Deghelt M, Mullier C, Sternon JF, Francis N, Laloux G, Dotreppe D, Van der Henst C,
828 Jacobs-Wagner C, Letesson JJ, De Bolle X. 2014. G1-arrested newborn cells are
829 the predominant infectious form of the pathogen *Brucella abortus*. *Nature*
830 *communications*. **5**:4366. doi: 10.1038/ncomms5366.
831 Demarre G, Galli E, Muresan L, Paly E, David A, Possoz C, Barre F-X. 2014.
832 Differential management of the replication terminus regions of the two *Vibrio*
833 *cholerae* chromosomes during cell division. *PLOS Genetics*. **10**:e1004557. doi:
834 10.1371/journal.pgen.1004557.
835 diCenzo G, Milunovic B, Cheng J, Finan TM. 2013. The tRNAarg gene and *engA* are
836 essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti*
837 and were translocated together from the chromosome in an ancestral strain.
838 *Journal of Bacteriology*. **195**:202-212. doi: 10.1128/JB.01758-12.
839 diCenzo GC, Finan TM. 2017. The divided bacterial genome: structure, function, and
840 evolution. *Microbiology and Molecular Biology Reviews*. **81**:e00019-00017. doi:
841 10.1128/MMBR.00019-17.
842 diCenzo GC, MacLean AM, Milunovic B, Golding GB, Finan TM. 2014. Examination
843 of prokaryotic nultipartite genome evolution through experimental genome
844 reduction. *PLOS Genetics*. **10**:e1004742. doi: 10.1371/journal.pgen.1004742.
845 diCenzo GC, Wellappili D, Golding GB, Finan TM. 2018. Inter-replicon gene flow

846 contributes to transcriptional integration in the *Sinorhizobium meliloti*
847 multipartite genome. *G3 (Bethesda)*. **8**:1711-1720. doi: 10.1534/g3.117.300405.
848 Dillon SC, Dorman CJ. 2010. Bacterial nucleoid-associated proteins, nucleoid structure
849 and gene expression. *Nat Rev Microbiol*. **8**:185-195. doi: 10.1038/nrmicro2261.
850 Drevinek P, Baldwin A, Dowson CG, Mahenthiralingam E. 2008. Diversity of the *parB*
851 and *repA* genes of the *Burkholderia cepacia* complex and their utility for rapid
852 identification of *Burkholderia cenocepacia*. *BMC microbiology*. **8**:44. doi:
853 10.1186/1471-2180-8-44.
854 Du D, Wang-Kan X, Neuberger A, van Veen HW, Pos KM, Piddock LJV, Luisi BF.
855 2018. Multidrug efflux pumps: structure, function and regulation. *Nature Review*
856 *in Microbiology*. **16**:523-539. doi: 10.1038/s41579-018-0048-6.
857 Dubarry N, Pasta F, Lane D. 2006. ParABS systems of the four replicons of
858 *Burkholderia cenocepacia*: new chromosome centromeres confer partition
859 specificity. *Journal of Bacteriology*. **188**:1489-1496. doi:
860 10.1128/JB.188.4.1489-1496.2006.
861 Egan ES, Lobner-Olesen A, Waldor MK. 2004. Synchronous replication initiation of the
862 two *Vibrio cholerae* chromosomes. *Current Biology*. **14**:R501-502. doi:
863 10.1016/j.cub.2004.06.036.
864 Egan ES, Waldor MK. 2003. Distinct replication requirements for the two *Vibrio*
865 *cholerae* chromosomes. *Cell*. **114**:521-530. doi: 10.1016/s0092-8674(03)00611-
866 1.
867 Enright AJ, Dongen S, Ouzounis C. 2002. An efficient algorithm for large-scale
868 detection of protein families. *Nucleic Acids Research*. **30**:1575-1584. doi:
869 10.1093/nar/30.7.1575.
870 Fang Y, Wang J. 2012. Selection of the number of clusters via the bootstrap method.

871 *Computational Statistics & Data Analysis*. **56**:468-477. doi:
872 10.1016/j.csda.2011.09.003.
873 Fiebig A, Keren K, Theriot JA. 2006. Fine-scale time-lapse analysis of the biphasic,
874 dynamic behaviour of the two *Vibrio cholerae* chromosomes. *Molecular*
875 *Microbiology*. **60**:1164-1178. doi: 10.1111/j.1365-2958.2006.05175.x.
876 Finn RD, Clements J, Eddy SR. 2011. HMMER web server: interactive sequence
877 similarity searching. *Nucleic Acids Research*. **39**:W29-37. doi:
878 10.1093/nar/gkr367.
879 Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, Potter SC, Punta
880 M, Qureshi M, Sangrador-Vegas A, Salazar GA, Tate J, Bateman A. 2016. The
881 Pfam protein families database: towards a more sustainable future. *Nucleic Acids*
882 *Research*. **44**:D279-285. doi: 10.1093/nar/gkv1344.
883 Frage B, Dohlemann J, Robledo M, Lucena D, Sobetzko P, Graumann PL, Becker A.
884 2016. Spatiotemporal choreography of chromosome and megaplasmids in the
885 *Sinorhizobium meliloti* cell cycle. *Molecular Microbiology*. **100**:808-823. doi:
886 10.1111/mmi.13351.
887 Galardini M, Mengoni A, Brilli M, Pini F, Fioravanti A, Lucas S, Lapidus A, Cheng JF,
888 Goodwin L, Pitluck S, Land M, Hauser L, Woyke T, Mikhailova N, Ivanova N,
889 Daligault H, Bruce D, Detter C, Tapia R, Han C, Teshima H, Mocali S,
890 Bazzicalupo M, Biondi EG. 2011. Exploring the symbiotic pangenome of the
891 nitrogen-fixing bacterium *Sinorhizobium meliloti*. *BMC Genomics*. **12**:235. doi:
892 10.1186/1471-2164-12-235.
893 Galardini M, Pini F, Bazzicalupo M, Biondi G, Mengoni A. 2013. Replicon-dependent
894 bacterial genome evolution: the case of *Sinorhizobium meliloti*. *Genome Biology*
895 *and Evolution*. **5**:542-558. doi: 10.1093/gbe/evt027.

896 Galardini M, Brilli M, Spini G, Rossi M, Roncaglia B, Bani A, Chiancianesi M, Moretto
897 M, Engelen K, Bacci G, Pini F, Biondi EG, Bazzicalupo M, Mengoni A. 2015.
898 Evolution of Intra-specific Regulatory Networks in a Multipartite Bacterial
899 Genome. *PLoS Computational Biology* **11**:e1004478. doi:
900 10.1371/journal.pcbi.1004478.
901 Galli E, Poidevin M, Le Bars R, Desfontaines JM, Muresan L, Paly E, Yamaichi Y,
902 Barre FX. 2016. Cell division licensing in the multi-chromosomal *Vibrio*
903 *cholerae* bacterium. *Nat Microbiology*. **1**:16094. doi:
904 10.1038/nmicrobiol.2016.94.
905 Gerding MA, Chao MC, Davis BM, Waldor MK. 2015. Molecular dissection of the
906 essential features of the origin of replication of the second *Vibrio cholerae*
907 chromosome. *MBio*. **6**:e00973. doi: 10.1128/mBio.00973-15.
908 Geurts P, Ernst D, Wehenkel L. 2006. Extremely randomized trees. *Machine Learning*.
909 **63**:3-42. doi: 10.1007/s10994-006-6226-1.
910 Guo FB, Ning LW, Huang J, Lin H, Zhang HX. 2010. Chromosome translocation and its
911 consequence in the genome of *Burkholderia cenocepacia* AU-1054. *Biochemical*
912 *and Bophysical Research Communications*. **403**:375-379. doi:
913 10.1016/j.bbrc.2010.11.039.
914 Guo X, Flores M, Mavingui P, Fuentes SI, Hernandez G, Davila G, Palacios R. 2003.
915 Natural genomic design in *Sinorhizobium meliloti*: novel genomic architectures.
916 *Genome Research*. **13**:1810-1817. doi: 10.1101/gr.1260903.
917 Hall JPJ, Brockhurst MA, Dytham C, Harrison E. 2017. The evolution of plasmid
918 stability: Are infectious transmission and compensatory evolution competing
919 evolutionary trajectories? *Plasmid*. **91**:90-95. doi:
920 10.1016/j.plasmid.2017.04.003.

921 Han J, Kamber M, Pei J. 2012. Data Mining: Concepts and Techniques, Third Edition:
922 Morgan kaufmann, Elsevier.
923 Harrison PW, Lower RP, Kim NK, Young JPW. 2010. Introducing the bacterial
924 'chromid': not a chromosome, not a plasmid. *Trends in Microbiology*. **18**:141-
925 148. doi: 10.1016/j.tim.2009.12.010.
926 Hennig C. 2007. Cluster-wise assessment of cluster stability. *Computational Statistics &*
927 *Data Analysis*. **52**:258-271. doi: 10.1016/j.csda.2006.11.025.
928 Hotelling H. 1933. Analysis of a complex of statistical variables into principal
929 components. *Journal of Educational Psychology*. **24**:417. doi:
930 10.1037/h0071325.
931 Izenman AJ. 2008. Modern multivariate statistical techniques: regression, classification,
932 and manifold learning: Springer-Verlag, New York Inc.
933 Jacomy M, Venturini T, Heymann S, Bastian M. 2014. ForceAtlas2, a continuous graph
934 layout algorithm for handy network visualization designed for the Gephi
935 software. *PLoS One*. **9**:e98679. doi: 10.1371/journal.pone.0098679.
936 Jiao J, Ni M, Zhang B, Zhang Z, Young JPW, Chan TF, Chen WX, Lam HM, Tian CF.
937 2018. Coordinated regulation of core and accessory genes in the multipartite
938 genome of Sinorhizobium fredii. *PLoS Genet*ics **14**:e1007428. doi:
939 10.1371/journal.pgen.1007428.
940 Johnson SC. 1967. Hierarchical clustering schemes. *Psychometrika*. **32**:241-254.
941 Kahng LS, Shapiro L. 2003. Polar localization of replicon origins in the multipartite
942 genomes of *Agrobacterium tumefaciens* and *Sinorhizobium meliloti*. *Journal of*
943 *Bacteriology*. **185**:3384-3391. doi: 10.1128/jb.185.11.3384-3391.2003.
944 Kanehisa M, Goto S, Sato Y, Furumichi M, Tanabe M. 2012. KEGG for integration and
945 interpretation of large-scale molecular data sets. *Nucleic Acids Research*.

946 **40**:D109-114. doi: 10.1093/nar/gkr988.

947 Kemter FS, Messerschmidt SJ, Schallopp N, Sobetzko P, Lang E, Bunk B, Sproer C,
948 Teschler JK, Yildiz FH, Overmann J, Waldminghaus T. 2018. Synchronous
949 termination of replication of the two chromosomes is an evolutionary selected
950 feature in Vibrionaceae. *PLOS Genetics*. **14**:e1007251. doi:
951 10.1371/journal.pgen.1007251.

952 Krawiec S, Riley M. 1990. Organization of the bacterial chromosome. *Microbiological*
953 *Reviews*. **54**:502-539.

954 Landeta C, Dávalos A, Cevallos MÁ, Geiger O, Brom S, Romero D. 2011. Plasmids
955 with a chromosome-like role in rhizobia. *Journal of Bacteriology*. **193**:1317-
956 1326. doi: 10.1128/JB.01184-10.

957 Lederberg J. 1998. Plasmid (1952-1997). *Plasmid*. **39**:1-9. doi: 10.1006/plas.1997.1320.

958 Leplae R, Lima-Mendez G, Toussaint A. 2010. ACLAME: a CLAssification of Mobile
959 genetic Elements, update 2010. *Nucleic Acids Research*. **38**:D57-61. doi:
960 10.1093/nar/gkp938.

961 Li Y, Canchaya C, Fang F, Raftis E, Ryan KA, van Pijkeren J-P, van Sinderen D,
962 O'Toole PW. 2007. Distribution of megaplasmids in *Lactobacillus salivarius* and
963 other lactobacilli. *Journal of Bacteriology*. **189**:6128-6139. doi:
964 10.1128/JB.00447-07.

965 Lioy VS, Cournac A, Marbouty M, Duigou S, Mozziconacci J, Espeli O, Boccard F,
966 Koszul R. 2018. Multiscale structuring of the *E. coli* chromosome by nucleoid-
967 associated and condensin proteins. *Cell*. **172**:771-783 e718. doi:
968 10.1016/j.cell.2017.12.027.

969 Liu G, Yong MY, Yurieva M, Srinivasan KG, Liu J, Lim JS, Poidinger M, Wright GD,
970 Zolezzi F, Choi H, Pavelka N, Rancati G. 2015. Gene essentiality is a

971 quantitative property linked to cellular evolvability. *Cell*. **163**:1388-1399. doi:
972 10.1016/j.cell.2015.10.069.
973 Livny J, Yamaichi Y, Waldor MK. 2007. Distribution of centromere-like parS sites in
974 bacteria: insights from comparative genomics. *Journal of Bacteriology*.
975 **189**:8693-8703. doi: 10.1128/JB.01239-07.
976 Lu C, Nakayasu ES, Zhang LQ, Luo ZQ. 2016. Identification of Fic-1 as an enzyme that
977 inhibits bacterial DNA replication by AMPylating GyrB, promoting filament
978 formation. *Science Signal*. **9**:ra11. doi: 10.1126/scisignal.aad0446.
979 Mackenzie C, Kaplan S, Choudhary M. 2004. Multiple chromosomes. *In: Miller RV and*
980 *Day MJ, editors. Microbial Evolution: Gene Establishment, Survival, and*
981 *Exchange: ASM press, Washington DC*:82-101.
982 Mackenzie C, Simmons AE, Kaplan S. 1999. Multiple chromosomes in bacteria: the Yin
983 and Yang of *trp* gene localization in *Rhodobacter sphaeroides* 2.4.1. *Genetics*.
984 **153**:525-538.
985 MacLellan SR, Sibley CD, Finan TM. 2004. Second chromosomes and megaplasmids in
986 bacteria. *In: Funnel BE and Phillips GJ, editors. Plasmid biology: ASM press,*
987 *Washington DC*:529-542.
988 MacLellan SR, Zaheer R, Sartor AL, MacLean AM, Finan TM. 2006. Identification of a
989 megaplasmid centromere reveals genetic structural diversity within the repABC
990 family of basic replicons. *Molecular Microbiology*. **59**:1559-1575. doi:
991 10.1111/j.1365-2958.2005.05040.x.
992 Maida I, Fondi M, Orlandini V, Emiliani G, Papaleo MC, Perrin E, Fani R. 2014.
993 Origin, duplication and reshuffling of plasmid genes: Insights from *Burkholderia*
994 *vietnamiensis* G4 genome. *Genomics*. **103**:229-238. doi:
995 10.1016/j.ygeno.2014.02.004.

996 McCullagh P, Nelder JA. 1989. Generalized linear models, Second Edition: Chapman &
997 Hall/CRC, London. 532 p.
998 Medema MH, Trefzer A, Kovalchuk A, van den Berg M, Muller U, Heijne W, Wu L,
999 Alam MT, Ronning CM, Nierman WC, Bovenberg RA, Breitling R, Takano E.
1000 2010. The sequence of a 1.8-mb bacterial linear plasmid reveals a rich
1001 evolutionary reservoir of secondary metabolic pathways. *Genome Biology and*
1002 *Evolution*. **2**:212-224. doi: 10.1093/gbe/evq013.
1003 Meier EL, Daitch AK, Yao Q, Bhargava A, Jensen GJ, Goley ED. 2017. FtsEX-
1004 mediated regulation of the final stages of cell division reveals morphogenetic
1005 plasticity in *Caulobacter crescentus*. *PLoS Genetics* **13**: e1006999. doi:
1006 10.1371/journal.pgen.1006999.
1007 Million-Weaver S, Camps M. 2014. Mechanisms of plasmid segregation: have
1008 multicopy plasmids been overlooked? *Plasmid*. **75**:27-36. doi:
1009 10.1016/j.plasmid.2014.07.002.
1010 Murray H, Errington J. 2008. Dynamic control of the DNA replication initiation protein
1011 DnaA by Soj/ParA. *Cell*. **135**:74-84. doi: 10.1016/j.cell.2008.07.044.
1012 Naito M, Ogura Y, Itoh T, Shoji M, Okamoto M, Hayashi T, Nakayama K. 2016. The
1013 complete genome sequencing of *Prevotella intermedia* strain OMA14 and a
1014 subsequent fine-scale, intra-species genomic comparison reveal an unusual
1015 amplification of conjugative and mobile transposons and identify a novel
1016 *Prevotella*-lineage-specific repeat. *DNA Research*. **23**:11-19. doi:
1017 10.1093/dnares/dsv032.
1018 Ohtani N, Tomita M, Itaya M. 2010. An extreme thermophile, *Thermus thermophilus*, is
1019 a polyploid bacterium. *Journal of Bacteriology*. **192**:5499-5505. doi:
1020 10.1128/JB.00662-10.

1021 Passot FM, Calderon V, Fichant G, Lane D, Pasta F. 2012. Centromere binding and
1022 evolution of chromosomal partition systems in the Burkholderiales. *Journal of*
1023 *Bacteriology*. **194**:3426-3436. doi: 10.1128/JB.00041-12.
1024 Pedregosa F, Weiss R, Brucher M. 2011. Scikit-learn : Machine Learning in Python.
1025 *Journal of Machine Learning Research*. **12**:2825-2830.
1026 http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf.
1027 Petersen J, Frank O, Göker M, Pradella S. 2013. Extrachromosomal, extraordinary and
1028 essential--the plasmids of the *Roseobacter* clade. *Applied Microbiology and*
1029 *Biotechnology*. **97**:2805-2815. doi: 10.1007/s00253-013-4746-8.
1030 Pinto UM, Pappas KM, Winans SC. 2012. The ABCs of plasmid replication and
1031 segregation. *Nature Review in Microbiology*. **10**:755-765. doi:
1032 10.1038/nrmicro2882.
1033 Pruitt KD, Tatusova T, Maglott DR. 2005. NCBI Reference Sequence (RefSeq): a
1034 curated non-redundant sequence database of genomes, transcripts and proteins.
1035 *Nucleic Acids Research*. **33**:D501-504. doi: 10.1093/nar/gki025.
1036 Rasmussen T, Jensen RB, Skovgaard O. 2007. The two chromosomes of *Vibrio cholerae*
1037 are initiated at different time points in the cell cycle. *EMBO Journal*. **26**:3124-
1038 3131. doi: 10.1038/sj.emboj.7601747.
1039 Reyes-Lamothe R, Tran T, Meas D, Lee L, Li AM, Sherratt DJ, Tolmasky ME. 2014.
1040 High-copy bacterial plasmids diffuse in the nucleoid-free space, replicate
1041 stochastically and are randomly partitioned at cell division. *Nucleic Acids*
1042 *Research*. **42**:1042-1051. doi: 10.1093/nar/gkt918.
1043 Rosenberg A, Hirschberg J. 2007. V-Measure: A conditional entropy-based external
1044 cluster evaluation measure. *Proceedings of the 2007 Joint Conference on*
1045 *Empirical Methods in Natural Language Processing and Computational Natural*

1046 *Language Learning (EMNLP-CoNLL)* p. 410-420.
1047 https://www.researchgate.net/publication/221012656_V-
1048 Measure_A_Conditional_Entropy-
1049 Based_External_Cluster_Evaluation_Measure.
1050 Rosvall M, Bergstrom CT. 2008. Maps of random walks on complex networks reveal
1051 community structure. *Proceedings of the National Academy of Sciences of the*
1052 *United States of America*. **105**:1118-1123. doi: 10.1073/pnas.0706851105.
1053 San Millan A, Peña-Miller R, Toll-Riera M, Halbert ZV, McLean aR, Cooper BS,
1054 MacLean RC. 2014. Positive selection and compensatory adaptation interact to
1055 stabilize non-transmissible plasmids. *Nature communications*. **5**:5208. doi:
1056 10.1038/ncomms6208.
1057 Slater SC, Goldman BS, Goodner B, Setubal JC, Farrand SK, Nester EW, Burr TJ,
1058 Banta L, Dickerman AW, Paulsen I, Otten L, Suen G, Welch R, Almeida NF,
1059 Arnold F, Burton OT, Du Z, Ewing A, Godsy E, Heisel S, Houmiel KL, Jhaveri
1060 J, Lu J, Miller NM, Norton S, Chen Q, Phoolcharoen W, Ohlin V, Ondrusek D,
1061 Pride N, Stricklin SL, Sun J, Wheeler C, Wilson L, Zhu H, Wood DW. 2009.
1062 Genome sequences of three *Agrobacterium* biovars help elucidate the evolution
1063 of multichromosome genomes in bacteria. *Journal of Bacteriology*. **191**:2501-
1064 2511. doi: 10.1128/JB.01779-08.
1065 Srivastava P, Fekete RA, Chattoraj DK. 2006. Segregation of the replication terminus of
1066 the two *Vibrio cholerae* chromosomes. *Journal of Bacteriology*. **188**:1060-1070.
1067 doi: 10.1128/JB.188.3.1060-1070.2006.
1068 Stalder T, Rogers LM, Renfrow C, Yano H, Smith Z, Top EM. 2017. Emerging patterns
1069 of plasmid-host coevolution that stabilize antibiotic resistance. *Scientific Reports*.
1070 **7**:4853. doi: 10.1038/s41598-017-04662-0.

1071 Stokke C, Waldminghaus T, Skarstad K. 2011. Replication patterns and organization of
1072 replication forks in *Vibrio cholerae*. *Microbiology*. **157**:695-708. doi:
1073 10.1099/mic.0.045112-0.
1074 Sýkora P. 1992. Macroevolution of plasmids: A model for plasmid speciation. *Journal*
1075 *of theoretical biology*. **159**:53-65. doi: 10.1016/s0022-5193(05)80767-2.
1076 Val M-E, Kennedy SP, El Karoui M, Bonné L, Chevalier F, Barre F-X. 2008. FtsK-
1077 dependent dimer resolution on multiple chromosomes in the pathogen *Vibrio*
1078 *cholerae*. *PLOS Genetics*. **4**:e1000201. doi: 10.1371/journal.pgen.1000201.
1079 Val M-E, Kennedy SP, Soler-Bistué AJ, Barbe V, Bouchier C, Ducos-Galand M,
1080 Skovgaard O, Mazel D. 2014. Fuse or die: how to survive the loss of Dam in
1081 *Vibrio cholerae*. *Molecular Microbiology*. **91**:665-678. doi: 10.1111/mmi.12483.
1082 Venkova-Canova T, Chattoraj DK. 2011. Transition from a plasmid to a chromosomal
1083 mode of replication entails additional regulators. *Proceedings of the National*
1084 *Academy of Sciences of the United States of America*. **108**:6199-6204. doi:
1085 10.1073/pnas.1013244108.
1086 Villaseñor T, Brom S, Davalos A, Lozano L, Romero D, Los Santos AG. 2011.
1087 Housekeeping genes essential for pantothenate biosynthesis are plasmid-encoded
1088 in *Rhizobium etli* and *Rhizobium leguminosarum*. *BMC microbiology*. **11**:66. doi:
1089 10.1186/1471-2180-11-66.
1090 Vuilleumier S, Chistoserdova L, Lee MC, Bringel F, Lajus A, Zhou Y, Gourion B,
1091 Barbe V, Chang J, Cruveiller S, Dossat C, Gillett W, Gruffaz C, Haugen E,
1092 Hourcade E, Levy R, Mangenot S, Muller E, Nadalig T, Pagni M, Penny C,
1093 Peyraud R, Robinson DG, Roche D, Rouy Z, Saenampechek C, Salvignol G,
1094 Vallenet D, Wu Z, Marx CJ, Vorholt JA, Olson MV, Kaul R, Weissenbach J,
1095 Medigue C, Lidstrom ME. 2009. *Methylobacterium* genome sequences: a

1096 reference blueprint to investigate microbial metabolism of C1 compounds from
1097 natural and industrial sources. *PLoS One*. **4**:e5584. doi:
1098 10.1371/journal.pone.0005584.
1099 Webber MA, Bailey AM, Blair JM, Morgan E, Stevens MP, Hinton JC, Ivens A, Wain
1100 J, Piddock LJ. 2009. The global consequence of disruption of the AcrAB-TolC
1101 efflux pump in *Salmonella enterica* includes reduced expression of SPI-1 and
1102 other attributes required to infect the host. *Journal of Bacteriology*. **191**:4276-
1103 4285. doi: 10.1128/JB.00363-09.
1104 Wisniewski-Dyé F, Borziak K, Khalsa-Moyers G, Alexandre G, Sukharnikov LO,
1105 Wuichet K, Hurst GB, McDonald WH, Robertson JS, Barbe V, Calteau A, Rouy
1106 Z, Mangenot S, Prigent-Combaret C, Normand P, Boyer M, Siguier P, Dessaux
1107 Y, Elmerich C, Condemine G, Krishnen G, Kennedy I, Paterson AH, González
1108 V, Mavingui P, Zhulin IB. 2011. *Azospirillum* genomes reveal transition of
1109 bacteria from aquatic to terrestrial environments. *PLOS Genetics*. **7**:e1002430.
1110 doi: 10.1371/journal.pgen.1002430.
1111 Wisniewski-Dyé F, Lozano L, Acosta-Cruz E, Borland S, Drogue B, Prigent-Combaret
1112 C, Rouy Z, Barbe V, Mendoza Herrera A, González V, Mavingui P. 2012.
1113 Genome sequence of *Azospirillum brasilense* CBG497 and comparative analyses
1114 of *Azospirillum* core and accessory genomes provide Insight into niche
1115 adaptation. *Genes (Basel)*. **3**:576-602. doi: 10.3390/genes3040576.
1116 Xie G, Johnson SL, Davenport KW, Rajavel M, Waldminghaus T, Detter JC, Chain PS,
1117 Sozhamannan S. 2017. Exception to the rule: genomic characterization of
1118 naturally occurring unusual *Vibrio cholerae* strains with a single chromosome.
1119 *International Journal of Genomics*. **2017**:8724304. doi: 10.1155/2017/8724304.
1120 Yamaichi Y, Fogel MA, McLeod SM, Hui MP, Waldor MK. 2007. Distinct centromere-

1121 like *parS* sites on the two chromosomes of *Vibrio* spp. *Journal of Bacteriology*.
1122 **189**:5314-5324. doi: 10.1128/JB.00416-07.
1123 Yamamoto S, Lee K-i, Morita M, Arakawa E, Izumiya H, Ohnishi M. 2018. Single
1124 circular chromosome identified from the genome sequence of the *Vibrio*
1125 *cholerae* O1 bv. El Tor Ogawa strain V060002. *Genome Announcements*. **6**. doi:
1126 10.1128/genomeA.00564-18.

1127 Yeoman CJ, Kelly WJ, Rakonjac J, Leahy SC, Altermann E, Attwood GT. 2011. The
1128 large episomes of *Butyrivibrio proteoclasticus* B316T have arisen through
1129 intragenomic gene shuttling from the chromosome to smaller *Butyrivibrio*-
1130 specific plasmids. *Plasmid*. **66**:67-78. doi: 10.1016/j.plasmid.2011.05.002.

1131 **SUPPLEMENTARY TABLES**

1132 Table 1. Replicon dataset

1133 Table 2. INFOMAP IS protein-based clustering solution of the 4928 replicons

1134 Table 3. PCA + WARD IS protein-based clustering solution of the 4928 replicons

1135 Table 4. PCA + WARD IS function-based clustering solution of the 4928 replicons