Supplementary Material

Genetic competence drives genome diversity in *Bacillus subtilis*

Patrícia H. Brito[1,2*], Bastien Chevreux[3], Cláudia R. Serra[4], Ghislain Schyns[3], Adriano O. Henriques[4], José B. Pereira-Leal[1,5*]

[1]Instituto Gulbenkian de Ciência, Oeiras, Portugal; [2]Nova Medical School, Faculdade de Ciências Médicas, Universidade Nova de Lisboa, Portugal; R&1D; [3]DSM Nutritional Products, Ltd., Kaiseraugst, Switzerland; [4]Instituto de Tecnologia Química e Biológica, Oeiras, Portugal; [5]Ophiomics - Precision Medicine, Lisboa, Portugal.

* Corresponding authors
Patrícia H. Brito: pbrito@igc.gulbenkian.pt
José B. Pereira-Leal: jleal@igc.gulbenkian.pt
Instituto Gulbenkian de Ciência, Oeiras, Rua da Quinta Grande 6, 2780-156 Oeiras, Portugal; Phone: + 351214407900; Fax: +351214407970

Supplementary methods

Orthology mapping

Gene orthology mapping for phylogenetic analysis was carried out using all-against-all blastp and three different sequence clustering algorithms, bidirectional best hit analysis (BDBH) (Contreras-Moreira & Vinuesa 2013), clusters of orthologous groups (COG) (Contreras-Moreira & Vinuesa 2013; Kristensen et al. 2010), and Markov cluster algorithm (OMCL) (Li et al. 2003) imposing a minimum pairwise sequence alignment coverage of 80% (SAC) and 50% of sequence identity (SI). From each of those analyses we retained all clusters characterized by the absence of inparalogs (sequences with best hits in its own genome), and the ubiquitous presence in all strains (core genes). The intersection of the clusters obtained from the three clustering algorithms was our final list of orthologous genes to be used in the phylogenetic analysis.

Orthology mapping to characterize the pangenome was performed with OMCL and following Tettelin *et al* (Tettelin et al. 2005) we imposed 50% of SAC and 80% of SI, we did not exclude clusters with inparalogs, and included all clusters independent of size. As in (Lefébure & Stanhope 2007) we only considered genes as taxon specific (clusters of size 1) if sequences were larger than 50 amino acids and had no blast hit with any other protein (E-value 1.0E-05). This was done to minimize the inclusion of truncated genes found at the end of contigs of open genomes. Still, because it is possible that assembly errors would truncate genes at the end of contigs precluding their annotation, we also performed an explicit test for the existence of non-annotated genes at both ends of the contigs of open genomes. This test was done as follows: for each genome, we extracted all nucleotide sequences hanging from the beginning of the contig to the beginning of the first annotated gene as well as the nucleotide sequences hanging from the end of the last annotated gene to the end of the contig. By definition, truncated genes should sit on those nucleotide positions. We BLASTN all those sequences against a database built with one representative sequence from each OMCL cluster (10137 in total). We considered a significant hit any BLASTN hit larger than 150 nucleotides, with more than 80% sequence identity in BLAST query/subject pairs, and 50% coverage in BLASTN pairwise alignments (alignment length/subject sequence length). This analysis lead to the identification of 510 false negatives that were corrected the final matrix (Supplementary dataset 2). Orthology mapping was done with get_homologues package (Contreras-Moreira & Vinuesa 2013).

Phylogenetic analysis of BmrA protein

For this analysis we extracted all 103632_bmrA homologs mapped in the KO (KEGG Orthology) database (K18104) and added homologous sequences from *B. subtilis* genomes not available in the

KEGG database totaling 498 sequences. We aligned this dataset with Mafft (G-ins-I, and Blosum 62) and run maximum likelihood and bootstrap analyses as before. From the results we identified the most inclusive well-supported clade that included *B. subtilis* sequences and its closest relatives. To this subset of sequences we added representatives of the remaining clades to build a simplified dataset that was realigned and reanalyzed with maximum likelihood and bootstrap to produce the final result presented here. Within *B. subtilis* we identified three homologous copies. One from the softcore genome (*B. subtilis* str. 168; NCBI_ProteinID: NP_391362.1; Genome location: NC_000964: 3577745..3579514) that has been transmitted vertically within *B. subtilis* group. The other two copies belong to the shell (*B. subtilis* str. XF-1; NCBI_ProteinID: AGE62337; Genome location: NC_020832:531386..533119) and the cloud (*B. subtilis* str. BSn5; NCBI_ProteinID: ADV95129; Genome location: NC_014976:2435115..2436872) genome and were recently acquired by *B. subtilis* strains from unrelated organisms. Since initial acquisition these later two genes have been transferred across strains sampled among Plant niches. These three gene copies are on average 58% divergent (between-group mean distance measured as the proportion of amino acid differences) hence too divergent to be detected as orthologous copies by the algorithm used in this study.

Supplementary results

Shared genes and core genome extrapolation (fig. 2d)

The average values of the shared genes were extrapolated by fitting the exponential decay function:

$$F_c(n) = 2119\ exp\left[\frac{-n}{18.45}\right] + 1659$$

[1]

where n is the number of distinct genomes, and the size of the core genome as $n \to \infty$ is 1659. Estimations were performed with get_homologues package (Contreras-Moreira & Vinuesa 2013) following Tettelin *et al* (Tettelin et al. 2005).

Strain-specific genes and pangenome extrapolation (fig. 2e)

The average number of specific genes per strain closely follows an exponential decay function with function:

$$P(n) = 4194 + 57.0(n-1) + 219\ exp\left[\frac{-2}{7.80}\right].\left[\frac{1 - exp\left(\frac{-(n-1)}{7.80}\right)}{1 - exp\left(\frac{-1}{7.80}\right)}\right]$$

[2]

where n is the number of distinct genomes, and 57.0 is the number of new genes asymptotically predicted for each new sequenced genome. For n=42 genomes the extrapolated pangenome size is $P(n) \approx 7932.9$. Estimations were performed with get_homologues package (Contreras-Moreira & Vinuesa 2013) following Tettelin *et al* (Tettelin et al. 2005).

Shared and strain-specific genes for core and pangenome extrapolation using only *B. subtilis* wild strains (32 genomes) (fig. S3)

Fitted functions are,

$$F_c(n) = 2097\ exp\left[\frac{-n}{12.77}\right] + 1803$$

[3]

where n is the number of distinct genomes, and the size of the core genome as $n \to \infty$ is 1803(SE 225.20), and

$$P(n) = 4169 + 73.0(n-1) + 276\ exp\left[\frac{-2}{6.35}\right].\left[\frac{1 - exp\left(\frac{-(n-1)}{6.35}\right)}{1 - exp\left(\frac{-1}{6.35}\right)}\right]$$

[4]

where n is the number of distinct genomes, and 73.0 is the number of new genes asymptotically predicted for each new sequenced genome. For n=32 genomes the extrapolated pangenome size is $P(n) \approx 7806.5$ (SE 202.77). Estimations were performed with get_homologues package (Contreras-Moreira & Vinuesa 2013) following Tettelin *et al* (Tettelin et al. 2005).

Supplementary references

Contreras-Moreira B, Vinuesa P. 2013. GET_HOMOLOGUES, a versatile software package for scalable and robust microbial pangenome analysis. Appl Environ Microbiol. 79:7696–7701. doi: 10.1128/AEM.02411-13.

Kristensen DM et al. 2010. A low-polynomial algorithm for assembling clusters of orthologous groups from intergenomic symmetric best matches. Bioinformatics. 26:1481–1487. doi: 10.1093/bioinformatics/btq229.

Krzywinski M et al. 2009. Circos: An information aesthetic for comparative genomics. Genome Res. 19:1639–1645. doi: 10.1101/gr.092759.109.

Lefébure T, Stanhope MJ. 2007. Evolution of the core and pan-genome of Streptococcus: positive selection, recombination, and genome composition. Genome Biol. 8:R71. doi: 10.1186/gb-2007-8-5-

r71.

Li L, Stoeckert CJ, Roos DS. 2003. OrthoMCL: identification of ortholog groups for eukaryotic genomes. Genome Res. 13:2178–2189. doi: 10.1101/gr.1224503.

Tettelin H et al. 2005. Genome analysis of multiple pathogenic isolates of *Streptococcus agalactiae*: Implications for the microbial 'pan-genome'. P Natl Acad Sci Usa. 102:13950–13955. doi: 10.2307/3376809?ref=search-gateway:46f22eed2f66acb6f02d383ffa46dc1e.

Supplementary figure legends

Fig. S1. Phylogenetic analysis of *B. subtilis* group with distant outgroups. Maximum-likelihood (LG+I+G) and bootstrap analyses were carried out on a concatenated dataset of 398 core proteins (104056 AA). This analysis was done to confirm the taxonomic identification of genomes available at the NCBI database and to confirm the outgroup selection. Five genomes (in red) were excluded from further analyses as they are more closely related to other species than to *B. subtilis*. Numbers on the branches are bootstrap support. Due to the small branch lengths at the tips of the tree we only show bootstrap support for the most basal branches. The scale bar indicates expected number of amino acid substitutions per site.

Fig. S2. Phylogenetic analysis of *B. subtilis* group and its closest outgroups. Maximum-likelihood (GTR+I+G+X) and bootstrap analyses were carried out on a concatenated dataset of 685 core genes (520227 nt). Numbers on the branches are bootstrap support. Due to the small branch lengths at the tips of the tree we only show bootstrap support for the most basal branches. The scale bar indicates expected number of nucleotide substitutions per site. Arrow indicates the branch that separates *B. subtilis* from other species.

Fig. S3. **Core and pangenome in wild strains**. Quantification of the core (a) and the pangenome (b) using only wild strains of *B. subtilis*. The number of shared (a) and novel (b) genes estimated after 100 random samples of the 32 genomes is plotted against the number of *n* strains sequentially added. Line in red is the fitted curve following exponential functions as in Tettelin (Tettelin et al. 2005).

Fig. S4. Distribution of intact prophages in *B. subtilis* shows the existence of three high frequency elements in the species pangenome. PBSX is ubiquitous among genomes in our dataset and SPβ and *skin* elements are largely restricted to *B. s. subtlis*. Detection of intact prophages was performed using PHAST and identification resulted from BLAST searches of prophage sequences against NCBI nr/nt databases. Presence and absence (black/white squares) is plotted on the core genome tree of figure 1. Numbers on the terminal branches indicate total numbers of intact prophages detected in each genome. The tree is not drawn to scale. Colors in the taxa names reflect niche at the site of sampling following fig. 1.

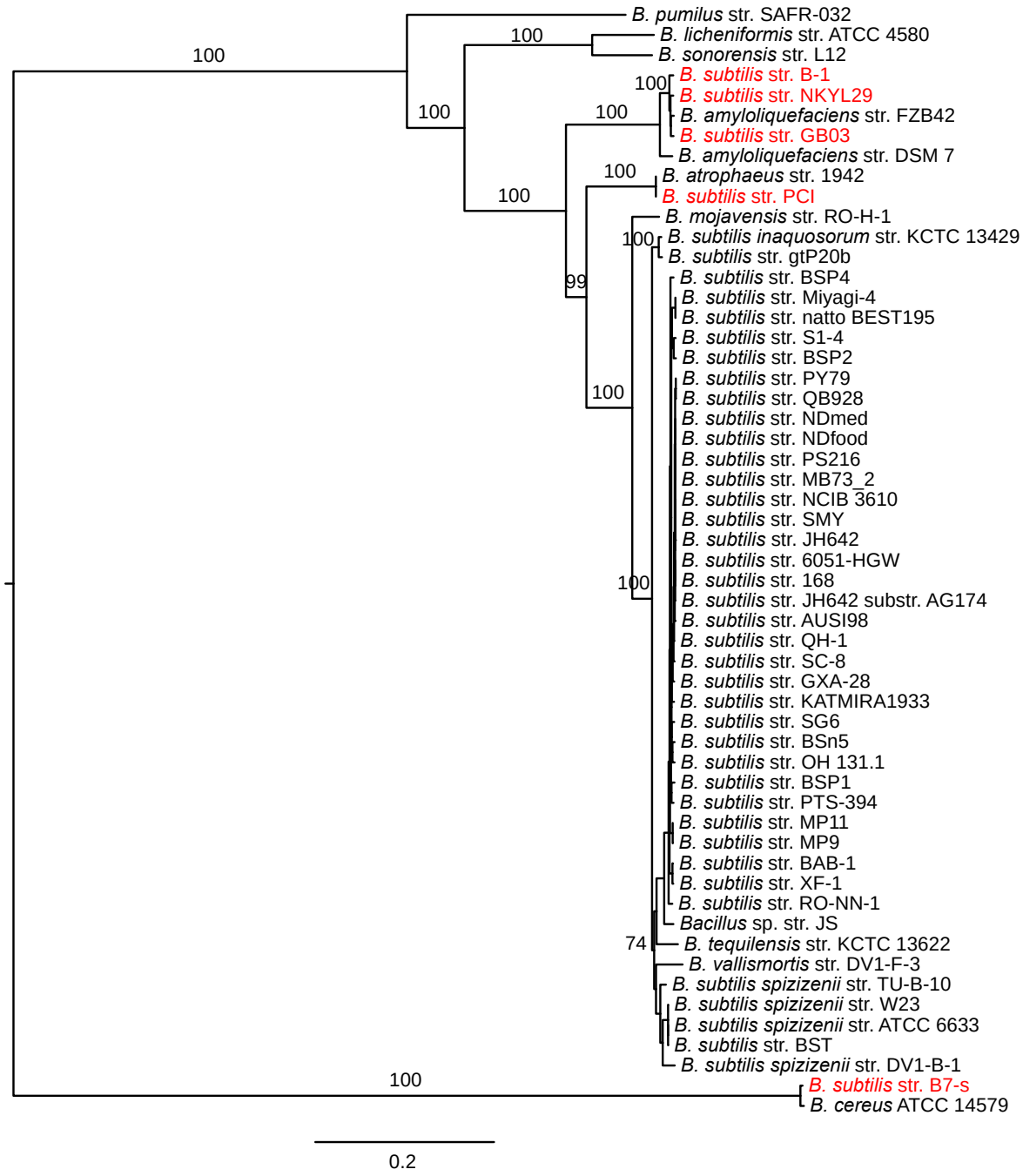
Fig. S5. Variable pangenome is distributed throughout the chromosome. Genome atlas of *B. subtilis* strains with closed genomes. The circles from the center represent the location of the following elements: coding sequences (dark grey), tRNA (red) and prophages (intact: dark green;

questionable: green; incomplete: light green), core genome (black), softcore genome (yellow), shell genome (orange), cloud genome (red). The outer circle represents relative frequency of each gene cluster in the dataset. Images were created with Circos v. 0.67 (Krzywinski et al. 2009). The radial histograms at the center of the genome atlas are rose diagrams showing the dispersion of cloud genes around the genome after transforming each gene middle point position into radians. The genome was divided into 100 bins and the radii of the sectors are equal to the square root of the relative frequencies of the cloud genome making the area of each sector proportional to its frequency. The rose diagram was scaled to the inner circle of the genome atlas.

Fig. S6. Lateral gene acquisition of paralogue copies of a *B. subtilis* softcore gene from unrelated organisms and posterior lateral transmission within the species. Phylogenetic history of the *bmrA* gene. This gene codes for a multidrug resistance ABC transporter ATP-binding protein, and its evolutionary history is characterized by frequent lateral transference events between gram-positive bacteria. In *B. subtilis* we identify three homologous copies, one from the softcore genome with vertical transmission, and two xenologue copies from the shell and the cloud genome that are recent LGT events from unrelated organisms. Since initial acquisition these later two genes have been transferred across strains sampled among Plant niches. The phylogenetic tree was computed with maximum-likelihood (LG+I+G), and numbers on the branches are bootstrap support. Analyses were carried out in RaxML. The scale bar indicates expected number of nucleotide substitutions per site.

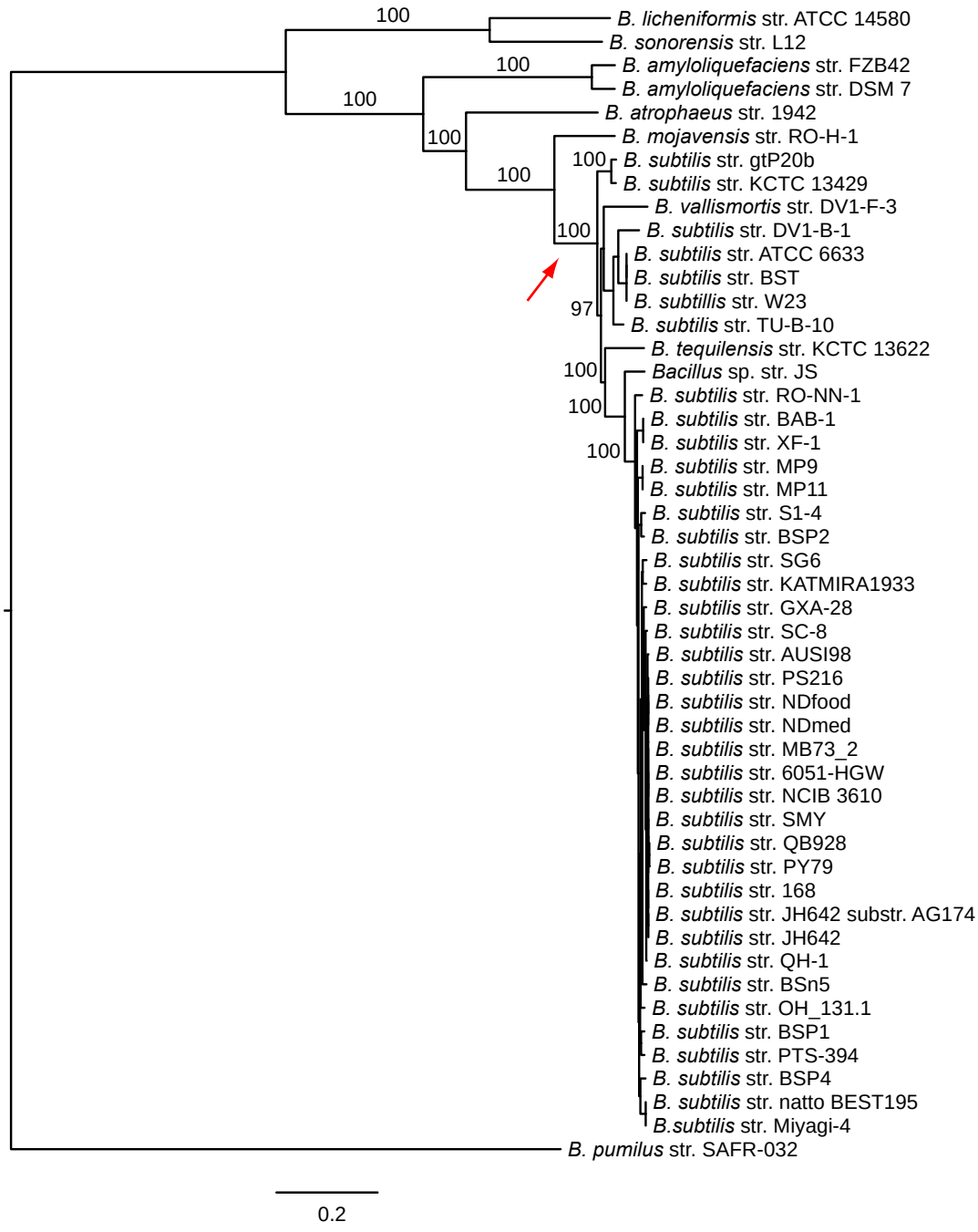
Supplementary figures

196 **Fig. S1.**



197

198 **Fig. S2.**


100
100
100
100
100
100
100
100
97
100
100
100
B. licheniformis str. ATCC 14580
B. sonorensis str. L12
B. amyloliquefaciens str. FZB42
B. amyloliquefaciens str. DSM 7
B. atrophaeus str. 1942
B. mojavensis str. RO-H-1
B. subtilis str. gtP20b
B. subtilis str. KCTC 13429
B. vallismortis str. DV1-F-3
B. subtilis str. DV1-B-1
B. subtilis str. ATCC 6633
B. subtilis str. BST
B. subtillis str. W23
B. subtilis str. TU-B-10
B. tequilensis str. KCTC 13622
Bacillus sp. str. JS
B. subtilis str. RO-NN-1
B. subtilis str. BAB-1
B. subtilis str. XF-1
B. subtilis str. MP9
B. subtilis str. MP11
B. subtilis str. S1-4
B. subtilis str. BSP2
B. subtilis str. SG6
B. subtilis str. KATMIRA1933
B. subtilis str. GXA-28
B. subtilis str. SC-8
B. subtilis str. AUSI98
B. subtilis str. PS216
B. subtilis str. NDfood
B. subtilis str. NDmed
B. subtilis str. MB73_2
B. subtilis str. 6051-HGW
B. subtilis str. NCIB 3610
B. subtilis str. SMY
B. subtilis str. QB928
B. subtilis str. PY79
B. subtilis str. 168
B. subtilis str. JH642 substr. AG174
B. subtilis str. JH642
B. subtilis str. QH-1
B. subtilis str. BSn5
B. subtilis str. OH_131.1
B. subtilis str. BSP1
B. subtilis str. PTS-394
B. subtilis str. BSP4
B. subtilis str. natto BEST195
B.subtilis str. Miyagi-4
B. pumilus str. SAFR-032
0.2

199

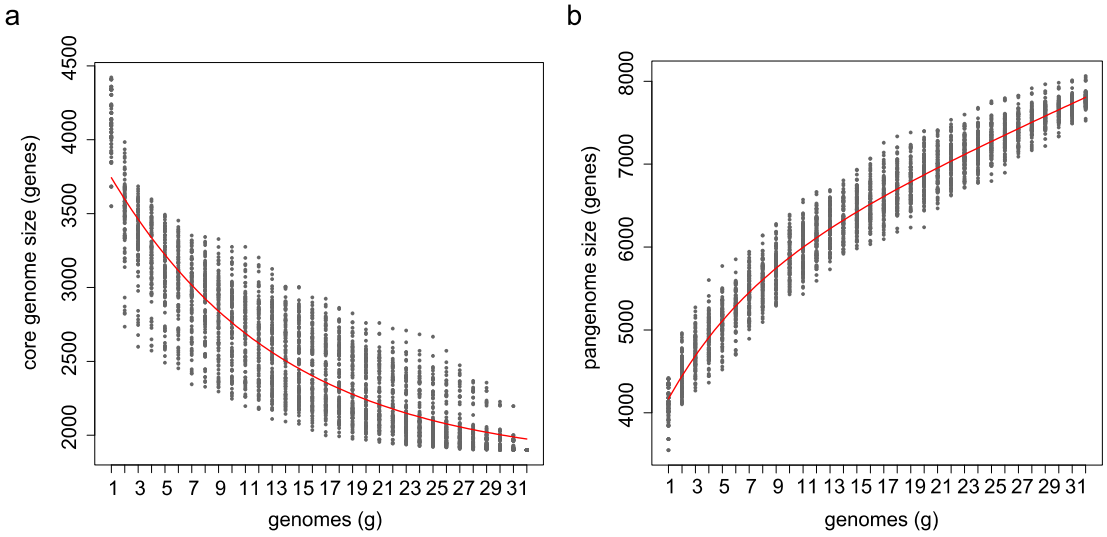

a
core genome size (genes)
2000
2500
3000
3500
4000
4500
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
genomes (g)
b
pangenome size (genes)
4000
5000
6000
7000
8000
1 3 5 7 9 11 13 15 17 19 21 23 25 27 29 31
genomes (g)

Fig. S3.

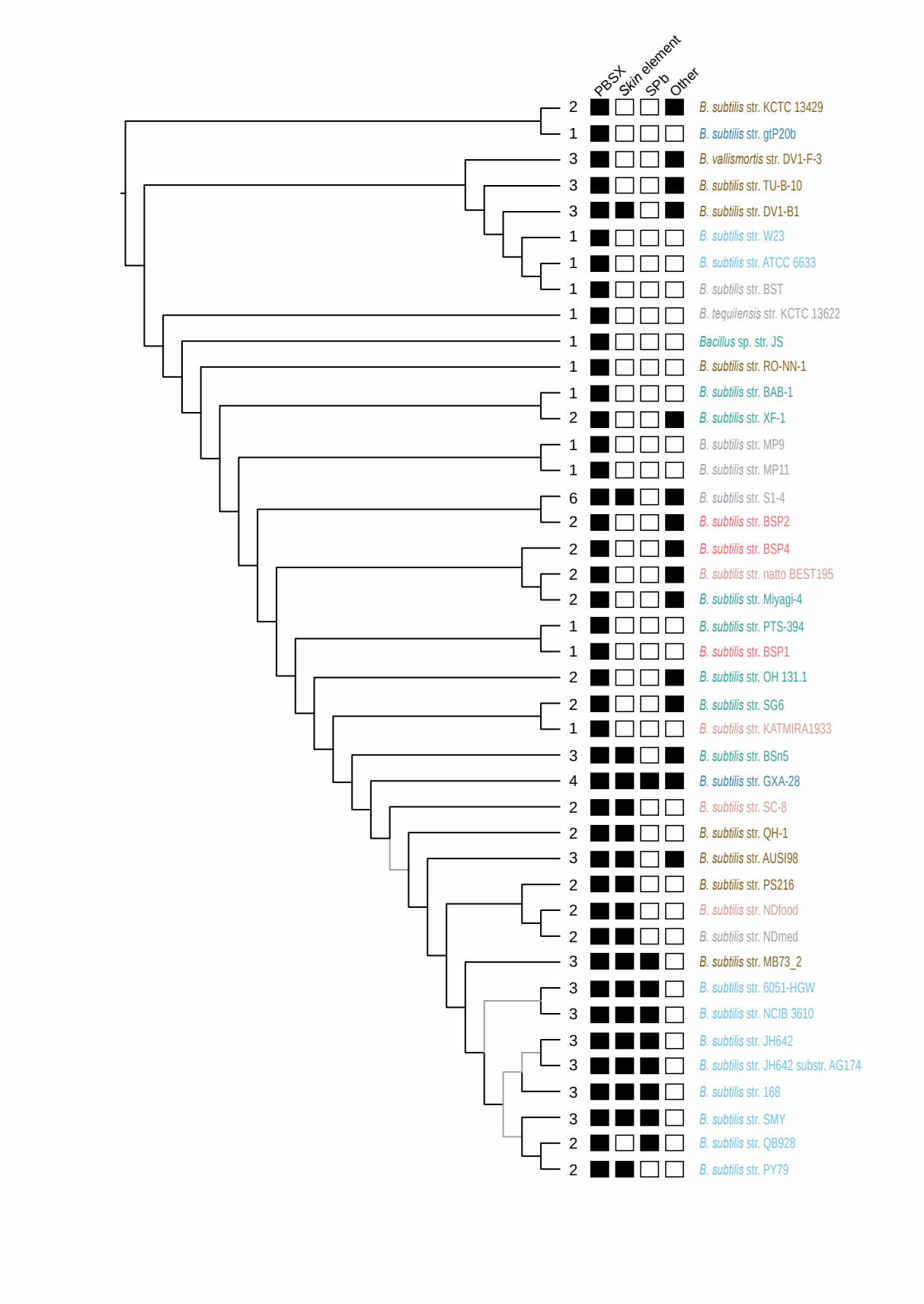

PBSX
Skin element
SPb
Other
2 B. subtilis str. KCTC 13429
1 B. subtilis str. gtP20b
3 B. vallismortis str. DV1-F-3
3 B. subtilis str. TU-B-10
3 B. subtilis str. DV1-B1
1 B. subtilis str. W23
1 B. subtilis str. ATCC 6633
1 B. subtilis str. BST
1 B. tequilensis str. KCTC 13622
1 Bacillus sp. str. JS
1 B. subtilis str. RO-NN-1
1 B. subtilis str. BAB-1
2 B. subtilis str. XF-1
1 B. subtilis str. MP9
1 B. subtilis str. MP11
6 B. subtilis str. S1-4
2 B. subtilis str. BSP2
2 B. subtilis str. BSP4
2 B. subtilis str. natto BEST195
2 B. subtilis str. Miyagi-4
1 B. subtilis str. PTS-394
1 B. subtilis str. BSP1
2 B. subtilis str. OH 131.1
2 B. subtilis str. SG6
1 B. subtilis str. KATMIRA1933
3 B. subtilis str. BSn5
4 B. subtilis str. GXA-28
2 B. subtilis str. SC-8
2 B. subtilis str. QH-1
3 B. subtilis str. AUSI98
2 B. subtilis str. PS216
2 B. subtilis str. NDfood
2 B. subtilis str. NDmed
3 B. subtilis str. MB73_2
3 B. subtilis str. 6051-HGW
3 B. subtilis str. NCIB 3610
3 B. subtilis str. JH642
3 B. subtilis str. JH642 substr. AG174
3 B. subtilis str. 168
3 B. subtilis str. SMY
2 B. subtilis str. QB928
2 B. subtilis str. PY79

Fig. S4

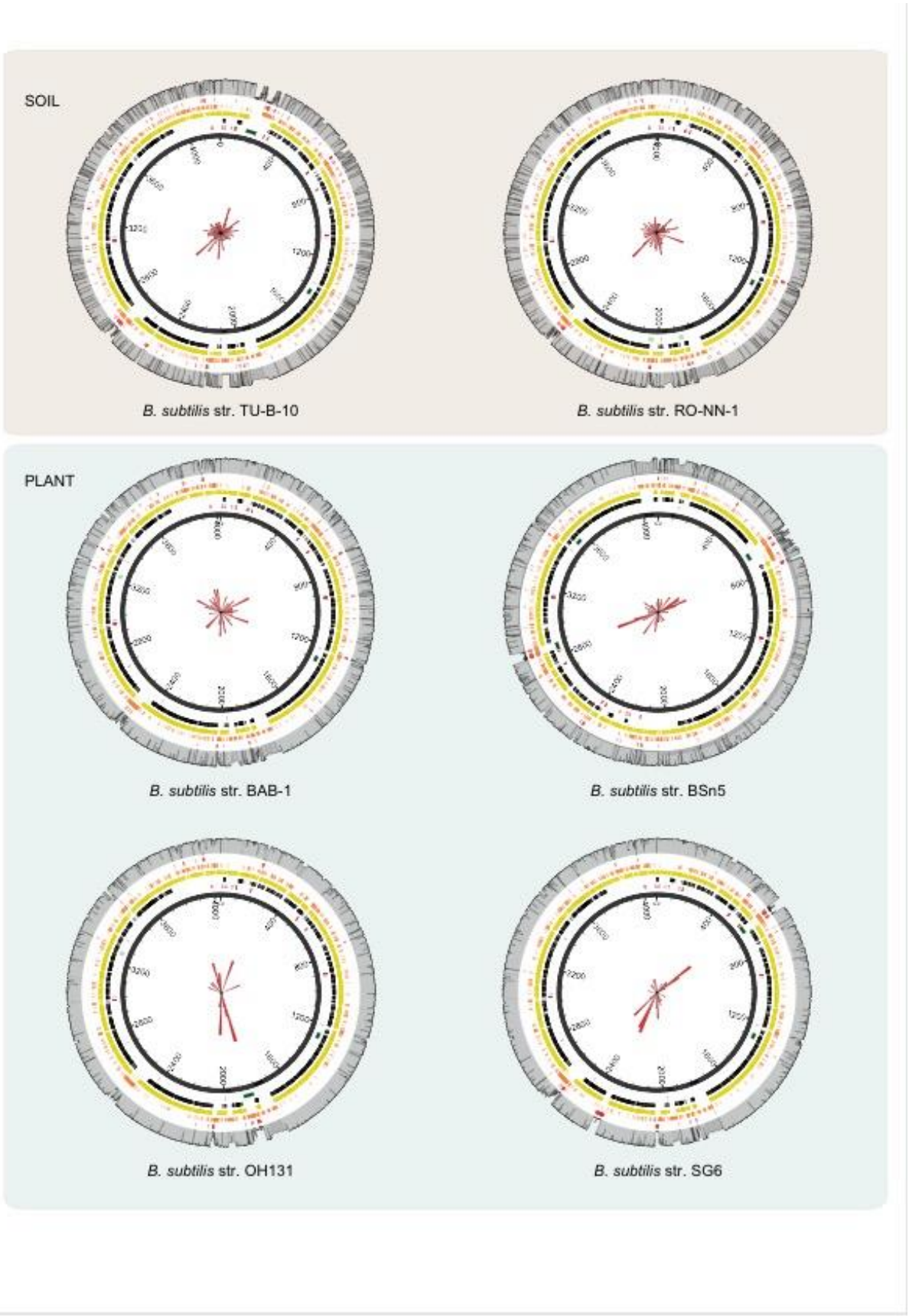

SOIL
B. subtilis str. TU-B-10
B. subtilis str. RO-NN-1
PLANT
B. subtilis str. BAB-1
B. subtilis str. BSn5
B. subtilis str. OH131
B. subtilis str. SG6

Fig. S5

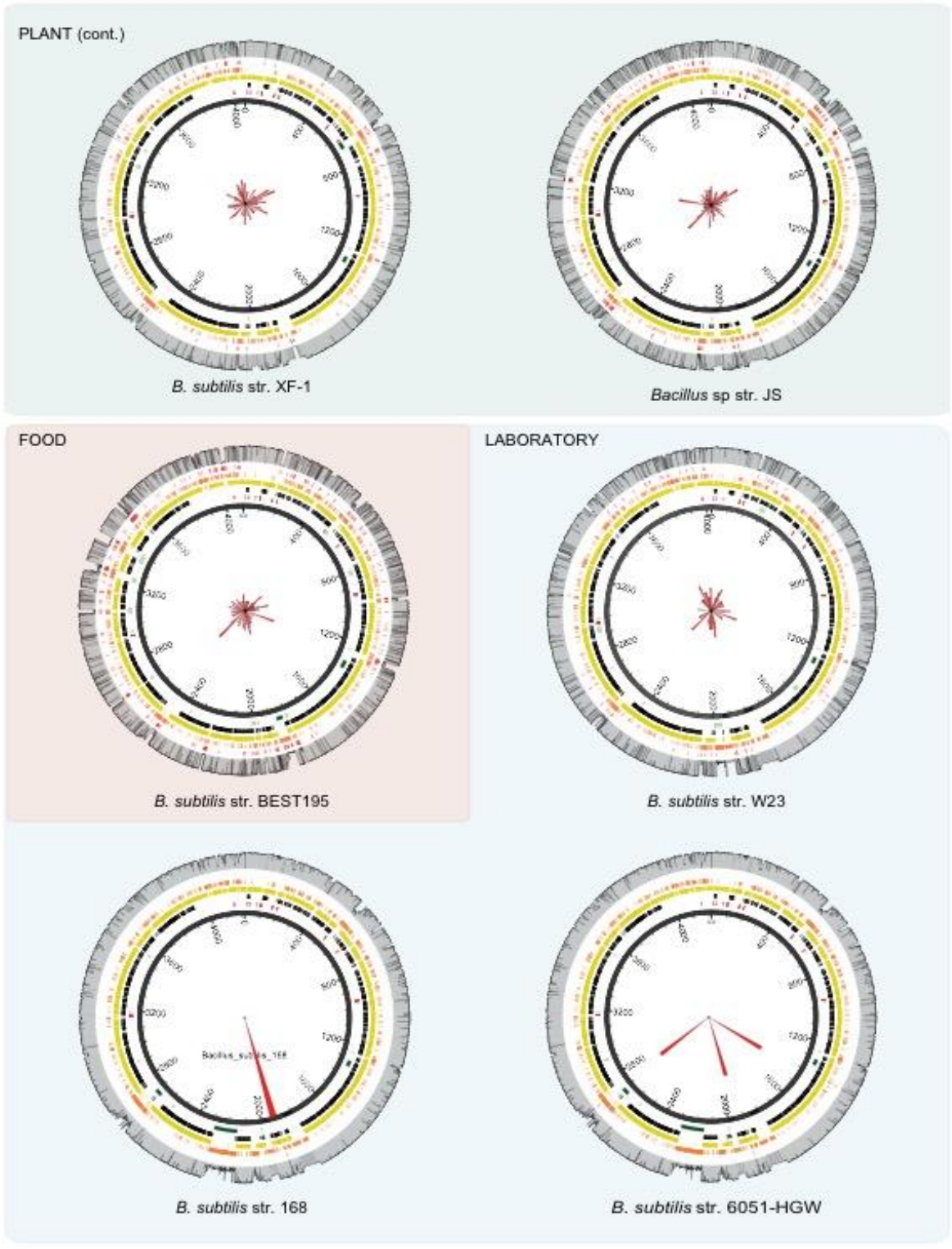

PLANT (cont.)
B. subtilis str. XF-1
Bacillus sp str. JS
FOOD
B. subtilis str. BEST195
LABORATORY
B. subtilis str. W23
B. subtilis str. 168
B. subtilis str. 6051-HGW

Fig. S5 (cont.)

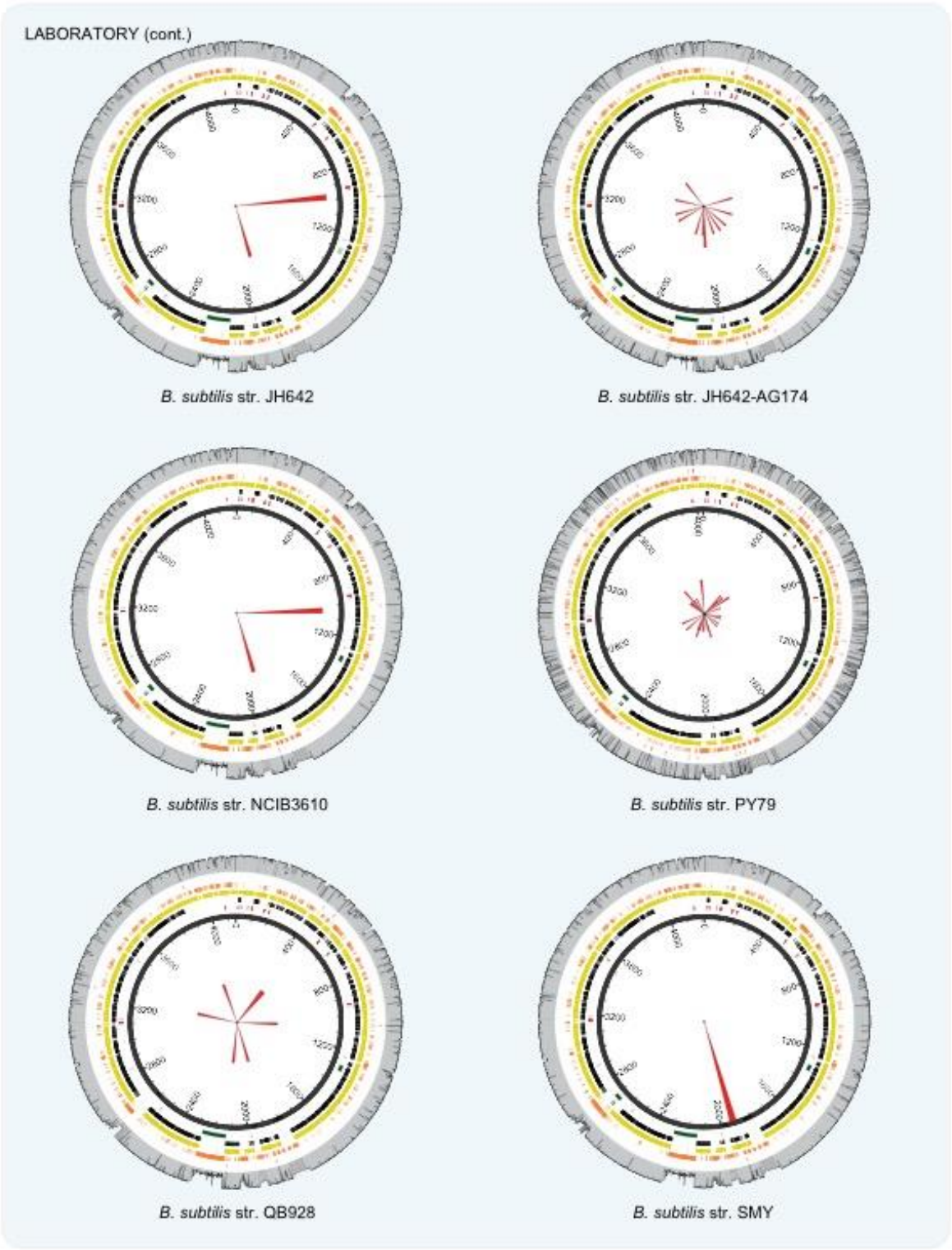

LABORATORY (cont.)
B. subtilis str. JH642
B. subtilis str. JH642-AG174
B. subtilis str. NCIB3610
B. subtilis str. PY79
B. subtilis str. QB928
B. subtilis str. SMY

Fig. S5 (cont.)


B.subtilis str. TU-B-10
B.vallismortis str. DV1-F-3
B.subtilis str. DV1-B-1
B.subtilis str. BST
B.subtilis str. W23
B.subtilis str. ATCC 6633
B.subtilis str. gtP20b
B.subtilis str. KCTC 13429
B.tequilensis str. KCTC 13622
B.subtilis str. D7XPN1
Bacillus sp. str. JS
B.subtilis str. BSP2
B.subtilis str. natto BEST 195
B.subtilis str. BSP4
B.subtilis str. SG6
B.subtilis str. KATMIRA 1933
B.subtilis str. MP9
B.subtilis str. MP11
B.subtilis str. BAB-1
B.subtilis str. XF-1
B.subtilis str. PTS-394
B.subtilis str. RO-NN-1
B.subtilis str. PS216
B.subtilis str. QH-1
B.subtilis str. 168
B.subtilis str. NDfood
B.subtilis str. NCIB 3610
B.subtilis str. AUSI98
B.subtilis str. MB73_2
B.subtilis str. QB928
B.subtilis str. NDmed
B.subtilis str. BSn5
B.subtilis str. OH 131.1
B.subtilis str. SC-8
B. subtilis str. BSP1
B.subtilis str. JH642
B. subtilis str. 6051-HGW
B. subtilis str. JH642 substr. AG174
B. subtilis str. SMY
B. subtilis str. PY79
B. subtilis str. GXA-28
B. atrophaeus str. NRS 1221A
B. atrophaeus str. AS43.3
B. amyloliquefaciens str. AS43.3
B. amyloliquefaciens str. NAU-B3
B. amyloliquefaciens str. CAU B946
B. amyloliquefaciens str. TA208
B. amyloliquefaciens str. DSM 7
Lysinibacillus sphaericus
Lysinibacillus fusiformis
B. cereus str. ATCC 10987
B. mycoides str. ATCC 6462
B. thuringiensis str. IS5056
B. cereus str. ATCC 14579
B. thuringiensis str. HD-789
B. cereus str. HD-771
B. toyonensis
B. thuringiensis str. HD-1011
Paenibacillus graminis
Paenibacillus sabinae
Paenibacillus polymyxa str. Sb3-1
Paenibacillus terrae
Paenibacillus johnsonii str. DPC 6026
Lactococcus lactis str. CV56
Enterococcus hirae
Enterococcus faecalis str. Symbioflor 1
Pediococcus acidilactici
Pediococcus pentosaceus str. ATCC 25745
Lactobacillus paraplantarum
Pediococcus claussenii
Lactobacillus casei str. BL23
Enterococcus gallinarum
Enterococcus casseliflavus
B. thuringiensis str. HD-789
B. thuringiensis str. HD-771
Lactobacillus johnsonii str. DPC 6026
Streptococcus pneumoniae str. A026
Streptococcus pneumoniae str. TCH8431_19A
Lactobacillus paraplantarum
Lactobacillus plantarum str. 16
B. amyloliquefaciens str. TA208
B. amyloliquefaciens str. DSM 7
Listeria ivanovii str. PAM 55
Listeria monocytogenes str. M7
Planococcus halocryophilus
Planococcus rifietoensis
B. licheniformis str. ATCC 14580
B. pseudofirmus
B. cellulosilyticus
B. thuringiensis str. IS5056
B. subtilis str. OH 131.1
B. subtilis str. BSn5
B. lehensis
Staphylococcus agnetis
Staphylococcus lugdunensis str. HKU09-01
Staphylococcus epidermidis str. ATCC 12228
Staphylococcus aureus str. MRSA252
Staphylococcus saprophyticus
B. subtilis str. PTS-394
B. subtilis str. XF-1
B. subtilis str. BAB-1
Solibacillus silvestris str. DSM 12223
Paenibacillus polymyxa str. Sb3-1
Solibacillus silvestris str. DSM 12223
B. lehensis
B. lehensis
0.3
B. subtilis softcore gene
B. subtilis cloud gene
B. subtilis shell gene

Fig. S6