Neo-formation of chromosomes in bacteria

Olivier 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 Bioogie Évolutive, Université Claude Bernard – Lyon 1, 43 boulevard du 11 novembre 1918, 69622 Villeurbanne cedex, France

* **Present address:** Cancer Epidemiology Program, University of Hawaii Cancer Center, University of Hawaii at Manoa, 701 Ilalo Street Honolulu, HI 96813, USA

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

Summary

The genome of bacteria is classically separated into the essential, stable and slow evolving chromosomes and the accessory, mobile and rapidly evolving plasmids. This distinction was blurred with the characterization of multipartite genomes constituted of a "standard" chromosome and additional essential replicons integrated in the cell cycle. Although these ("megaplasmids", "secondary chromosomes" or "chromids") are considered to be adapted plasmids, their true nature and evolution are yet to be determined. We performed a global comparative genomic analysis using all available bacterial genomes using genes functionally related to the genomic information inheritance systems as parameters. We show that the bacterial replicons can be discriminated into chromosomes, plasmids, and another class of genomic elements that we tentatively name “neochromosomes” for their functioning as chromosomes and their coming into existence secondarily to the classical chromosome. Whereas some neochromosomes are clearly plasmidic in origin (most frequent case), a few might be deriving from the cleavage of an ancestral “true” chromosome.

Chromosomes are the only components of the genome, common to all members of a species, that encode the necessary information for replication and continued life of the cell/organism under normal growth conditions. Whereas their number varies in eukaryotes, a single, generally circular, chromosome has been the norm for bacterial genomes until recently[1]. Evidence that bacterial genomes can be distributed on more than one chromosome-like autonomously replicating genomic element (replicon) has accumulated for the past twenty-five years[2,3,4]. Several unique, essential, replicons (ERs) are constitutively present in, and required for the functioning of, the bacterial cell. Only the largest, primary, ER is expected to be a *bona fide* chromosome, *i.e.*, homologous to the single chromosome constituting the stable genome of most bacteria. The additional, secondary, ERs (SERs) are proposed to derive from accessory replicons (plasmids[5,6]) that were acquired by a mono-chromosome progenitor bacterium and stabilized through the transfer from the chromosome of genes essential to the cell viability[7,8]. The existence of plasmid-like replication and partition systems in SERs[7,9-14] as well as experimental results[15] support this view. Yet, the duplication and maintenance processes of SERs are in stark contrast to the common behaviour of plasmids for which both the timing of replication initiation and the centromere movement are random[16,17]. The SERs appear to share many characteristic features with chromosomes: enrichment in Dam methylation sites of the replication origin[9,18], presence of initiator titration sites[9,19], synchronization of the replication with the cell cycle[9,20-28], KOPS-guided FtsK translocation[29], FtsK-dependent dimer resolution system[29], MatP/matS macrodomain organisation system[30], and similar fine-scale segregation dynamics[22,28]. Furthermore, although initiated at different time points[22-28] using replicon-specific systems[7,9-11,31,32], the replications of the chromosome and the SER(s) within a multipartite genome are coordinated, hence maintaining the genome stoichiometry[21,22,25,26,28]. In the few species where this was studied, the replication of the SER is initiated after that of the chromosome[22-28]. In *V. cholerae* N16961, the SER duplication is licensed by the replication of a short region of the chromosome[33,34] and the divisome assembly is by the advancement of the SER replication and segregation[35]. In turn, in α-proteobacterium *Ensifer/Sinorhizobium meliloti*, the altering of the chromosome replication does not affect the replication initiation control of the SER[28]. Beside the exploration of the replication/segregation mechanistic, studies of multipartite genomes, targeting a single bacterial species or genus[3,7,8,10,15] or using a more extensive set of taxa[4,36], relied on debatable[4,36,37] (replicon size, nucleotide composition, coding of core essential genes for growth and survival; Extended data Fig. 1) and/or oriented[36] (presence of plasmid-type systems for genome maintenance and replication initiation) criteria to characterize the SER(s). While clarifying the functional and evolutionary contributions of each type of replicon to a multipartite genome in a given bacterial lineages[7,10,32,36], these studies did not produce an incontrovertible definition of SERs[4,36] or a universal model for their emergence[4,7,15,32,36]. We thus attempted to investigate the nature(s) and origin(s) of these replicons using as few assumptions as possible.

Replicon inheritance systems as diagnostic features

We reasoned that the key property discriminating chromosomes from plasmids is their vertical inheritance during the bacterial cell cycle, hence that the level of integration of a replicon into the host cycle relies on the nature of its inheritance systems (ISs), *i.e.*, the functions involved in its replication, partition and maintenance. We did not limit our study to multipartite genomes but performed a global analysis encompassing all bacterial replicons whose complete sequence was available in public sequence databases, and explored the structuration of the bacterial replicon space according to the IS usage of the replicons using a data mining and machine learning approach (Fig. 1). Since the inheritance of genetic information requires functionally diverse and heterogeneous actuators depending on the replicon type and the characteristics of the organism, we faced the challenge to identify as many as possible (possibly all) sequence orthologues corresponding to IS functional homologues whilst avoiding false positives (*e.g.*, sequence paralogues), bearing in mind that remote sequence homology most likely prevails among chromosome/plasmid protein-homologue pairs. From an initial dataset of 5125 replicons, 358,624 putative IS functional homologues, overall corresponding to 1711 Pfam functional domains (Extended data Fig. 2a), were retrieved using a query set of 47,604 chromosomal and plasmidic IS-related proteins selected from the KEGG and ACLAME databases and clustered into 7013 homology groups named after the most frequent annotation found among their proteins (Extended data Fig. 2b,c). Most clusters were characterized by a single annotation whilst the remaining few (4.7%) each harboured from 2 to 710 annotations, the most frequent annotation in a cluster generally representing more than half of all annotations (Extended data Fig. 2c). The removal of false positives left 267.497 IS protein homologues distributed in 6096 clusters (Extended data Fig. 2d) and coded by 4928 replicons (Supplementary data Table S1) out of the initial replicon dataset. Following the Genbank/RefSeq annotations, our final dataset comprised 2016 complete genome sets corresponding to 3592 replicons (2016 chromosomes, 129 SERs, and 1447 plasmids) and 1336 plasmid genomes, irregularly distributed across the bacterial phylogeny (Fig. 2a). Multi-ER genomes are observed in 5.0% of all represented bacterial genera, and constitute 5.7% of the complete genomes available at the time of study (Fig. 2b). They are yet to be observed in most bacterial phyla, possibly as a result of the poor representation of some lineages. They are merely incidental (0.2% in Firmicutes) or reach up to almost one third of the genomes (30.1% in β-Proteobacteria) depending on the lineage. Although found in ten phyla, they occur more than once per genus in only three of them, Bacteroidetes, Proteobacteria and Spirochaetae.

Exploration of the replicon space

The 4928 replicons were each ascribed a vector according to their IS usage profile, *i.e.*, the number of IS protein occurrences in each of the 6096 retained clusters, and transformed into bipartite graphs. These display both the vectors (replicons) and the variables (protein clusters) together with their respective connections, and support the

interactive mining of the data. The majority of the replicons are interconnected, as testimony of the shared evolutionary history of their ISs. Chromosomes and plasmids form overall distinct groups and communities with varying degree of connectivity depending on their functional specificities (Fig. 3a) as well as the bacterial taxonomy of their hosts (Fig. 3c). Thus, the occurrence of poorly IS cluster-connected plasmids within a group of chromosomes may rather result from shared connections to a very small number (as low as one) of common ISs than reflect a true relationship. The chromosomes and the plasmids nevertheless share numerous ISs, bearing witness to the continuity of the genomic material and the extensive exchange of genetic material within bacterial genomes. While being interconnected to both chromosomes and plasmids *via* numerous IS clusters, the SERs generally stand apart from either these types of replicons and gather at the chromosome-plasmid interface (Fig. 3a,b). Indeed, their IS usage is neither chromosome-like nor plasmid-like, suggesting that they may constitute a separate category of replicons. This is most tangible in the case of the proteobacterial lineages where SERs occur most frequently (Fig. 3b). All SERs in the β- and γ-Proteobacteria, and most in the α-Proteobacteria are linked to some remarkable chromosome-type IS clusters, among which AcrA, IciA, FtsE, HN-S and Lrp, as well to plasmid-like ParA/ParB-, Rep- and PSK-like IS clusters. A similar pattern is observed for the SERs in actinobacterium *Nocardiopsis dassonvillei,* firmicute *Butyrivibrio proteoclasticus*, and Chloroflexi *Sphaerobacter thermophilus* and *Thermobaculum terrenum* (Fig. 3b). Interestingly, the connection of the SERs present in the genomes of all but one *Burkholderia* species (β-Proteobacteria) to DnaG-annotated clusters that are linked to the chromosomes of all other bacteria is suggestive of a single origination of this SER prior to the diversification of the genus. The second SERs harboured by only some *Burkholderia* species exhibit a higher level of interconnection to plasmids, pointing to an incomplete stabilisation into the genome and a separate, more recent, integration event. The SERs in α-proteobacterium *Sphingobium*, cyanobacterium *Cyanothece*, and *Deinococcus radiodurans* (Deinococcus-Thermus) similarly localise closer to plasmids and are enriched in plasmid-like ISs, possibly reflecting the differing selective pressures at play depending on the bacterial lineages rather than a recent emergence. At odds with these observations, some SERs group unambiguously with chromosomes. The SERs in *Chloroacidobacterium thermophilum* (Acidobacteria), *Prevotella intermedia* and *P. melaninogenica* (Bacteroidetes), *Anabaena* sp. (Cyanobacteria), and α-Proteobacteria *Asticcacaulis excentricus* and *Paracoccus denitrificans* bear higher levels of interconnection to chromosomes than to plasmids or other SERs. The SERs in *Prevotella* spp. are in fact hardly linked to plasmids, and the few plasmid-like IS proteins that *C. thermophilum* SER codes (mostly Rep, Helicase and PSK) are observed in plasmids occurring in other phyla but in none of the Acidobacteria plasmids. An extreme situation is met in *Leptospira* spp. (Spirochaetae) whose SERs are each linked to only three or four (out of six) chromosome-like IS clusters, always including ParA and ParB. Interestingly, the ParA cluster appears to be specific to Spirochaetae chromosomes with the notable exception of one plasmid found in Leptospiraceae *Turneriella parva*.

IS-based relationships of the replicons

The bipartite graphs of the whole dataset or after genus-normalisation were submitted to a community structure detection algorithm (INFOMAP) that performs a random walk along the edges connecting the graph vertices. The underlying assumption is that the random walk will be trapped in high-density regions of the graph (*i.e.*, the community structure). We also performed a dimension reduction by Principal Component Analysis followed by a hierarchical clustering procedure (WARD). That latter was able to detect groups of replicons strongly connected to a small number of IS protein clusters and correlated to the host taxonomy, whereas the former performed better with replicons highly connected to clusters of IS proteins. The clustering solutions were meaningful (high values reached by the stability criterion scores), and biologically relevant (efficient separation of the chromosomes from the plasmids; high *homogeneity* values) using either method (Extended data Table 1). The genus-normalisation, aimed at controlling for the disparity in taxon representation of the replicons, produced overall similar albeit slightly less stable clusters (lower *homogeneity* values). Taxonomically homogeneous clusters of chromosomes were best retrieved using the coupling of dimension reduction and hierarchical clustering with a large enough number of clusters (*homogeneity* scores up to 0.93). In turn, the community detection algorithm was more efficient in recovering the underlying taxonomy of replicons (higher value of *completeness*), and was sole able to identify small and scattered plasmid clusters (Supplementary data Tables S6 and S7). The plasmid clusters obtained using PCA+WARD lacked taxonomical patterning and, albeit highly stable, only reflected the small Euclidian distances existing between the plasmid replicons (*e.g.*, one cluster of 2656 plasmids had a stability score of 0.975). Those resulting from INFOMAP probable walks on the graph mirrored the taxonomical structure of the data, suggesting that the taxonomic signal, chiefly expected from the chromosomes, is preserved among the protein families functionally specifying the plasmids. The grouping of a majority of SERs within INFOMAP-generated communities of chromosomes confirmed the sharing of characteristic features by these two types of replicons, hence the clear individuation of the SERs from the plasmids. In turn, the difference in numbers of IS clusters linked to either SERs or chromosomes often led to the placement of SERs into plasmid clusters by the PCA+WARD methodology. The SERs in *Butyrivibrio*, *Deinococcus*, *Leptospira* and *Rhodobacter* spp. grouped stably and consistently with plasmids while the SERs in *Vibrionaceae* and *Brucellaceae* formed specific clusters (Extended data Table 2). Burkholderiales and *Agrobacterium* SERs, whose homogenous clusters tended to be unstable, exhibited a higher affinity to plasmids overall. In turn, the SERs of *Asticaccaulis*, *Paracoccus* and *Prevotella* grouped stably with chromosomes using both clustering methods (Extended data Table 2a,b) and appear to possess IS profiles that set them apart from both the plasmids and the other SERs. Similar conclusions were reached when performing a PCA+WARD unsupervised classification using the 117 functional annotations of the IS protein clusters rather than the IS clusters *per se* (Extended data Tables 1 and 3). The SERs are thus determined by specific IS

functionalities that distinguish them from either chromosomes or plasmids. They are nevertheless heterogeneous insofar as they present varying distance from the other types of replicons, some showing affinity to plasmids and others to chromosomes, with the singularity of the SERs found in *Asticaccaulis*, *Paracoccus* and *Prevotella* being confirmed. The *Prevotella* multipartite genomes (*P. intermedia* and *P. melaninogenica*) are further remarkable, their chromosomes being closer to plasmids than those of monopartite genomes in this genus (*P. denticola* and *P. ruminicola*).

SER-specifying IS functions

Next, we searched which IS functions could specify the SERs with respect to chromosomes and plasmids. Logistic regression analyses were used to identify over- or underrepresented ISs and to assess their respective relevance to each class of replicons. Because of their comparatively small number, all SERs were assembled into a single group despite their disparity. A hundred and one IS functionalities (96% of KEGG-annotated chromosome-like functions and 72% of ACLAME-annotated plasmid-like functions) were significantly enriched in one replicon category over the other (Table 1). The large majority of the IS functions differentiates the chromosomes from the plasmids. These are only determined by ISs corresponding to ACLAME annotations Rep, Rop and TrfA, involved in initiation of plasmid replication, and ParA and ParB, dedicated to plasmid partition. Some KEGG-annotated functions (*e.g.,* DnaA, DnaB or FtsZ,) appear to be more highly specific to chromosomes than others, such as DnaC, FtsE or H-NS (lower *OR* values). Numerous functions specifying the chromosomes over the plasmids discriminate only poorly, or not at all, the SERs from the chromosomes. Furthermore, the IS distribution bias toward chromosomes is weaker overall in the comparison to SERs than in that to plasmids. Chromosome-signature ISs such as DnaA, DnaB, DnaG and FtsZ are also present on the SERs, and a few of them are enriched to the same order of magnitude in both these classes of replicons when compared to plasmids (signalled by stars in Table 1). The main divergence between SERs and chromosomes lies in the distribution patterns of the ACLAME-annotated ISs, the SER biases toward plasmid-type ISs suggesting a stronger link to plasmids. This pattern, rather than constitute a global trend shared by all multipartite genomes and a diagnostic feature of this genome architecture, may arise from the unbalanced taxon representation in our SER dataset and reflect the adaptation of a single bacterial lineage (*e.g.*, RepC in Rhizobiales SERs[38]).

Identification of candidate (S)ERs

Since the IS profiles constitute replicon-type signatures, we tried and identified putative SERs or chromosomes among the extra-chromosomal replicons, using IS function-based supervised classifications and various training sets (Extended Data Table 4). The discriminative power of the ISs in these procedures was ascertained by the overall high probabilities for SERs to be classified in their own class (Extended Data Tables 4 and 5). The relatively low average probabilities observed for a few bacterial genera (Extended Data Table 5) likely arose from the structural scarcity of

the ISs carried by their SERs (*e.g., Leptospira*) or the initial discarding of lineage-specific ISs (*e.g.,* RtcB, the idiosyncratic replication initiator of Vibrionaceae SERs). We identified a number of candidate SERs among the plasmids (Table 2a), most of which are expected to be essential to the cell functioning and/or to the fitness of the organism (*cf.* Supplementary data). Whereas most belong to bacterial lineages known to host multipartite genomes, novel taxa emerge as putative host to complex genomes (Rhodospirillales and, to a lesser extent, Actinomycetales). In contrast, only one SER among those proposed for members of the *Roseobacter* clade (Rhodobacterales)[39] is confirmed by our analyses (*Ruegeria* sp. TM1040). Our analyses also revealed eight candidate chromosomes: two plasmids (also identified as candidate SERs) that encode ISs hardly found in extra-chromosomal elements (*e.g.*, DnaG, DnaB, ParC and ParE), and six SERs that part of or all our analyses associate to standard chromosomes (Table 2b). Notably, *P. intermedia* SER shows a very high probability (> 0.98) to be a chromosome and its annotated chromosome, unique of its kind, falls within the plasmid class.

To be, or not, a chromosome

Regardless of their host lineage, SERs exhibit a distributional bias of IS-related genes that specifies them functionally and enables their identification. They clearly stand apart from plasmids, be these occurring consistently in a bacterial species (*e.g.*, *Lactobacillus salivarius* pMP118-like plasmids[40]). In contrast to plasmid that are stabilized by compensatory adaption alleviating the fitness cost incurred by their presence in the cell[41-43], they encode chromosome-type proteins involved in the partition and segregation systems and the cell division as well as proteins acting on the DNA structure and cell cycle regulators, all functions that are hardly found on plasmids. The SERs thence qualify as essential replicons irrespective of the eco-phenotypic, possibly essential, functions that they encode (varying across host taxa) and of their size. The replicon size that was proposed as primary classification criterion by some authors[4,36] proves inoperative. The IS profiles did not discriminate the megaplasmids (>350 kb[4]) from smaller plasmids, and accurately identified *Leptospira* and *Butyrivibrio* SERs despite their plasmid-like size. Neither the small size of chromosomes (down to 139 kb) in the reduced genomes of endosymbionts precluded their assignment to the chromosome class, nor did the 1.12 Mb length of *Rhodococcus jostii* RHA1 pRHL1 to the plasmid class. In turn, chromosome-like replication initiators are rarely found on SERs. When they are (*Paracoccus denitrificans*, *Prevotella intermedia* and *P. melaninogenica*), the chromosome in the corresponding genome does not code for one. Similarly, *bona fide* chromosomal centromeres (*parS*) are found on a single replicon in a multipartite genome, which is the chromosome in all but one genome: in *P. intermedia* (GCA_000261025.1), both replication initiation and partition systems define the secondary chromosome (CP003503) as the *bona* fide chromosome and the annotated chromosome (CP003502) as the SER.

Regarding the replication initiation and the partition, chromosomes and plasmids alike.

Whereas multiple copies of the chromosome are known to cohabit constitutively in polyploid bacteria (*e.g.*, *Thermus thermophilus*[44]), different replicons bearing identical chromosomal-like replication initiation and partition systems are incompatible, which implies that SERs do not derive from a whole duplication of the chromosome. The presence in SERs of plasmid-type replication initiation systems signs the involvement of plasmids in the coming into being of these replicons. Yet, the existence of several types of SERs characterized by specific features, and the extreme similitude between chromosomes and SERs within a few genomes provides evidence that processes other than the capture and amelioration of megaplasmids[7,8,10,13,15,36] are at play in the stabilization of additional essential replicons and the formation of multipartite genomes. Indeed, the presence on *Asticaccaulis, Paracoccus* and *Prevotella* SERs of high numbers of IS genes found on chromosomes in closely related species, and their complementarity between the SER and chromosome (Extended Data Table 6), paralleled by their whole gene contents[45,46], are best explained by postulating a chromosome splitting event. A novel replicon thus formed could only be sustainable if it acquired compatible replication-partition systems somehow.

The sum of these observations can be reconciled into a single model of emergence implicating at least one cointegration step of one (or several) plasmid(s) and the progenitor chromosome. Plasmids are known to form cointegrates and to integrate chromosomes in both experimental settings[46,47] and the natural environment[48,49], and to disintegrate into novel genomic architectures[48,49] with plasmid acquiring large fragment of the chromosome[48]. The replicon-cointegrate configuration would thus facilitate the exchange of genes and the translocation of replicon fragments and, through several cell division and possibly several rounds of integration-disintegration, provide time for the stabilisation process to take place, finally resulting in the separation of the shuffled cointegrate into multiple viable essential replicons forming a divided genome. They will differ from the chromosome and plasmid(s) that they were at the start, one replicon retaining the chromosome-like origin of replication and centrosome, and the other(s) the plasmidic counterparts. Depending on the level of reorganisation, the number of generations or cycle spent as cointegrate, and the advent of genetic material from other sources, the final replicons will show varying level of genome integration. Such a fusion-fission model of genome evolution can account for the variety of stabilization modalities seen in existing SERs and the extreme plasticity met in multipartite genomes. Having a divided genome appears to extend and accelerate the exploration of the genome evolutionary landscape, producing complex expression regulation and leading to novel eco-phenotypes and species diversification (*e.g.*, Burkholderiaceae and Vibrionaceae).

Irrespective of the mode of evolution, bacteria repeatedly evolve divided genome constituted of multiple essential replicons. One most resembles the classical bacterial chromosome while the other(s) show plasmid features. They do not strictly correspond to the ancestral replicons that gathered in their progenitor host, and rather constitute neo-chromosomes.

References

1. Krawiec, S. & Riley, M. Organization of the bacterial chromosome. *Microbiol. Rev.* 54, 502-39 (1990).
2. Casjens, S. The diverse and dynamic structure of bacterial genomes. *Annu. Rev. Genet.* **32,** 339-77 (1998).
3. Mackenzie, C., Kaplan, S., & Choudhary, M. Multiple chromosomes. *In* Microbial evolution: gene establishment, survival, and exchange. Pp. 82-101 (2004).
4. diCenzo, G. C. & Finan, T. M. The Divided Bacterial Genome: Structure, Function, and Evolution. *Microbiol Mol Biol Rev* **81**, doi:10.1128/MMBR.00019-17 (2017).
5. Lederberg J. Plasmid (1952-1997). *Plasmid* **39,** 1-9 (1998).
6. Smillie, C., Garcillán-Barcia, M. P., Victoria Francia, M., Rocha, E. P. C., & de la Cruz, F. Mobility of plasmids. *Microbiol. Mol. Biol. Rev.* **74,** 434-52 (2010).
7. Slater, S. C. *et al.* Genome sequences of three *Agrobacterium* biovars help elucidate the evolution of multichromosome genomes in bacteria. *Journal of bacteriology***191,** 2501–2511 (2009).
8. diCenzo, G., Milunovic, B., Cheng, J. & Finan, T. The tRNAarg gene and *engA* are essential genes on the 1.7-Mb pSymB megaplasmid of *Sinorhizobium meliloti* and were translocated together from the chromosome in an ancestral strain. *Journal of Bacteriology* **195,** 202–212 (2013).
9. Egan, E. S. & Waldor, M. K. Distinct replication requirements for the two *Vibrio cholerae* chromosomes. *Cell* **114,** 521–30 (2003).
10. MacLellan, S. R., Sibley, C. D., & Finan T. M. Second chromosomes and megaplasmids in bacteria, p 529–542. In Funnell BE, Phillips GJ (ed), Plasmid biology. ASM Press, Washington, DC. (2004).
11. MacLellan, S. R., Zaheer, R., Sartor, A. L., MacLean, A. M. & Finan, T. M. Identification of a megaplasmid centromere reveals genetic structural diversity within the repABC family of basic replicons. *Mol. Microbiol.* **59,** 1559–75 (2006).
12. Dubarry, N., Pasta, F. & Lane, D. ParABS systems of the four replicons of *Burkholderia cenocepacia*: new chromosome centromeres confer partition specificity. *J. Bacteriol.* **188,** 1489–96 (2006).
13. Livny, J., Yamaichi, Y. & Waldor, M. K. Distribution of centromere-like *parS* sites in bacteria: insights from comparative genomics. *J. Bacteriol.* **189,** 8693–703 (2007).
14. Yamaichi, Y., Fogel, M. A., McLeod, S. M., Hui, M. P. & Waldor, M. K. Distinct centromere-like parS sites on the two chromosomes of *Vibrio* spp. *J. Bacteriol.***189,** 5314–24 (2007).
15. diCenzo, G. C., MacLean, A. M., Milunovic, B., Golding, G. B. & Finan, T. M. Examination of prokaryotic multipartite genome evolution through experimental genome reduction. *PLoS Genet.* **10,** e1004742 (2014).

16. Million-Weaver, S. & Camps, M. Mechanisms of plasmid segregation: have multicopy plasmids been overlooked? *Plasmid* **75,** 27–36 (2014).
17. Reyes-Lamothe, R. *et al.* High-copy bacterial plasmids diffuse in the nucleoid-free space, replicate stochastically and are randomly partitioned at cell division. *Nucleic Acids Res* **42,** 1042–1051 (2014).
18. Gerding, M., Chao, M., Davis, B. & Waldor, M. Molecular dissection of the essential features of the origin of replication of the second *Vibrio cholerae* chromosome. *mBio* **6,** e00973–15 (2015).
19. Venkova-Canova, T. & Chattoraj, D. K. Transition from a plasmid to a chromosomal mode of replication entails additional regulators. *Proc. Natl. Acad. Sci. U.S.A.* **108,** 6199–204 (2011).
20. Kahng, L. S. & Shapiro, L. Polar localization of replicon origins in the multipartite genomes of *Agrobacterium tumefaciens* and *Sinorhizobium meliloti*. *Journal of bacteriology* **185,** 3384–91 (2003).
21. Egan, E. S., Løbner-Olesen, A. & Waldor, M. K. Synchronous replication initiation of the two *Vibrio cholerae* chromosomes.*Curr. Biol.* **14,** R501–2 (2004).
22. Fiebig, A., Keren, K. & Theriot, J. A. Fine-scale time-lapse analysis of the biphasic, dynamic behaviour of the two *Vibrio cholerae* chromosomes. *Mol. Microbiol.* **60,** 1164–78 (2006).
23. Srivastava, P., Fekete, R. A. & Chattoraj, D. K. Segregation of the replication terminus of the two *Vibrio cholerae* chromosomes. *J. Bacteriol.* **188,** 1060–70 (2006).
24. Rasmussen, T., Jensen, R. B. & Skovgaard, O. The two chromosomes of *Vibrio cholerae* are initiated at different time points in the cell cycle. *EMBO J.* **26,** 3124–31 (2007).
25. Stokke, C., Waldminghaus, T. & Skarstad, K. Replication patterns and organization of replication forks in *Vibrio cholerae*. *Microbiology (Reading, Engl.)* **157,** 695–708 (2011).
26. Deghelt, M. *et al.* G1-arrested newborn cells are the predominant infectious form of the pathogen *Brucella abortus*. *Nat Commun* **5,** 4366 (2014).
27. De Nisco, N. J., Abo, R. P., Wu, C. M., Penterman, J. & Walker, G. C. Global analysis of cell cycle gene expression of the legume symbiont *Sinorhizobium meliloti*. *Proc. Natl. Acad. Sci. U.S.A.* **111,** 3217–24 (2014).
28. Frage, B. *et al.* Spatiotemporal choreography of chromosome and megaplasmids in the *Sinorhizobium meliloti* cell cycle. *Mol Microbiol* **100,** 808–823 (2016).
29. Val, M.-E. *et al.* FtsK-dependent dimer resolution on multiple chromosomes in the pathogen *Vibrio cholerae*. *PLoS Genet.* **4,** e1000201 (2008).
30. Demarre, G. *et al.* Differential management of the replication terminus regions of the two *Vibrio cholerae* chromosomes during cell division. *PLoS Genet.* **10,** e1004557 (2014).
31. Drevinek, P., Baldwin, A., Dowson, C. G. & Mahenthiralingam, E. Diversity of the *parB* and *repA* genes of the *Burkholderia cepacia* complex and their

utility for rapid identification of *Burkholderia cenocepacia*. *BMC Microbiol.* **8,** 44 (2008).
32. Galardini, M., Pini, F., Bazzicalupo, M., Biondi, E. G. & Mengoni, A. Replicon-dependent bacterial genome evolution: The case of *Sinorhizobium meliloti*. *Genome Biol Evol* **5,** 542–558 (2013).
33. Baek, J. H. & Chattoraj, D. K. Chromosome I controls chromosome II replication in *Vibrio cholerae*. *PLoS Genet.* **10,** e1004184 (2014).
34. Val, M.-E. E. *et al.* A checkpoint control orchestrates the replication of the two chromosomes of *Vibrio cholerae*. *Sci Adv* **2,** e1501914 (2016).
35. Galli, E. *et al.* Cell division licensing in the multi-chromosomal *Vibrio cholerae* bacterium. *Nat Microbiol* **1**, 16094, doi:10.1038/nmicrobiol.2016.94 (2016).
36. Harrison, P., Lower, R., Kim, N. & Young, J. W. P. Introducing the bacterial 'chromid': not a chromosome, not a plasmid. *Trends Microbiol.* **18,** 141–8 (2010).
37. Liu, G. *et al.* Gene Essentiality Is a Quantitative Property Linked to Cellular Evolvability. *Cell* **163,** 1388–99 (2015).
38. Pinto, U. M., Pappas, K. M. & Winans, S. C. The ABCs of plasmid replication and segregation. *Nature reviews. Microbiology* **10**, 755-765, doi:10.1038/nrmicro2882 (2012).
39. Petersen, J., Frank, O., Göker, M. & Pradella, S. Extrachromosomal, extraordinary and essential-the plasmids of the *Roseobacter* clade. *Applied Microbiology and Biotechnology* **97**, 2805-2815, doi:10.1007/s00253-013-4746-8 (2013).
40. Li, Y. *et al*. Distribution of megaplasmids in *Lactobacillus salivarius* and other lactobacilli. *J. Bacteriol*. **189**, 6128-6139, doi:10.1128/JB.00447-07 (2007).
41. San Millan, A. *et al.* Positive selection and compensatory adaptation interact to stabilize non-transmissible plasmids. *Nature communications* **5**, 5208, doi:10.1038/ncomms6208 (2014).
42. Hall, J. P. J., Brockhurst, M. A., Dytham, C. & Harrison, E. The evolution of plasmid stability: Are infectious transmission and compensatory evolution competing evolutionary trajectories? *Plasmid* **91**, 90-95, doi:10.1016/j.plasmid.2017.04.003 (2017).
43. Stalder, T. *et al.* Emerging patterns of plasmid-host coevolution that stabilize antibiotic resistance. *Sci Rep* **7**, 4853, doi:10.1038/s41598-017-04662-0 (2017).
44. Ohtani, N., Tomita, M. & Itaya, M. An extreme thermophile, *Thermus thermophilus*, is a polyploid bacterium. *J Bacteriol* **192**, 5499-5505, doi:10.1128/JB.00662-10 (2010).
45. Poirion, O. Synteny plots in Chapter 8 in *Discrimination analytique des génomes bactériens* PhD thesis, Université de Lyon, 157-214, https://tel.archives-ouvertes.fr/tel-01174161 (2014).
46. Naito, M. *et al*. The complete genome sequencing of *Prevotella intermedia* strain OMA14 and a subsequent fine-scale, intra-species genomic comparison reveal an unusual amplification of conjugative and mobile transposons and identify a novel *Prevotella*-lineage-specific repeat. *DNA Res* **23**, 11-19,

doi:10.1093/dnares/dsv032 (2016).
47. Guo, X. *et al.* Natural genomic design in *Sinorhizobium meliloti*: novel genomic architectures. *Genome Res* **13**, 1810-1817, doi:10.1101/gr.1260903 (2003).
48. Val, M.-E. *et al.* Fuse or die: how to survive the loss of Dam in *Vibrio cholerae*. *Molecular microbiology* **91**, 665-678, doi:10.1111/mmi.12483 (2014).
49. Cervantes, L. *et al.* The conjugative plasmid of a bean-nodulating *Sinorhizobium fredii* strain is assembled from sequences of two *Rhizobium* plasmids and the chromosome of a *Sinorhizobium strain*. *BMC microbiology* **11**, 149, doi:10.1186/1471-2180-11-149 (2011).
50. Lovell H.C. et al. *In planta* conditions induce genomic changes in *Pseudomonas syringae* pv. *phaseolicola. Mol Plant Pathol* 12, 167-76 ,doi: 10.1111/j.1364-3703.2010.00658.x (2011).

Methods

To understand the pattern of relationships between the chromosomal and plasmidic replicons in bacteria, we focused on the distribution of IS-related genes for each replicon and built networks linking the replicons given their IS functional orthologues (Extended data Fig. 2).

Retrieval of IS functional homologues

A sample of proteins involved in the replication and segregation of bacterial replicons and of the bacterial cell cycle was constructed using datasets available from the ACLAME[51] and the KEGG[52] databases. Gene ontologies for "replication", "partition", "dimer resolution", and "genome maintenance" (Supplementary data Table S2) were used to select related ACLAME plasmid protein families (Supplementary data Table S3) using a semi-automated procedure. KEGG orthology groups were selected following the KEGG BRITE hierarchical classification (Supplementary data Table S4). The proteins belonging to the relevant 92 ACLAME protein families and 71 KEGG orthology groups (3,847 and 43,757 proteins, respectively) were retrieved and pooled. Using this query set amounting to a total of 47,604 proteins, a *blastp* search of the 6,903,452 protein sequences available from the 5,125 complete sequences of bacterial replicons available in the NCBI Reference Sequence database (RefSeq)[53] on 30/11/2012 identified 358,624 putative homologues using BLAST default parameters[54] and a 10^{-5} significance cut-off value. This *E*-value threshold was chosen to enable the capture of similarities between chromosome and plasmid proteins whilst minimizing the production of false positives, *i.e.*, proteins in a given cluster exhibiting small *E*-values despite not being functionally homologous. Using RefSeq ensured the annotation consistency of the genomes included in our study.

Clustering of IS functional homologues

Using this dataset, clusters of IS functional homologues were computed through the coupling of an *all-versus-all blastp* search using a 10^{-2} *E*-value threshold, followed by a TRIBE-MCL[55] clustering procedure based on a matrix of the *E*-values obtained from the comparisons of all possible protein pairs $d(p_i, p_j) = -\log_{10}\left(e_{value}(p_i, p_j)\right)$. Using a granularity value of 4.0 (see below),

the 358,624 IS homologues were organized into 7013 clusters, each comprising from a single to 1990 proteins (Extended data Fig. 3A). IS homologues were annotated according to their best match (BLAST hit with the lowest *E*-value) among the proteins of the query set, *i.e.*, according to one of the 117 functions of the query set (71 from KEGG and 46 from ACLAME). The clusters of functional homologues were then named after the annotation most frequently occurring among the proteins in the cluster. The number of protein annotations in a cluster determined the cluster quality, a single annotation being optimal. To select the best granularity and to estimate the consistency of the clusters in terms of functional homologues, the weighted Biological Homogeneity Index (*wBHI*, modified from the *BHI*[56], each cluster being weighted by its size) and the Conservation Consistency Measure (*CCM*, similar to the *BHI* but using the functional domains of the proteins to define the reference classes), which both take into account the size distribution of the clusters, were computed (*cf.* Supplementary Data for details on index calculation). The former gives an estimation of the overall consistency of clusters annotations according to the protein annotations whereas the latter gives an estimation of cluster homogeneity according to the protein domains identified beforehand. To build the sets of functional domains, an *hmmscan*[57] procedure against the Pfam database[58] was applied to each of the 358,624 putative IS homologues. Each protein was annotated according to the domain match(es) with *E*-value $< 10^{-5}$ (individual *E*-value of the domain) and *c-E*-value $< 10^{-5}$ (conditional *E*-value that measures the statistical significance of each domain). If two domains overlapped, only the domain exhibiting the smallest *E*-value was considered. *wBHI* and *CCM* indices were estimated for the clustering of the IS homologues and compared with values obtained for random clusters simulated according to the cluster size distribution of the IS proteins, irrespective of their length or function. For each of the clustering obtained for different granularities, a random clustering was constructed following the original cluster size distribution (assessed with a χ^2 test) and composed with simulated proteins according to the distributions of the type and number of functional domains of the data collected from the 358,624 IS homologues. Overall, the clusters obtained using a granularity of 4.0 with the TRIBE-MCL algorithm appeared to be homogenous in terms of proteins similarities toward their best BLAST hits and their functional domain distributions (Extended data Fig. 4).

Assessment of the homogeneity of IS functional homologues

The homogeneity toward the functions of the proteins in the query set relied on the assumption that the first BLAST cut-off (10^{-5} *E*-value) was stringent enough to capture only functional homologues to the query proteins. Potential bias might nevertheless arise from query proteins possessing a supplementary functional domain unrelated to the IS role, or from the selection of proteins belonging to the same superfamily but differing in function. The functional vectors associated to each KEGG group or ACLAME family of the query set, as well as those for all obtained clusters were calculated. For a protein P_i, the associated functional vector was defined with respect to its set of identified domains D_{P_i} and to the set of all identified domains $D=\{d_1,...,d_X\}$ as:

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

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 cluster of proteins C_i could then be defined as:

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

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

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

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

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

The $d_{cosine}\left(v_{C_a}, v_{C_0}\right)$ values were compared with those obtained using random clusters C_r of the same size than C_0. For each C_0 and its corresponding C_a, 200 random clusters and their associated distances $d_{cosine}\left(v_{C_a}, v_{C_r}\right)$, from which the corresponding empirical distribution D_e was constructed, were computed. C_0 is then considered as noise if $d_{cosine}\left(v_{C_a}, v_{C_0}\right) \notin Q_{10\%}^{D_e}$ where $Q_{10\%}^{D_e}$ is the 0.1-quantile of D_e.

Unsupervised analyses of the replicon space

Bacterial replicons (Supplementary data Table S1) were coded as vectors according to their content in IS genes. The number of IS protein clusters retained for the analysis determined the vector dimension and the number of proteins in a replicon assigned to each cluster gave the value of each vector component. Matrices $P = \begin{bmatrix} p_{1,1} & \cdots & p_{1,m} \\ \vdots & \ddots & \vdots \\ p_{n,1} & \cdots & p_{n,m} \end{bmatrix}$ were built, where n is the number of replicons, m the number of protein clusters, and $p_{i,j}$ the number of proteins of the jth cluster encoded by a gene present on the ith replicon. Several datasets were constructed to explore both the replicon type and the host taxonomy effects on the separation of the replicons in the performed analyses (Supplementary Data Table S5). The taxonomic representation bias was taken into account by normalizing the data with regard to the host genus: a consensus vector was built for each bacterial genus present in the datasets. The value of each vector attribute was calculated as the mean of the attribute values in the replicon vectors that belong to the same genus.

As a first approach, the data were transformed into bipartite graphs whose vertices are the replicons and the proteins clusters, and the graphs were spatialized using the force-directed layout algorithm ForceAtlas2[59] implemented in Gephi[60]. Bipartite graphs are a powerful way of representing the data by naturally drawing the links between the replicons while enabling the detailed analysis of the IS cluster-based connections of each replicon by applying forces to each node with regard to its connecting edges. To investigate further the IS-based relationships of the replicons, the community structure detection algorithm INFOMAP[61] was applied to the bipartite graph using the *igraph* python library[62]. A WARD hierarchical clustering[63] was also performed after a dimension reduction of the data using a Principal Component Analysis[64]. Selection of an optimal number of principal components to use in the second clustering procedure relied on the comparing of the cluster stabilities, given by Hennig *stability criterion*[65], estimated using variable number of principal components. The first 30 principal components (57% of total variance) were thus retained. For consistency purpose, the number of clusters in the WARD analysis was chosen to match that obtained with the INFOMAP procedure. The used number of clusters was assessed by the stability index by Fang & Wang[66] (Extended Data Table 1). The quality of the projection and clustering results were assessed using the V-measure

indices[67] (*homogeneity*, *completeness*, *V-measure*) as external cluster evaluation measures (Extended Data Table 1). The *homogeneity* indicates how uniform clusters are towards a class of reference. The *completeness* indicates whether reference classes are embedded within clusters. The *V-measure* is the harmonic mean between these two indices and indicates the quality of a clustering solution relative to the classes of reference. These three indices vary between 0 and 1, with values closest to 1 reflecting the good quality of the clustering solution. The type of replicons (*i.e.,* plasmid and chromosome) and the taxonomic affiliation (phylum or class) for chromosomes or plasmids were used as references classes (Supplementary Data Table S5). Additionally, the *stability criterion*[65] of individual clusters, weighted by their size, for a given clustering result was evaluated using the bootstrapping of the original dataset as re-sampling scheme. Individual Jaccard coefficient for each replicon were computed as the number of times that a given replicon of a cluster in a clustering solution is also present in the closest cluster in the resampled datasets.

Functional characterization of the replicons and genomes

In order to characterize the functional bias of the replicons, 117 IS functionalities (46 from ACLAME and 71 from KEGG) were considered. When equivalent in plasmids and chromosomes, functions from ACLAME and KEGG databases were considered to be 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 number of IS functionalities, was used as input to the projection algorithms. $f_{i,j}$ represents the number of times that genes coding for proteins annotated with the jth function are present on the ith replicon. Several datasets were analysed using PCA dimension reduction of the data followed by WARD hierarchical clustering (Extended Data Table 1).

Logistic regression analyses

Several reference classes of replicons and complete genomes were considered for comparison (Supplementary data Table S8). Ambiguous, *i.e.*, potentially adapted, plasmids belonging to INFOMAP clusters of plasmid replicons partially composed of SERs and/or chromosomes were removed from the plasmid class. When appropriate, the taxonomic representation bias was taken into account by normalizing the data with regard to the host genus as before. Logistic regressions[68] were performed for the 117

IS functions using the R glm package coupled to the python binder rpy2. The computed P_{value} measured the probability of a functionality to be predictive of a given group of replicons/genomes and the *Odd-Ratio* estimated how the functionality occurrence influenced the belonging of a replicon/genome to a given group.

Supervised classification of replicons and genomes

In order to identify putative ill-defined SERs and chromosomes amongst plasmids, we performed supervised classification analyses using random forest procedures[69]. We used the IS functionalities as the set of features and the whole sets of chromosomes, plasmids and SER as sets of samples to build four classification studies (Extended Data Table 4) and detect SER candidates (plasmids *vs.* SERs) and chromosome candidates (chromosomes *vs.* SERs or chromosomes *vs.* plasmids). Because of the unbalanced size of the training classes (SERs *vs.* chromosomes and plasmids), iterative sampling procedures were performed using 1000 random subsets of the largest class, with a size similar to that of the smallest class. The ensuing results were averaged to build the class probabilities and relative importance of the variables. We also used the whole set of plasmids when compared to SERs, to identify more robust SER candidates. The discarding of plasmids in the iterative procedure increases the classifier sensitivity while reducing the rate of false negatives by including more plasmid-annotated putative true SERs, whereas it decreases the classifier precision while increasing the rate of false positives. The ExtraTreeClassifier (a classifier similar to Random Forest) class from the Scikit-learn python library[70] was used to perform the classifications, with *K=1000*, *max_feat=sqrt(number of variables)* and *min_split=1*. For each run, the *feature_importances* and *estimate_proba* functions were used to compute, respectively, the relative contribution of the input variables and the class probabilities of replicons/genomes. The statistical probability of a replicon/genome belonging to a class was calculated as the average predicted class of the trees in the forest. The relative contribution of the input variables was estimated according to Breiman[71]. The choices of the number of trees in the forest *K*, the number of variables selected for each split *max_feat*, and the minimum number of samples required to split an internal node *min_split* were cross-validated using a *Leave-One-Out* scheme. The performance of the *Extremely-randomized-trees* classification procedures was assessed using a stratified 10-fold cross-validation

procedure following Han et al. [72], and the out-of-bag estimate (OOB score)[70,73] computed using the *oob_score* function of Scikit-learn python library.

References

51. Leplae, R., Lima-Mendez, G. & Toussaint, A. ACLAME: A CLAssification of Mobile genetic Elements, update 2010. *Nucleic Acids Res* **38,** D57–D61 (2010).
52. Kanehisa, M., Goto, S., Sato, Y., Furumichi, M. & Tanabe, M. KEGG for integration and interpretation of large-scale molecular data sets. *Nucleic Acids Res* **40,** D109–D114 (2012).
53. Pruitt, K. D., Tatusova, T. & Maglott, D. R. NCBI reference sequences (RefSeq) : a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res* **35,** D61–D65 (2007).
54. Camacho, C. *et al.* BLAST+: architecture and applications. *BMC Bioinformatics* **10**, 421 (2009).
55. Enright, A., Dongen, S. & Ouzounis, C. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res* **30,** 1575–84 (2002).
56. Datta, S. & Datta, S. Methods for evaluating clustering algorithms for gene expression data using a reference set of functional classes. *BMC Bioinformatics* **7,** 1–9 (2006).
57. Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res.* **39,** W29–37 (2011).
58. Finn, R. D. *et al.* The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Res.* **44,** D279–85 (2016).
59. Jacomy, M., Venturini, T., Heymann, S. & Bastian, M. ForceAtlas2, a continuous graph layout algorithm for handy network visualization designed for the Gephi software. *PLoS ONE* **9,** e98679 (2014).
60. Bastian, M., Heymann, S. & Jacomy, M. Gephi : An open source software for exploring and manipulating networks. In Third International AAAI Conference on Weblogs and Social Media, San Jose Mc Enery Convention Center, May 17, 2009 – May 20, 2009. AAAI Publications (2009).
61. Rosvall, M. & Bergstrom, C. T. Maps of random walks on complex networks reveal community structure. *Proceedings of the National Academy of Sciences U.S.A* **105,** 1118–1123 (2008). [infomap]
62. Csardi, G. & Nepusz, T. The igraph software package for complex network research. *InterJournal, Complex Systems (1695)* (2006).
63. Johnson, S. C. Hierarchical clustering schemes. *Psychometrika*, **32**, 241-254 (1967).
64. Hotelling, H. Analysis of a complex of statistical variables into principal components. *Journal of Educational Psychology* **24,** 417 (1933).
65. Hennig, C. Cluster-wise assessment of cluster stability. *Computational Statistics & Data Analysis* **52,** 258–271 (2007).

66. Fang, Y. & Wang, J. Selection of the number of clusters via the bootstrap method. Computational Statistics & Data Analysis **56,** 468–477 (2012).
67. Rosenberg, A. & Hirschberg, J. V-Measure : A conditional entropy-based external cluster evaluation measure. In Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL), Number June, pp. 410–420. (2007).
68. McCullagh, P. & Nelder, J. A. Generalized linear models, Second Edition. Chapman & Hall, London (1989). [logit]
69. Geurts, P., Ernst, D. & Wehenkel, L. Extremely randomized trees. *Machine Learning* **63,** 3–42 (2006).
70. Pedregosa, F. *et al.* Scikit-learn : Machine Learning in Python. *Journal of Machine Learning Research* **12**, 2825-2830 (2011).
71. Breiman, L. Random forests. *Machine Learning* **45,** 5–32 (2001). [ERT, OOB]
72. Han, J., Kamber, M. & Pei, J. Data Mining : Concepts and Techniques, Third Edition. The Morgan Kaufmann Series in Data Management Systems. Morgan kaufmann, Elsevier. (2012). [CV_{score}]
73. Izenman, A. Modern multivariate statistical techniques : regression. Classification, and Manifold Learning Series, Springer Texts in Statistics. Springer-Verlag, New York Inc. (2008). [$\mathrm{OOB}_{\mathrm{score}}$]

Acknowledgment: OP was supported by French Ministry of Research

Author Contributions: OP and BL designed and performed the work, and wrote the article.

Figures

Figure 1. Analytical procedure
Figure 2. Taxonomic structure of the replicon dataset
Figure 3. Visualisation of the replicon IS-based relationships
Figure 4. Model

Captions

Figure 1. Analytical procedure

Figure 2. Taxonomic structure of the replicon dataset

Numbers of replicons (a) or complete genomes (b), and represented bacterial genera are shown according to datasets and host phylum or class (Proteobacteria). 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).

Figure 3. 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).

Figure 1

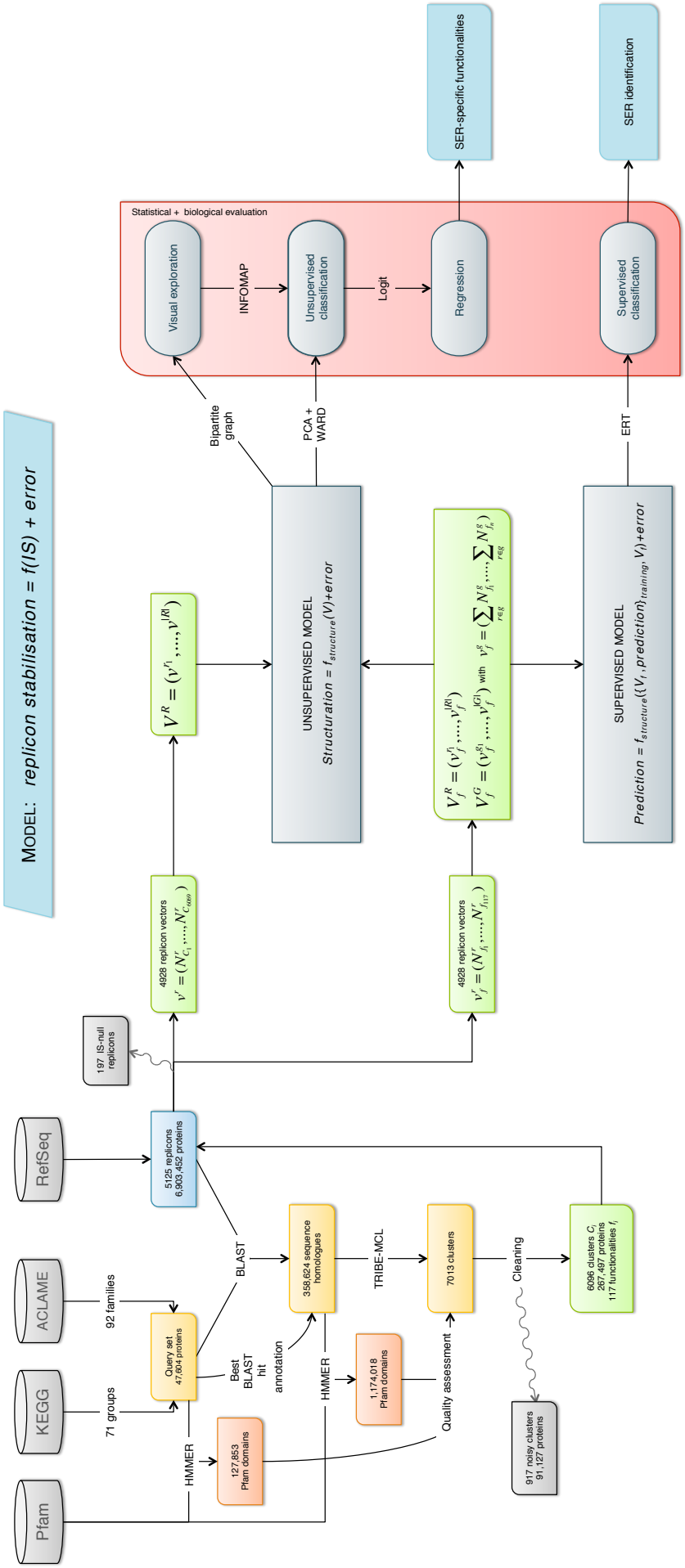

MODEL: replicon stabilisation = f(IS) + error
Pfam
KEGG
ACLAME
RefSeq
71 groups
92 families
HMMER
Query set
47,604 proteins
127,853
Pfam domains
BLAST
Best
BLAST
hit
annotation
HMMER
358,624 sequence
homologues
1,174,018
Pfam domains
TRIBE-MCL
7013 clusters
Quality assessment
Cleaning
917 noisy clusters
91,127 proteins
6096 clusters C_i
267,497 proteins
117 functionalities f_i
5125 replicons
6,903,452 proteins
197 IS-null
replicons
4928 replicon vectors
$v^r = (N^r_{C_1}, \ldots, N^r_{C_{6096}})$
4928 replicon vectors
$v^r_f = (N^r_{f_1}, \ldots, N^r_{f_{117}})$
$V^R = (v^{r_1}, \ldots, v^{|R|})$
$V^R_f = (v^{r_1}_f, \ldots, v^{|R|}_f)$
$V^G_f = (v^{g_1}_f, \ldots, v^{|G|}_f)$ with $v^g_f = (\sum_{r \in g} N^g_{f_1}, \ldots, \sum_{r \in g} N^g_{f_n})$
Bipartite
graph
PCA +
WARD
ERT
UNSUPERVISED MODEL
Structuration = $f_{structure}(V)$+error
SUPERVISED MODEL
Prediction = $f_{structure}(\{V_i, prediction\}_{training}, V_i)$+error
Statistical + biological evaluation
Visual exploration
INFOMAP
Unsupervised
classification
Logit
Regression
Supervised
classification
SER-specific functionalities
SER identification

Figure 2

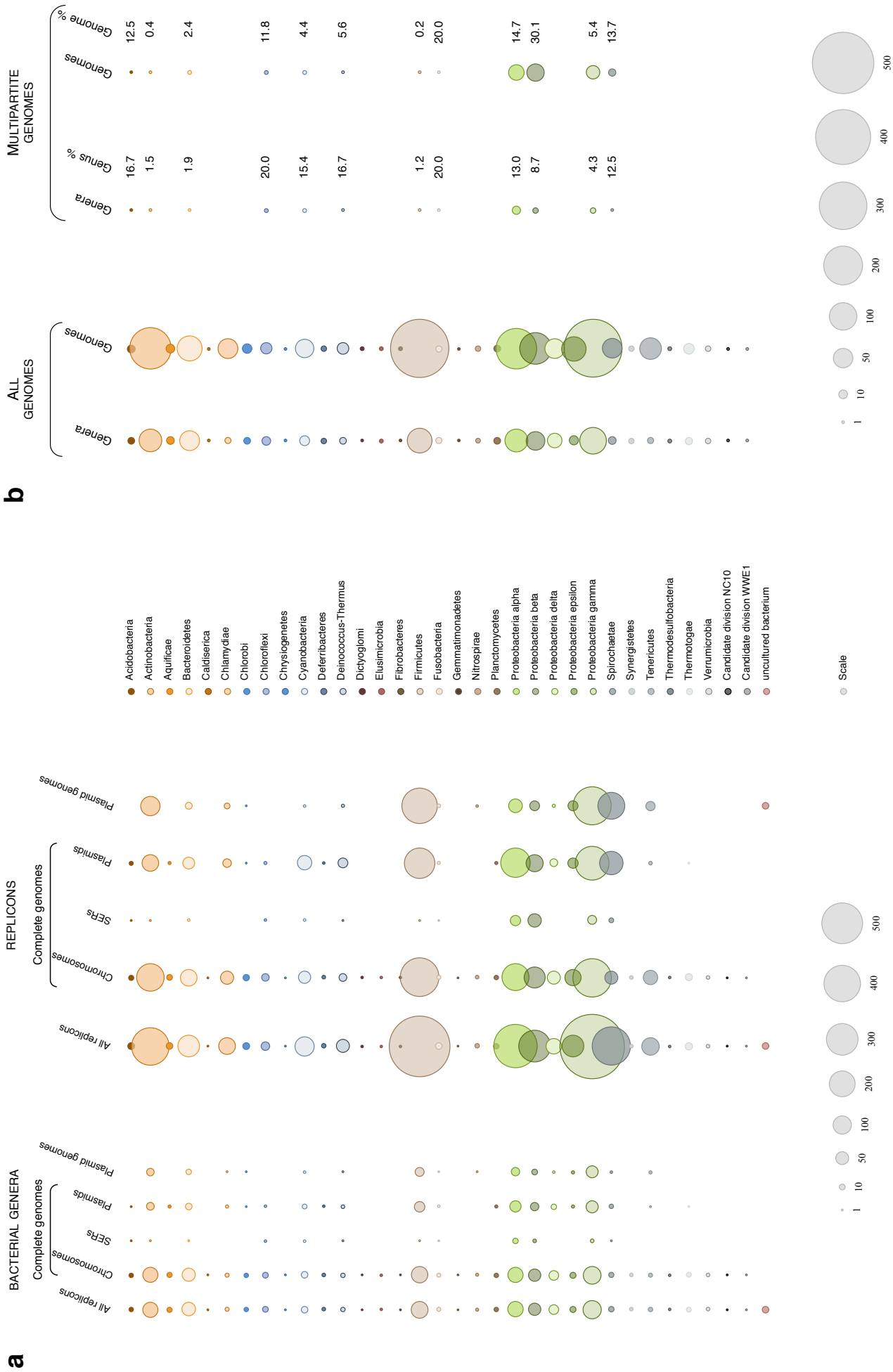
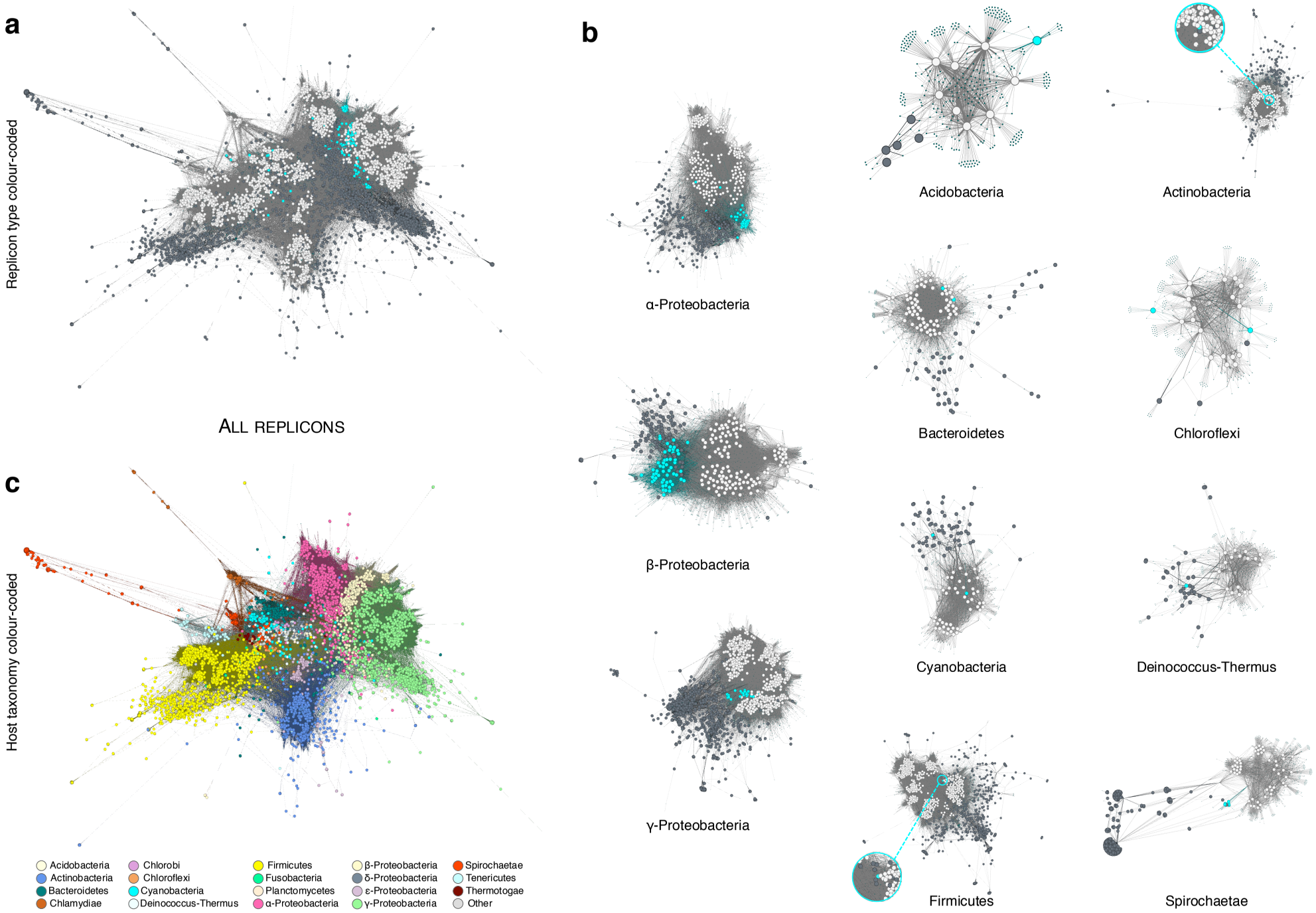

a
BACTERIAL GENERA
Complete genomes
All replicons
Chromosomes
SEFRs
Plasmids
Plasmid genomes
REPLICONS
Complete genomes
All replicons
Chromosomes
SEFRs
Plasmids
Plasmid genomes
Acidobacteria
Actinobacteria
Aquificae
Bacteroidetes
Caldiserica
Chlamydiae
Chlorobi
Chloroflexi
Chrysiogenetes
Cyanobacteria
Deferribacteres
Deinococcus-Thermus
Dictyoglomi
Elusimicrobia
Fibrobacteres
Firmicutes
Fusobacteria
Gemmatimonadetes
Nitrospirae
Planctomycetes
Proteobacteria alpha
Proteobacteria beta
Proteobacteria delta
Proteobacteria epsilon
Proteobacteria gamma
Spirochaetae
Synergistetes
Tenericutes
Thermodesulfobacteria
Thermotogae
Verrumicrobia
Candidate division NC10
Candidate division WWE1
uncultured bacterium
Scale
1
10
50
100
200
300
400
500
b
ALL GENOMES
Genera
Genomes
MULTIPARTITE GENOMES
Genera
Genus %
Genomes
Genome %
16.7
12.5
1.5
0.4
1.9
2.4
20.0
11.8
15.4
4.4
16.7
5.6
1.2
0.2
20.0
20.0
13.0
14.7
8.7
30.1
4.3
5.4
12.5
13.7
1
10
50
100
200
300
400
500

Figure 3


a
Replicon type colour-coded
ALL REPLICONS
b
α-Proteobacteria
β-Proteobacteria
γ-Proteobacteria
Acidobacteria
Actinobacteria
Bacteroidetes
Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Spirochaetae
c
Host taxonomy colour-coded
Acidobacteria
Actinobacteria
Bacteroidetes
Chlamydiae
Chlorobi
Chloroflexi
Cyanobacteria
Deinococcus-Thermus
Firmicutes
Fusobacteria
Planctomycetes
α-Proteobacteria
β-Proteobacteria
δ-Proteobacteria
ε-Proteobacteria
γ-Proteobacteria
Spirochaetae
Tenericutes
Thermotogae
Other

Tables

Table 1. IS usage comparison between types of replicons
Between classes of replicons logistic regressions for each IS function

Table 2. Identification of ERs among the extra-chromosomal replicons
a. Candidate SERs identified among plasmids
b. Candidate chromosomes identified among ERs

Table 1. IS usage comparison between classes of replicons

	IS function[a]		Chromosomes *vs.* Plasmids		Chromosomes *vs.* SERs		SERs *vs.* Plasmids	
			P_value[b]	*OR*[c]	*P_value*[b]	*OR*[c]	*P_value*[b]	*OR*[c]
KEGG entries (chromosome-like)	CbpA		8.20E-27	2,608.4	9.90E-13	22.8	5.60E-07	36.1
	Dam		6.90E-16	16.7	3.60E-02	2.0	2.40E-02	4.3
	DiaA		1.50E-15	81.9	1.20E-03	38.4		
	DnaA		3.00E-44	2,118.9	1.10E-19	239.6	3.50E-03	8.3
	DnaB		1.10E-43	1,992.9	5.10E-19	429.4	8.20E-03	3.7
	DnaB2		6.70E-03	12.6				
	DnaC	*	6.00E-12	2.6			4.60E-02	1.5
	DnaG		2.10E-50	1,861.5	1.90E-21	205.3	2.50E-03	4.5
	DnaI		5.20E-03	18.0				
REPLICATION	Dps		9.10E-21	65.3	3.50E-05	8.4	8.70E-03	6.7
	Fis		5.80E-07	180.9	3.30E-03	7.9	1.40E-02	25.1
	Hda		7.30E-07	149.1	5.30E-03	7.9	1.90E-02	18.0
	Hfq		1.40E-12	121.7	3.00E-04	6.9	8.10E-04	19.3
	H-NS	*	1.10E-05	2.8			3.80E-04	2.8
	HupA		2.70E-04	15.1				
	HupB		1.20E-53	97.6	2.30E-08	6.7	2.40E-07	11.6
	IciA	*	7.10E-20	3.2			4.50E-07	1.8
	IhfA, HimA		1.70E-12	63.8	1.40E-03	10.5	4.90E-02	6.9
	IhfB, HimD		1.20E-14	68.4	4.90E-04	8.4	8.40E-03	9.9
	Lrp	*	1.60E-19	8.4			5.40E-11	8.1
	Rob	*	6.30E-19	5.3			3.40E-08	4.2
	SeqA		1.60E-03	25.9				
	ssb		5.90E-41	298.3	5.00E-18	160.6		
	Fic	*	3.10E-09	10.3			8.60E-03	7.2
	GidA, MnmG, MTO1		5.20E-13	1,477.2	2.90E-08	110.6	4.30E-02	18.2
	GidB, RsmG		6.70E-17	6,059.9	2.20E-15	252.5	9.00E-03	32.2
	MreB		1.30E-21	1,598.2	3.90E-12	40.1	1.40E-05	24.1
	MreC		2.90E-11	1,311.2	1.30E-08	46.3	8.90E-03	32.8
	MreD		1.80E-08	459.2	6.80E-05	19.8	1.90E-02	18.2
	Mrp		6.60E-17	2,599.3	1.30E-14	35.0	2.50E-05	86.2
PARTITION	MukB	*	2.30E-03	27.4			1.90E-02	18.2
	MukE	*	3.10E-03	21.0			1.90E-02	18.2
	MukF	*	3.70E-03	19.6			1.90E-02	18.2
	ParA, Soj		2.70E-38	9.9	9.00E-06	2.6	8.40E-06	3.8
	ParB, Spo0J		2.50E-44	13.7	3.00E-03	2.1	2.30E-06	4.1
	ParC		3.00E-27	4,149.3	3.00E-16	134.0	4.60E-04	12.3
	ParE		7.30E-26	5,842.4	5.70E-15	350.1	2.40E-04	15.8
	RodA, MrdB		2.80E-12	1,233.1	9.70E-10	33.0	2.60E-03	55.3
	TrmFO, Gid		1.50E-06	182.5	4.40E-03	8.3	1.90E-02	18.0
	XerC		1.70E-43	55.0	3.10E-08	8.8	1.80E-04	6.7
	XerD		1.30E-38	26.6	4.10E-08	3.4	2.50E-06	6.2
	ScpA		1.40E-11	789.4	5.70E-07	42.9	2.10E-02	16.6
SEGREGATION	ScpB		7.50E-32	102.5	1.80E-07	25.8		
	SepF		1.80E-07	68.8	1.40E-02	12.3		
	SlmA, Ttk		3.80E-09	52.3	1.20E-02	4.6	1.50E-02	7.5
	Smc		1.60E-08	3,090.5	1.40E-05	131.9		
	SulA	*	3.30E-06	17.5			1.50E-02	10.7
	AcrA		6.60E-19	2.8	1.70E-02	1.1	5.30E-10	2.7
	AmiA, AmiB, AmiC		6.40E-36	46.4	2.90E-10	8.9	4.60E-03	3.0
	DivIC, DivA		4.90E-05	90.5	4.70E-02	8.1		
	DivIVA		4.10E-06	128.0	1.10E-02	13.4		
	EzrA		1.00E-02	13.7				
	FtsA		9.50E-12	742.7	2.20E-08	24.6	2.50E-03	41.7
	FtsB		1.10E-06	167.2	5.40E-03	16.1		
	FtsE		4.20E-24	2.3	1.30E-06	1.1	4.00E-11	1.9
	FtsI		9.80E-09	47.0	7.00E-16	3.9	2.20E-07	76.7
CELL DIVISION	FtsK, SpoIIIE		2.80E-37	76.9	2.70E-08	15.8	1.40E-02	4.2
	FtsL		1.20E-05	91.5	2.70E-02	9.8		
	FtsN		1.60E-04	53.0				
	FtsQ		1.70E-15	2,135.0	1.30E-11	99.3	9.00E-03	28.8
	FtsW, SpoVE		5.70E-16	4,266.4	4.40E-16	87.7	8.20E-04	55.0
	FtsX		9.30E-12	972.9	1.30E-08	13.8	4.80E-04	146.2
	FtsZ		3.10E-31	2,747.0	1.20E-19	101.6	9.70E-04	16.5
	MinC		4.40E-09	172.3	1.20E-02	3.0	5.80E-05	76.8
	MinD		3.10E-19	42.8	1.60E-04	2.3	5.40E-11	81.5
	MinE		9.00E-09	152.9	3.10E-02	2.6	5.90E-05	75.2
	ZapA		8.20E-09	602.8	7.40E-06	17.3	7.30E-03	56.1
	ZipA		7.90E-05	66.0				
ACLAME Families (plasmid-like)	CdsD				4.40E-02	0.1		
	DNA helicase		5.80E-21	33.6	2.70E-04	4.1	1.30E-04	9.8
	Helicase-1		1.60E-27	71.1	1.90E-13	20.0	1.10E-04	4.6
	DNA repair		2.20E-04	34.0			5.70E-04	43.6
	primase, LtrC		3.10E-02	1.8				
REPLICATION	RepA		5.90E-03	0.7				
	RepAEB		1.70E-16	0.0	1.90E-04	0.1		
	RepC						9.60E-03	2.7
	RepCJE				4.40E-02	0.1		
	RepRSE		1.30E-02	0.0	4.90E-02	0.1		
	RNA polymerase		3.20E-02	6.3				
	Rop		3.20E-02	0.0	4.40E-02	0.1		
	RuvB		1.20E-08	433.0	5.70E-08	17.7	1.40E-05	37.8
	TrfA		1.40E-02	0.3				
	ATPase, TyrK, ExoP		2.20E-20	19.4			8.50E-07	9.3
	CopG				2.70E-02	0.2	4.60E-03	23.1
	DNA-binding protein				4.40E-02	0.1		
PARTITION	FtsK, SpoIIIE		1.90E-07	6.0			9.90E-05	9.8
	ParA, ParM		1.50E-10	0.4	4.00E-06	0.3		
	ParB		5.70E-12	0.1	1.40E-05	0.2		
	serine recombinase		2.50E-06	1.4	1.50E-03	2.9	1.80E-02	0.4
	Tyrosine recombinase		3.40E-04	3.3			7.40E-04	8.7
	Xer-like Tyrosine		7.60E-11	2.0			6.30E-03	1.6
	Ccd (PSK)		4.60E-02	3.9				
	HicAB (PSK)		4.30E-05	25.2			4.80E-03	15.1
	HigBA (PSK)		3.30E-15	3.4	2.40E-02	1.5	1.20E-03	2.5
	MazEF (PSK)		1.20E-11	5.2	2.90E-02	2.6		
	ParC (PSK)				4.40E-02	0.1		
MAINTENANCE	ParDE (PSK)		5.50E-08	2.3			7.80E-05	3.4
	PhD, Doc (PSK)		3.20E-07	11.9			2.90E-03	8.8
	plasmid maintenance				4.40E-02	0.1		
	RelBE (PSK)		2.70E-08	3.5			6.10E-04	4.2
	VapBC/Vag (PSK)		1.20E-09	3.9			1.40E-05	5.8

OR colour code

- $10^4 \leq OR$
- $10^3 \leq OR < 10^4$
- $10^2 \leq OR < 10^3$
- $10^1 \leq OR < 10^2$
- $10^0 \leq OR < 10^1$
- $10^{-1} \leq OR < 10^0$
- $10^{-2} \leq OR < 10^{-1}$
- $10^{-3} \leq OR < 10^{-2}$
- $10^{-4} \leq OR < 10^{-3}$
- $OR < 10^{-4}$

[a] Stars indicate chromosome-like KEGG-annotated functionalities which level of bias is of the same order of magnitude in chromosomes and SERs when compared to plasmids.

[b] Model significance. $0 < P_value < 0.01$: significant (darker shade of grey); $0.01 < P_value < 0.05$: poorly significant (lighter shade of grey); $0.05 < P_value$: non significant (not shown)

[c] *OR*: Odd-ratio favouring the first class (blue) or the second class (green)

Table 2. Identification of ERs among the extra-chromosomal replicons

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] (AP0109)	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

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

[a] Probability for an extra-chromosomal replicon, *i.e.*, plasmid or SER, to belong to the SER (A) or Chromosome (B) class according to the supervised classification procedures. Colour code varies from probability value = 0.500 (yellow) to probability value = 1.000 (blue).

Extended data Figures

Figure 1. Structural features of the replicons

Figure 2. Properties of the IS clustering

Figure 3. Influence of granularity on clustering

Legends

Figure 1. Structural features of the replicons

Boxplots of the length, number of protein-coding genes, number of rRNA genes, number of tRNA genes and number of pseudogenes for the 2016 chromosomes (green), 129 SERs (blue), and 2783 plasmids (yellow) included in the final dataset (4928 replicons).

Figure 2. 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 across the 358,624 putative IS homologues (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) Pourcentage 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).

Figure 3. Influence of granularity on the clustering

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

Figure 1

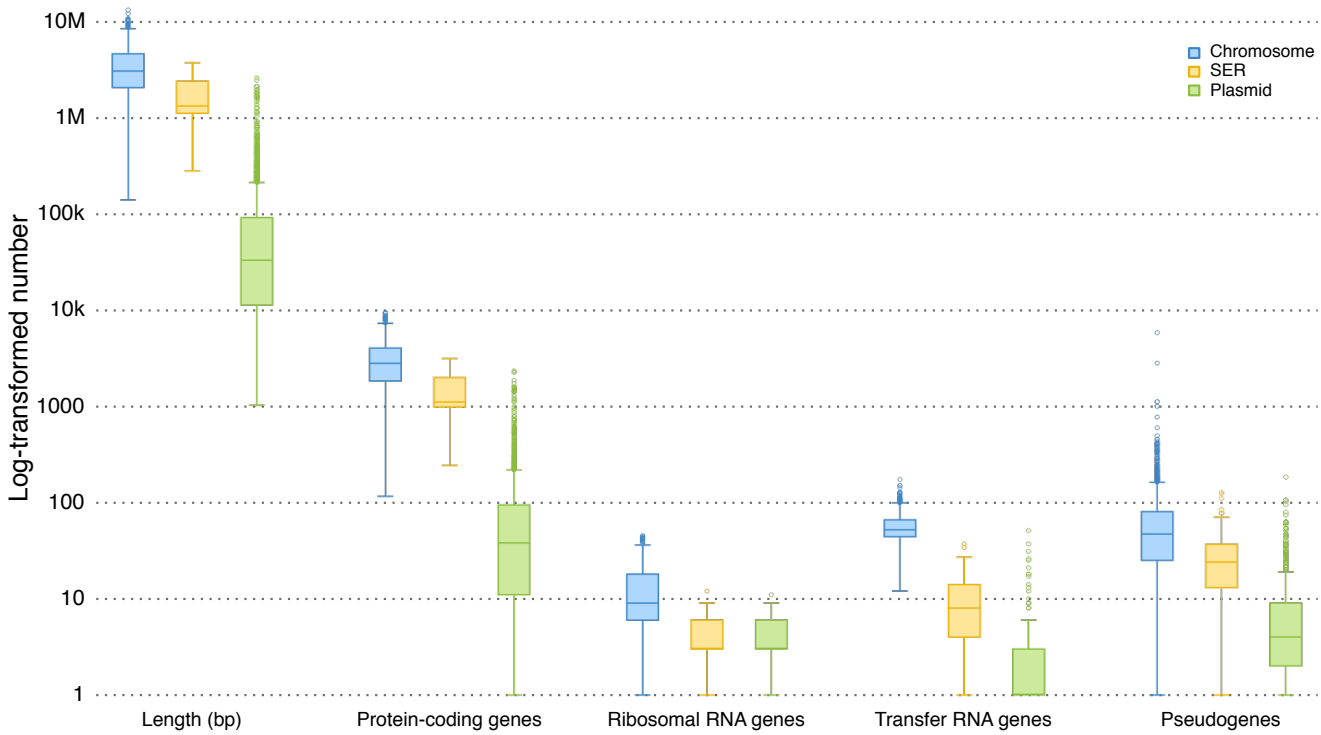

Chromosome
SER
Plasmid
Log-transformed number
10M
1M
100k
10k
1000
100
10
1
Length (bp)
Protein-coding genes
Ribosomal RNA genes
Transfer RNA genes
Pseudogenes

Figure 2

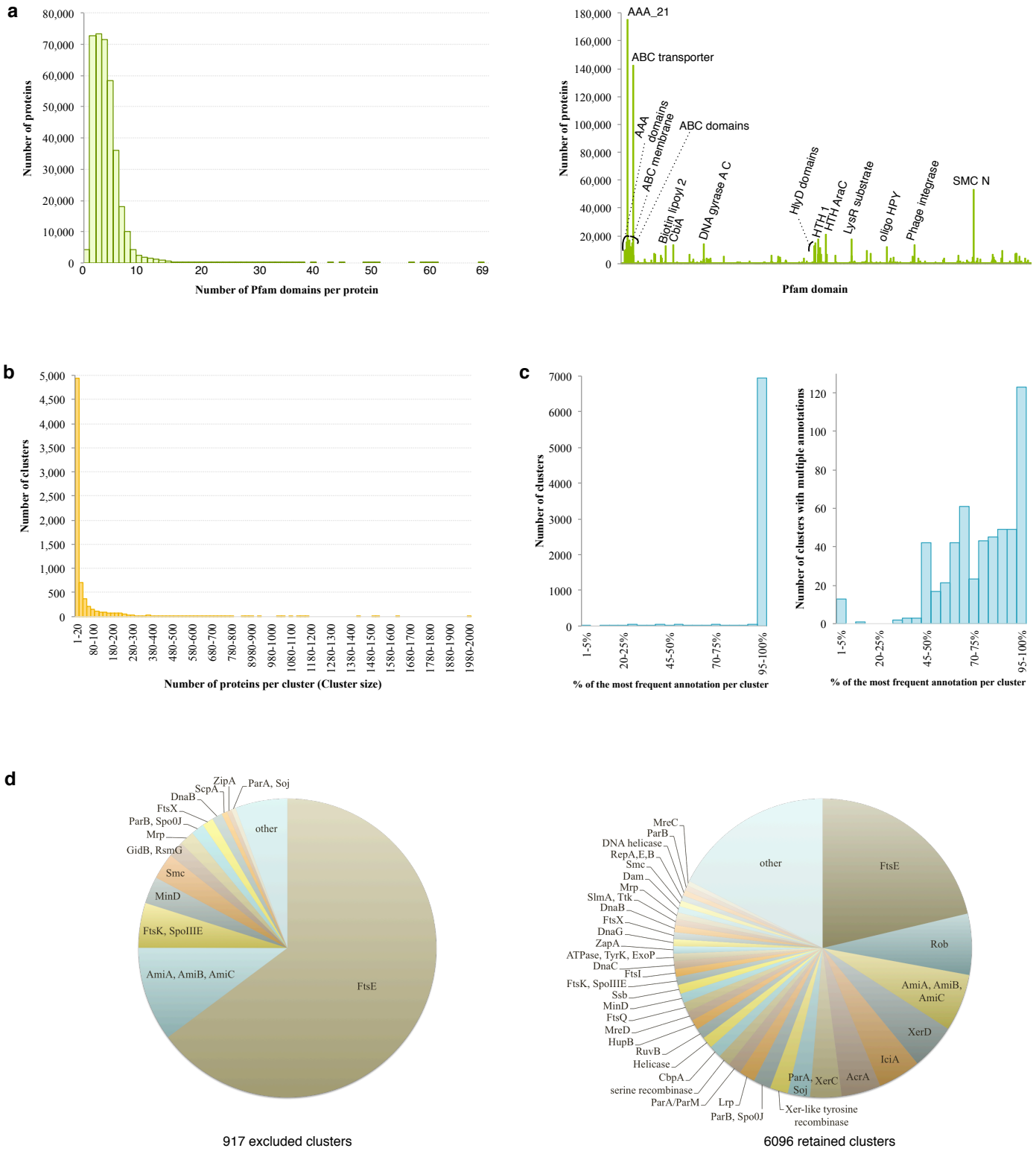
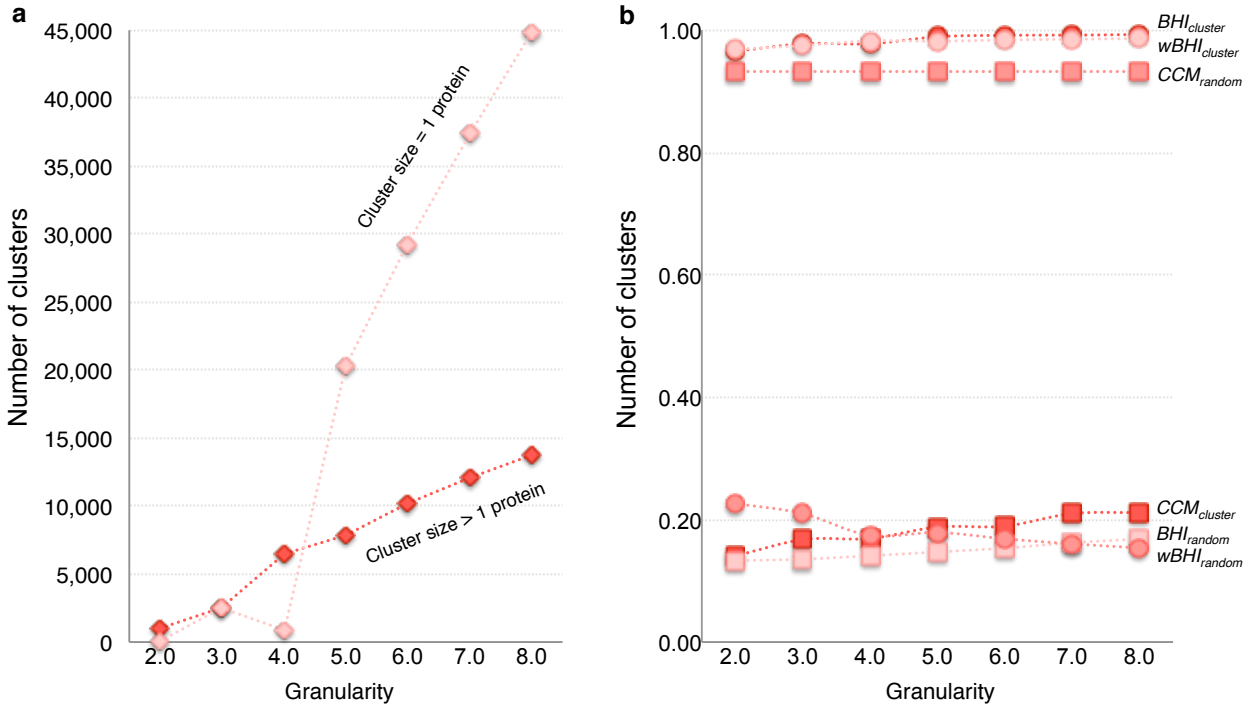

a
Number of proteins
Number of Pfam domains per protein
AAA_21
ABC transporter
AAA domains
ABC membrane
ABC domains
Biotin lipoyl 2
CbiA
DNA gyrase A C
HlyD domains
HTH 1
HTH AraC
LysR substrate
oligo HPY
Phage integrase
SMC N
Pfam domain
b
Number of clusters
Number of proteins per cluster (Cluster size)
c
Number of clusters
% of the most frequent annotation per cluster
Number of clusters with multiple annotations
d
other
FtsE
Rob
AmiA, AmiB, AmiC
XerD
IciA
AcrA
XerC
ParA, Soj
Xer-like tyrosine recombinase
ParB, Spo0J
Lrp
ParA/ParM
serine recombinase
CbpA
Helicase
RuvB
HupB
MreD
FtsQ
MinD
Ssb
FtsK, SpoIIIE
FtsI
DnaC
ATPase, TyrK, ExoP
ZapA
DnaG
FtsX
DnaB
SlmA, Ttk
Mrp
Dam
Smc
RepA,E,B
DNA helicase
ParB
MreC
917 excluded clusters
6096 retained clusters

Figure 3


a
45,000
40,000
35,000
30,000
25,000
20,000
15,000
10,000
5,000
0
Number of clusters
Cluster size = 1 protein
Cluster size > 1 protein
2.0
3.0
4.0
5.0
6.0
7.0
8.0
Granularity
b
1.00
0.80
0.60
0.40
0.20
0.00
Number of clusters
BHI cluster
wBHI cluster
CCM random
CCM cluster
BHI random
wBHI random
2.0
3.0
4.0
5.0
6.0
7.0
8.0
Granularity

Extended data Tables

Table 1. Evaluation of the replicon IS-based clusterings

Table 2. IS cluster-based unsupervised classification of SERs

Table 3. Function-based unsupervised classification of SERs

Table 4. Performance of the ERT classification procedures

Table 5. SER probability to belong to the *SER* class

Table 6. IS profiles of *Paracoccus denitrificans vs. Rhodobacter sphaeroides* (Rhodobacterales).

Table 2. IS protein cluster-based unsupervised classification of SERs

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

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

[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

Table 3. 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

[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 sizes of the clusters
[d] mean value of the SER stability estimator for a given bacterial genus

Table 4. Performance of the ERT classification procedures

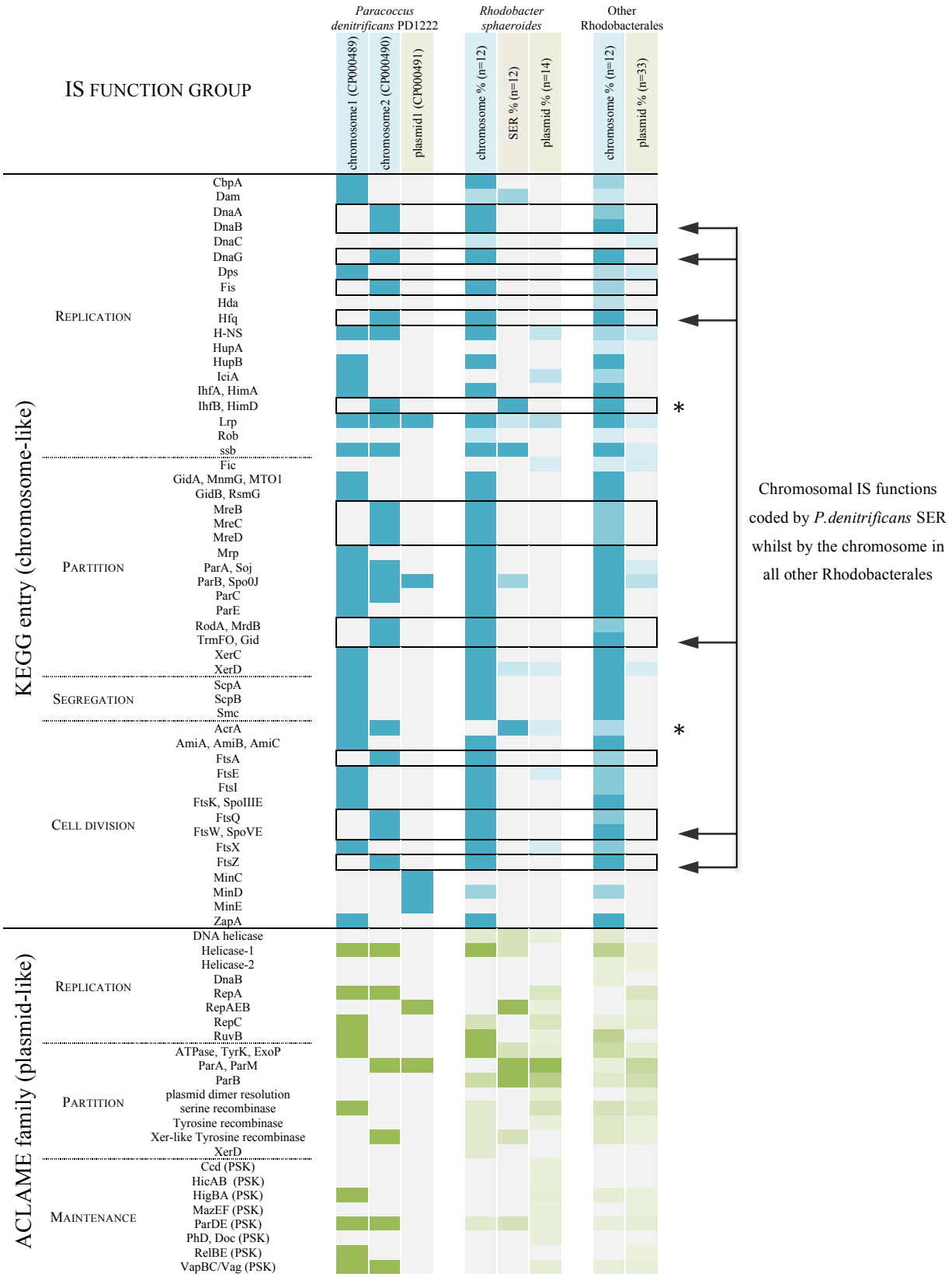
TRAINING SET [a]	CV_{score} [b]	$\sigma_{CV_{score}}$ [c]	OOB_{score} [d]	$\sigma_{OOB_{score}}$ [e]
$\{E_{SER}, E'_{plasmid}\}$	0.96	-	0.96	-
$\{E_{SER}, E'_{plasmid}\}^{it}$	0.92	0.02	0.93	0.02
$\{E_{chromosome}, E'_{plasmid}\}$	1.00	-	1.00	-
$\{E_{SER}, E_{chromosome}\}^{it}$	0.98	0.00	0.98	0.01

[a] $E_{chromosome}$ and E_{SER} are host genus-normalized sets of chromosomes or SERs, respectively, as in Table S8. $E'_{plasmid}$ is derived from the INFOMAP clustering solution, by discarding plasmids belonging to clusters also harbouring SERs or chromosomes, and normalized according to host genus. *it* designates the iterative procedure.
[b] Cross-validation score or mean of iteration cross-validation scores.
[c] Standard deviation of iteration cross-validation scores.
[d] Out-of bag estimate or mean of iteration Out-of bag estimates.
[e] Standard deviation of iteration Out-of bag estimates.

Table 5. 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

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

Table 6. IS profiles of *Paracoccus denitrificans vs. Rhodobacter sphaeroides* (Rhodobacterales).

Frames show functions that are coded only by the SER in *P. denitrificans* whilst by the chromosome in all other Rhodobacterales. The asterisks indicate the two chromosome-like functions coded only by the SER in *R. sphaeroides*.